**1. What is Statistics? Explain its main types.**

Statistics is the branch of mathematics that deals with **collecting, organizing, analyzing, and interpreting data** to make decisions and predictions.

**Types:**

1. **Descriptive Statistics** – Summarizes and describes data (e.g., mean, median, charts).
2. **Inferential Statistics** – Makes predictions or generalizations about a population based on a sample (e.g., hypothesis testing).

**2. Define population and sample with examples.**

* **Population**: The entire group of individuals or data of interest.  
   Example: All students in a school.
* **Sample**: A smaller group selected from the population for analysis.  
   Example: 100 students chosen from the school for a survey.

**3. Difference between descriptive and inferential statistics**

* **Descriptive**: Describes data using numbers, graphs, and summaries.  
   Example: Average height of 100 students = 160 cm.
* **Inferential**: Uses sample data to make predictions about the population.  
   Example: Predicting the average height of all students in the school based on a sample.

**4. Data types (qualitative vs quantitative, discrete vs continuous)**

* **Qualitative (Categorical)**: Non-numeric, descriptive data.
* Example: Gender (Male/Female), Colors.
* **Quantitative (Numerical)**: Numeric values.
  + **Discrete**: Countable numbers. Example: Number of students.
  + **Continuous**: Measurable values, can take decimals. Example: Height, Weight.

**5. What is a variable in statistics? Give examples.**

A **variable** is a characteristic that can take different values.  
 Examples: Age, salary, temperature, exam scores.

**6. Define mean, median, and mode. How are they different?**

* **Mean**: Arithmetic average.
* **Median**: Middle value when data is ordered.
* **Mode**: Most frequently occurring value.

Example: Data = [2, 3, 3, 5, 7]

* Mean = (2+3+3+5+7)/5 = 4
* Median = 3
* Mode = 3

**7. How do you calculate the range of a dataset?**

Range = **Maximum value – Minimum value**.  
 Example: For [4, 8, 15, 20], Range = 20 – 4 = 16.

**8. What is the standard deviation, and why is it important?**

Standard deviation (SD) measures the **spread of data around the mean**.

* Low SD → Data is close to the mean.
* High SD → Data is widely spread.  
   Important because it shows data variability.

**9. Explain variance and how it relates to standard deviation.**

Variance = Average of squared differences from the mean.

Variance=σ2\text{Variance} = \sigma^2Variance=σ2

Standard deviation is the **square root of variance**.

**10. What is a frequency distribution? Give an example.**

Frequency distribution shows how often each value (or range of values) occurs.

Example: Student test scores:

* 0–10 → 2 students
* 11–20 → 5 students
* 21–30 → 8 students

**11. Explain the concept of normal distribution and its characteristics.**

Normal distribution = Bell-shaped curve where data is symmetrically distributed around the mean.

**Characteristics:**

* Mean = Median = Mode.
* 68% of data within 1 SD, 95% within 2 SD, 99.7% within 3 SD (Empirical Rule).

**12. What is skewness, and how does it affect data interpretation?**

Skewness measures asymmetry in data.

* **Positive skew (right-skewed)**: Tail is longer on the right.
* **Negative skew (left-skewed)**: Tail is longer on the left.  
   It affects whether mean > median or mean < median.

**13. What is kurtosis, and what does it tell us about a dataset?**

Kurtosis measures the **peakedness or flatness** of a distribution.

* **High kurtosis**: More outliers, sharp peak.
* **Low kurtosis**: Flatter distribution.

**14. Differentiate between probability and statistics.**

* **Probability**: Starts with known data to predict outcomes.  
   Example: Tossing a coin, probability of heads = 0.5.
* **Statistics**: Starts with data and makes inferences about the population.  
   Example: Collecting coin toss data and estimating probability.

**15. What is a z-score, and how is it calculated?**

Z-score = Number of standard deviations a value is from the mean.

Z=X−μσZ = \frac{X - \mu}{\sigma}Z=σX−μ​

Example: If mean = 50, SD = 10, X = 70 → Z = (70–50)/10 = 2.

**16. Difference between population standard deviation and sample standard deviation**

* **Population SD (σ):** Uses N (entire population).
* **Sample SD (s):** Uses n–1 (sample correction, Bessel’s correction).

**17. What is the Central Limit Theorem, and why is it important?**

The CLT states that the sampling distribution of the sample mean becomes approximately **normal**, regardless of population distribution, if the sample size is large enough (n ≥ 30).

Important for hypothesis testing and confidence intervals.

**18. What is correlation? Differentiate between positive and negative correlation.**

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

* **Positive correlation**: As one increases, the other increases (e.g., height vs weight).
* **Negative correlation**: As one increases, the other decreases (e.g., exercise vs body fat).

**19. Difference between correlation and causation.**

* **Correlation**: Two variables are related but not necessarily cause-effect.
* **Causation**: One variable directly affects the other.  
   Example: Ice cream sales & drowning are correlated (summer), but ice cream doesn’t cause drowning.

**20. What is regression analysis, and when is it used?**

Regression analysis is used to model the relationship between a **dependent variable** and one or more **independent variables**.  
 Example: Predicting house price (dependent) based on size, location, and rooms (independent).

**21. Explain hypothesis testing and its steps.**

Hypothesis testing is a statistical method to make decisions about population parameters using sample data.

**Steps:**

1. Formulate **Null (H₀)** and **Alternative (H₁)** hypotheses.
2. Choose significance level (α).
3. Calculate test statistic.
4. Find p-value or compare with critical value.
5. Accept or reject H₀.

**22. What is a null hypothesis and an alternative hypothesis?**

* **Null Hypothesis (H₀):** Assumes no effect or no difference.  
   Example: "The new drug has no effect."
* **Alternative Hypothesis (H₁):** Assumes effect or difference exists.  
   Example: "The new drug improves recovery."

**23. Explain p-value in hypothesis testing.**

The p-value measures the probability of getting results as extreme as observed if H₀ is true.

* **Low p-value (≤ 0.05):** Reject H₀ → Evidence supports H₁.
* **High p-value (> 0.05):** Fail to reject H₀.

**24. Difference between Type I and Type II errors.**

* **Type I Error (False Positive):** Rejecting H₀ when it is true.  
   Example: Saying a drug works when it doesn’t.
* **Type II Error (False Negative):** Failing to reject H₀ when H₁ is true.  
   Example: Saying a drug doesn’t work when it actually does.

**25. What is a confidence interval, and how is it interpreted?**

A confidence interval (CI) gives a range of values within which the true population parameter lies with a certain probability (usually 95%).

Example: "The average height of students is 160–170 cm with 95% confidence."

**26. Explain t-test and when to use it.**

A **t-test** compares the means of two groups to see if they are significantly different.  
 Used when sample size is small (< 30) or population SD is unknown.

**27. Explain chi-square test and its applications.**

The **Chi-square test** checks if there is a significant association between categorical variables.

Example: Checking if gender and voting preference are related.

**28. What is ANOVA, and when is it used?**

**ANOVA (Analysis of Variance):** Compares the means of **3 or more groups** to check if at least one is significantly different.  
Example: Comparing exam scores of students from three different teaching methods.

**29. How do you handle missing data in statistics?**

Methods:

* Remove rows with missing values (if few).
* Replace with mean/median/mode.
* Use regression or ML algorithms to predict missing values.
* Use advanced methods (Multiple Imputation, KNN imputation).

**30. What is sampling bias, and how can it be reduced?**

Sampling bias occurs when the sample is not representative of the population.  
 Example: Only surveying young people to generalize about all ages.

**How to reduce:**

* Use random sampling.
* Increase sample size.
* Avoid selective data collection.