Day_13_Assignment

1.what is statistics?

Statistics is the science of collecting, analyzing, interpreting, and presenting data.

Example:

- Finding average salary
- Predicting sales
- Comparing student marks

2. Explain briefly about the types of statistics?

Types of Statistics

1. Descriptive Statistics

- Summarizes data
- Mean, Median, Mode
- Variance, Standard Deviation

Example:
Average marks of class.

2. Inferential Statistics

- Makes predictions about population
- Uses sample data
- Hypothesis testing, t-test, ANOVA

Example:
Predicting election result from sample survey.

3. What is central tendency?explain with hands on code?

Types:

- Mean (average)
- Median (middle value)
- Mode (most frequent value)

Hands-on Python Code

```
import numpy as np

import statistics as stats
```

```
data = [10, 20, 30, 40, 50, 50]
mean = np.mean(data)
median = np.median(data)
mode = stats.mode(data)
print("Mean:", mean)
print("Median:", median)
print("Mode:", mode)
```

Output:

```
Mean: 33.33
Median: 35.0
Mode: 50
```

4. What is measures of dispersion?explain about range,var,std,iqr?

Measures of Dispersion

Shows how spread out data is.

Range

Highest – Lowest

Variance

Average squared difference from mean

Standard Deviation (STD)

Square root of variance

IQR (Inter Quartile Range)

Q3 – Q1

Code

```
import numpy as np
data = [10, 20, 30, 40, 50]
print("Range:", max(data) - min(data))
print("Variance:", np.var(data))
print("Standard Deviation:", np.std(data))
print("IQR:", np.percentile(data, 75) - np.percentile(data, 25))
```

5. what is t-test?

Used to compare means of two groups when:

- Sample size is small
- Population variance unknown

Example:
Comparing marks of two classes.

6. what is z-test?

Used when:

- Sample size is large (>30)
- Population variance known

7. what is annova? what kind of the assumptions we can see?

ANOVA = Analysis of Variance
Used to compare means of 3 or more groups

Example:
Compare marks of Class A, B, C

Assumptions:

- Normal distribution
- Equal variances
- Independent samples

8. what is p-value?

Probability of getting results if null hypothesis is true.

If:

- $p < 0.05$ → Reject null hypothesis
- $p > 0.05$ → Accept null hypothesis

9. what is hypothesis testing?

Process to test assumption about population.

Steps:

1. Null hypothesis (H_0)
2. Alternative hypothesis (H_1)
3. Choose significance level
4. Calculate test statistic
5. Compare p-value

10. what is z-score?

Shows how many standard deviations a value is from mean.

Formula:

$Z = (X - \text{Mean}) / \text{Std}$

Example:
If Z = 2 → value is 2 std above mean

11. why significance level is important explain briefly about it?

Usually 0.05

It controls:

- Type 1 error (False positive)

Lower significance level → more strict test

12. what is chisquare? explain the description about it?

Used for:

- Categorical data
- Testing independence

Example:
Gender vs Product choice

13. how the data is distributed explain with different examples?

◈ Normal Distribution

Bell-shaped curve

◈ Skewed Distribution

- Right skewed
- Left skewed

◈ Uniform Distribution

All values equal probability.

14. what is inferential statistics?

Making predictions about population using sample data.

Example:
Predicting total sales from sample region.

15. what is correlation?

Measures relationship between variables.

Range:

- -1 to +1

+1 → Perfect positive
-1 → Perfect negative.

16. what is regression analysis explain briefly about it?

Regression Analysis is a statistical method used to predict or estimate a dependent variable (output) based on one or more independent variables (inputs).

In simple words:
It helps us understand the relationship between variables and make predictions.

◈ Example

Suppose:

- Experience (Independent Variable – X)
- Salary (Dependent Variable – Y)

We can predict salary based on experience using regression.

◈ Types of Regression

Linear Regression

Used when the relationship between variables is linear (straight line).

Equation:

$$Y = a + bX$$

Where:

- Y → Dependent variable
- X → Independent variable
- a → Intercept
- b → Slope

Multiple Linear Regression

Used when there are multiple independent variables.

Example:

- Salary depends on experience, education, and skills.

Logistic Regression

Used for classification problems (Yes/No, 0/1).

Example:

- Will customer buy product? (Yes/No)

Simple Hands-on Example (Linear Regression in Python)

```
import numpy as np
from sklearn.linear_model import LinearRegression
# Experience (X)
X = np.array([1, 2, 3, 4, 5]).reshape(-1,1)
# Salary (Y)
Y = np.array([20000, 25000, 30000, 35000, 40000])
 model = LinearRegression()
model.fit(X, Y)
 # Predict salary for 6 years experience
prediction = model.predict([[6]])
print("Predicted Salary:", prediction[0])
```

17. what are independent and dependent variables?

Independent Variable (X)

The independent variable is the variable that you control, change, or use as input to see its effect.

It is also called:

- Predictor variable
- Input variable
- Feature (in ML)

It does NOT depend on other variables.

Dependent Variable (Y)

The dependent variable is the variable that depends on the independent variable.

It is also called:

- Output variable
- Target variable

- Response variable

It changes because of the independent variable.

18. how statistics will impact on ml models?

Data Understanding (Descriptive Statistics)

Before building any ML model, we analyze:

- Mean
- Median
- Standard deviation
- Distribution
- Skewness

Helps in:

- Understanding data behavior
- Detecting imbalance
- Detecting anomalies

Example:
If salary data is highly skewed → We may apply log transformation.

Handling Outliers

Statistics helps detect outliers using:

- Z-score
- IQR method

Outliers can:

- Mislead regression models
- Affect distance-based models (KNN, K-means)

Feature Selection

Using statistical techniques:

- Correlation
- Chi-square test
- ANOVA
- p-values

- Helps in selecting important features
  Reduces overfitting
  Improves model performance

Probability Theory

Many ML algorithms are based on probability:

- Naive Bayes
- Logistic Regression
- Bayesian Models

Probability helps in:

- Predicting class likelihood
- Handling uncertainty

Hypothesis Testing

Used to:

- Check feature importance
- Validate assumptions
- Compare models

Example:
Is this feature statistically significant?
Check p-value.

19. what are data attributes?

haracteristics of data.

Examples:

- Age
- Salary
- Gender
- City

20. what is qualitative and quantitative?

Qualitative data describes qualities or categories.
It is non-numerical.

It tells *what type* or *which category*.

Examples:

- Gender (Male/Female)

- City (Hyderabad, Delhi)
- Color (Red, Blue)
- Product Type (Electronics, Clothing)

Types of Qualitative Data:

Nominal

- No order
- Example: Blood group, Gender

Ordinal

- Has order
- Example: Rating (Poor, Good, Excellent)

Quantitative Data (Numerical Data)

Quantitative data represents numbers and measurable quantities. It tells *how much* or *how many*.

Examples:

- Age
- Salary
- Height
- Temperature
- Marks

Types of Quantitative Data:

Discrete

- Countable numbers
- Example: Number of students, number of cars

Continuous

- Can take any value in a range
- Example: Height, Weight, Temperature

21. what is the difference between continuous and categorical data?

1.Continuous Data

Continuous data can take any value within a range (including decimals).

It is numeric and measurable.

Examples:

- Height (170.5 cm)
- Weight (65.3 kg)
- Temperature (36.7°C)
- Salary (25000.75)

Key Characteristics:

- Infinite possible values
- Includes decimals
- Measured using instruments
- Used in regression models

2.Categorical Data

Categorical data represents groups or categories.

It is non-numeric (or numbers used as labels).

Examples:

- Gender (Male/Female)
- Blood Group (A, B, O)
- City (Delhi, Mumbai)
- Pass/Fail

Key Characteristics:

- Fixed categories
- No decimal values
- Used in classification models

Comparison Table

Continuous Data	Categorical Data
Numeric	Non-numeric
Measurable	Group-based
Infinite values	Limited categories
Example: Height	Example: Gender

simple Real Example

If you are analyzing students:

- Marks = Continuous
- Grade (A/B/C) = Categorical

In Machine Learning

- Continuous → Used in regression
- Categorical → Used in classification
- Categorical needs encoding (One-hot encoding)
- 22.what is data?explain about different types in it?

Data is raw facts, figures, or information collected for analysis.

It can be numbers, text, images, sounds, or observations.
Data becomes information after processing and analysis.

Example:

- 25, 30, 45 → Data
- Average age = 33 → Information

◈ Types of Data

Data can be classified in different ways:

Based on Nature

A) Qualitative Data (Categorical)

Describes qualities or categories.

Examples:

- Gender
- City
- Color
- Product type

Types:

- Nominal (No order) → Gender
- Ordinal (Has order) → Ratings (Low, Medium, High)

B) Quantitative Data (Numerical)

Represents measurable numbers.

Examples:

- Age
- Salary
- Height
- Marks

Types:

- Discrete → Countable (Number of students)
- Continuous → Measurable (Weight, Temperature)

Based on Structure

A) Structured Data

- Organized in rows & columns
- Stored in databases, Excel, SQL tables

Example:

Name Age Salary

B) Unstructured Data

- No fixed format
- Hard to organize

Examples:

- Images
- Videos
- Emails
- Social media posts

C) Semi-Structured Data

- Partially organized
- JSON, XML files

Based on Source

Primary Data

Collected directly by researcher.

Example:
Survey, Interview

Secondary Data

Collected by someone else.

Example:
Government reports, Websites

23. explain about structured and unstructured data?

Structured Data

Structured data is data that is organized in a fixed format, usually in rows and columns. It follows a predefined schema (table format).
Easy to store, search, and analyze.

Examples:

- Excel sheets
- SQL databases
- CSV files
- Banking transaction records

Example Table:

Name	Age	Salary
Ram	25	30000
Sita	28	35000

Characteristics:

- Organized
- Easy to query (using SQL)
- Mostly numeric & categorical
- Used in traditional ML models

Unstructured Data

Unstructured data has no predefined format or structure.

It cannot be stored easily in tables.

Examples:

- Images

- Videos
- Audio files
- Emails
- Social media posts
- PDFs

Example:

- A WhatsApp message
- A YouTube video
- An Instagram image

Characteristics:

- No fixed format
- Hard to analyze directly
- Requires preprocessing
- Used in NLP & Computer Vision

24. what are outliers?why these are important?

Outliers are data points that are very different from the rest of the dataset.

They are unusually high or low values compared to other observations.

Example:

Dataset:

10, 12, 15, 18, 20, 200

200 is an outlier because it is far away from other values.

Why Are Outliers Important?

Outliers are important because they can:

Affect Mean and Standard Deviation

Outliers can distort the average.

Example:

Without 200 → Mean ≈ 15

With 200 → Mean becomes very high

Impact Machine Learning Models

Some models are sensitive to outliers:

- Linear Regression
- KNN
- K-Means Clustering

Outliers can reduce model accuracy.

Indicate Errors

Outliers may represent:

- Data entry mistakes
- Measurement errors
- Incorrect data collection

Example:
Age = 250 years (data error)

Reveal Important Insights

Sometimes outliers are meaningful!

Example:

- Fraud detection
- Abnormal medical reports
- Rare events

In fraud detection, outliers are actually useful.

Types of Outliers

1.Global Outliers → Very different from all data
2.Contextual Outliers → Different in specific condition
3.Collective Outliers → Group of unusual points.

25. how to find the outliers in the dataset?

Outliers can be detected using statistical methods and visualization techniques.

Using Z-Score Method

Concept:

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

$$Z = \frac{X - \text{Mean}}{\text{Standard Deviation}}$$

If $|Z| > 3$ → It is usually considered an outlier.

Python Example

```
import numpy as np
from scipy import stats
data = np.array([10, 12, 15, 18, 20, 200])
z_scores = np.abs(stats.zscore(data))
outliers = data[z_scores > 3]
print("Outliers:", outliers)
```

Using IQR (Interquartile Range) Method

Concept:

$IQR = Q3 - Q1$

Outlier limits:

- Lower limit = $Q1 - 1.5 \times IQR$
- Upper limit = $Q3 + 1.5 \times IQR$

Values outside these limits are outliers.

Python Example

```
import numpy as np
data = np.array([10, 12, 15, 18, 20, 200])
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
IQR = Q3 - Q1
lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR
outliers = data[(data < lower_limit) | (data > upper_limit)]
print("Outliers:", outliers)
```

Using Box Plot (Visualization)

Boxplot visually shows outliers as points outside the whiskers.

```
import matplotlib.pyplot as plt
plt.boxplot(data)
plt.show()
```

Using Scatter Plot

Useful for detecting outliers in two variables.

```
plt.scatter(range(len(data)), data)
plt.show()
```