1. What is the difference between mean, median, and mode?

- **Definition** :
  - **Mean** : The average of all values.
  - **Median** : The middle value in an ordered dataset.
  - **Mode** : The most frequently occurring value.
- **Formulas** :
  - Mean: $x^{-}=n\sum xi$
  - Median: Middle value (or average of two middle values if n is even).
  - Mode: Value with the highest frequency.
- **Example** :
  For [2,3,3,5,7]:
  - Mean = 4, Median = 3, Mode = 3.

---

2. How do outliers affect the mean and median?

- **Definition** :
  - **Outliers** : Extreme values that deviate significantly from other observations.
- **Effect** :
  - **Mean** : Sensitive to outliers (pulled toward extreme values).
  - **Median** : Robust to outliers.
- **Example** :
  For [1,2,3,4,100]:
  - Mean = 22, Median = 3.

---

3. What is the formula for calculating variance?

- **Definition** :
  - **Variance** : Measures the spread of data around the mean.
- **Formulas** :
  - Population variance: $\sigma 2=N\sum(xi-\mu)2$
  - Sample variance: $s2=n-1\sum(xi-x^{-})2$
- **Example** :
  For [1,2,3,4,5], sample variance = 2.5.

---

4. What does a standard deviation of zero indicate?

- **Definition** :
  - **Standard Deviation** : Measures the average distance of data points from the mean.
- **Interpretation** :
  - If $\sigma=0$, all values in the dataset are identical.
- **Example** :
  Dataset [5,5,5] has $\sigma=0$.

---

5. Explain the interquartile range (IQR) and its use.

- **Definition** :
    - **IQR** : The range between the first quartile (Q1) and third quartile (Q3).
- **Formula** :$IQR = Q3 - Q1$
- **Use** : Identifies outliers and summarizes spread.
- **Example** :
  For [1,3,5,7,9], IQR = $7 - 3 = 4$.

---

6. What is skewness? Describe left-skewed and right-skewed distributions.

- **Definition** :
    - **Skewness** : Asymmetry in the data distribution.
- **Types** :
    - **Left-skewed** : Tail on the left (mean < median).
    - **Right-skewed** : Tail on the right (mean > median).
- **Example** :
  Income data is typically right-skewed.

---

7. How do you identify outliers using the IQR method?

- **Definition** :
    - **Outliers** : Values outside $Q1 - 1.5 \times IQR$ or $Q3 + 1.5 \times IQR$.
- **Steps** :
    - Calculate Q1 and Q3.
    - Compute IQR.
    - Define bounds:
        - Lower bound: $Q1 - 1.5 \times IQR$
        - Upper bound: $Q3 + 1.5 \times IQR$
- **Example** :
  For [1,3,5,7,9], outliers are values < -3 or > 13.

---

8. What is kurtosis, and how does it differ from skewness?

- **Definition** :
    - **Kurtosis** : Measures the "tailedness" of the distribution.
    - **Skewness** : Measures asymmetry.
- **Types of Kurtosis** :
    - **Leptokurtic** : Heavy tails (high kurtosis).
    - **Platykurtic** : Light tails (low kurtosis).

---

9. What is the empirical rule (68-95-99.7 rule)?

- **Definition** :
    - Rule describing the percentage of data within standard deviations in a normal distribution:
        - 68% within $\mu \pm \sigma$,
        - 95% within $\mu \pm 2\sigma$,
        - 99.7% within $\mu \pm 3\sigma$.

---

10. What is the difference between a population and a sample?

- **Definition** :
    - **Population** : Entire group of interest.
    - **Sample** : Subset of the population used for analysis.

---

11. What is a percentile? How is it different from a quartile?

- **Definition** :
    - **Percentile** : Value below which a given percentage of data falls.
    - **Quartile** : Specific percentiles (25th, 50th, 75th).

---

12. How do you calculate the coefficient of variation?

- **Definition** :
    - **Coefficient of Variation (CV)** : Measures relative variability.
- **Formula** :$CV = \frac{\sigma}{\mu} \times 100\%$
- **Example** :
  If $\sigma = 10$, $\mu = 50$, CV = 20%.

---

13. What is the difference between range and IQR?

- **Definition** :
    - **Range** : Max−Min.
    - **IQR** : Q3−Q1.
- **Key Difference** : IQR is robust to outliers; range is not.

---

14. What does a box plot show?

- **Definition** :
    - A visual summary of data showing:
        - Median (line),
        - Quartiles (box),
        - Outliers (dots).

---

15. What is the relationship between variance and standard deviation?

- **Definition** :
    - **Standard Deviation** : Square root of variance.
- **Formula** :$\sigma = \sigma 2$

---

16. When would you prefer the median over the mean?

- **Definition** :
    - Use the **median** for skewed data or datasets with outliers.

---

17. What is the 5-number summary?

- **Definition** :
    - Summary of data using:
        - Minimum, Q1, Median, Q3, Maximum.

---

18. How do you interpret a z-score?

- **Definition** :
    - **Z-score** : Number of standard deviations a value is from the mean.
- **Formula** :$z = \sigma x - \mu$
- **Example** :
  If z=2, the value is 2 SDs above the mean.

---

19. What does a covariance of zero imply?

- **Definition** :
    - **Covariance** : Measures the direction of the linear relationship between two variables.
- **Implication** :
    - Covariance = 0 suggests no linear relationship.

---

20. What is the difference between descriptive and inferential statistics?

- **Definition** :
    - **Descriptive** : Summarizes data (e.g., mean, median).
    - **Inferential** : Makes predictions/generalizations (e.g., hypothesis testing).

---

21. What is the trimmed mean, and when is it useful?

- **Definition** :
    - **Trimmed Mean** : Removes a percentage of extreme values before calculating the mean.

- **Use Case** : Reducing outlier influence (e.g., Olympic scoring).

---

22. How does the geometric mean differ from the arithmetic mean?

- **Definition** :
    - **Geometric Mean** : Useful for multiplicative relationships (e.g., growth rates).
    - **Arithmetic Mean** : Sum of values divided by count.
- **Formula** :Geometric Mean=nx1x2…xn

---

23. What is the harmonic mean, and where is it applied?

- **Definition** :
    - **Harmonic Mean** : Used for rates/ratios (e.g., average speed).
- **Formula** :HM=∑xi1n

---

24. Explain Winsorizing and its purpose.

- **Definition** :
    - **Winsorizing** : Replaces outliers with the nearest non-outlier values.
- **Purpose** : Reduces outlier impact while retaining sample size.

---

25. What is the midrange, and why is it rarely used?

- **Definition** :
    - **Midrange** : Average of the minimum and maximum values.
- **Formula** :Midrange=2Min+Max
- **Drawback** : Highly sensitive to outliers.

26. What is Chebyshev’s inequality, and how is it used?

- **Definition** :
  A theorem stating that **at least 1−k21** of data lies within k standard deviations from the mean, regardless of the distribution.
- **Formula** :P(∣X−μ∣≥kσ)≤k21
- **Example** :
  For k=2, at least 1−41=75% of data lies within 2 SDs of the mean.

---

27. What is the Durbin-Watson statistic?

- **Definition** :
  A test for **autocorrelation** in regression residuals.
- **Range** : 0 to 4.
    - 2: No autocorrelation.

    - <2: Positive autocorrelation.
    - >2: Negative autocorrelation.
- **Use** : Detect serial correlation in time series data.

---

28. What is the difference between sample size and standard error?

- **Sample Size (n)** : Number of observations.
- **Standard Error (SE)** :SE=nσ
  Relationship : Larger n reduces SE, increasing precision.
- **Example** :
  If σ=10 and n=100, SE=1.

---

29. What is the Gini coefficient, and how is it calculated?

- **Definition** :
  Measures **inequality** in a distribution.
- **Formula** :G=2n2μ∑i=1n∑j=1n|xi−xj|
    - 0: Perfect equality.
    - 1: Maximum inequality.
- **Example** :
  Income inequality in a population.

---

30. What is entropy in the context of data distributions?

- **Definition** :
  Measures **uncertainty** or disorder in a distribution.
- **Formula** :H(X)=−i=1∑np(xi)logp(xi)
- **Example** :
  A fair coin toss has entropy H=1 bit.

---

31. What is the Jaccard similarity coefficient?

- **Definition** :
  Measures similarity between two sets.
- **Formula** :J(A,B)=|A∪B||A∩B|
- **Example** :
  For sets A={1,2} and B={2,3}, J=31.

---

32. How do you calculate weighted averages?

- **Formula** :x¯=∑wi∑wixi

- **Example** :
  Course grade = $1(0.3\times80)+(0.7\times90)=87$.

---

33. What is Spearman's footrule?

- **Definition** :
  Measures rank correlation by summing absolute differences in ranks.
- **Formula** :$D=i=1\sum n|Ri-Si|$
- **Use** : Less sensitive to outliers than Pearson's correlation.

---

34. What is Kendall's tau?

- **Definition** :
  Measures ordinal association between two variables.
- **Formula** :$\tau=2n(n-1)C-D$
  - C: Concordant pairs.
  - D: Discordant pairs.
- **Range** : -1 to 1.

---

35. What is Cronbach's alpha?

- **Definition** :
  Measures **internal consistency** of a test.
- **Formula** :$\alpha=\sigma X2k\sigma2$
  - k: Number of items.
  - $\sigma2$: Variance of item scores.
- **Threshold** : $\alpha>0.7$ is acceptable.

---

36. What is item response theory (IRT)?

- **Definition** :
  A framework to model the relationship between latent traits (e.g., ability) and observed responses.
- **Example** :
  Used in standardized testing to score examinees.

---

37. What is principal component analysis (PCA)?

- **Definition** :
  Reduces dimensionality by transforming data into orthogonal components.
- **Steps** :
  1. Standardize data.

2. Compute covariance matrix.
3. Extract eigenvectors (principal components).

- **Example** :
  Compress 10 variables into 2 PCs.

---

38. What is factor analysis?

- **Definition** :
  Identifies latent factors that explain correlations among observed variables.
- **Example** :
  Grouping survey questions into underlying traits (e.g., "satisfaction").

---

39. What is canonical correlation?

- **Definition** :
  Measures the correlation between two sets of variables.
- **Example** :
  Relationship between health metrics (e.g., BMI, BP) and lifestyle factors (e.g., diet, exercise).

---

40. What is multiple correspondence analysis?

- **Definition** :
  Extends correspondence analysis to categorical variables.
- **Use** : Visualizes associations between categories.

---

41. What is t-distributed stochastic neighbor embedding (t-SNE)?

- **Definition** :
  A tool to visualize high-dimensional data in 2D/3D.
- **Key Feature** : Preserves local structures (clusters).
- **Example** :
  Visualizing clusters in gene expression data.

---

42. What is multidimensional scaling (MDS)?

- **Definition** :
  Represents pairwise distances between objects in lower dimensions.
- **Example** :
  Mapping customer similarity based on purchasing behavior.

---

43. What is cluster analysis?

- **Definition** :
  Groups similar data points into clusters.
- **Methods** :
  - **K-means** : Minimizes within-cluster variance.
  - **Hierarchical** : Builds a dendrogram.

---

44. What is discriminant analysis?

- **Definition** :
  Classifies data into predefined groups.
- **Types** :
  - **Linear (LDA)** : Assumes normality.
  - **Quadratic (QDA)** : Allows non-linear boundaries.

---

45. What is survival analysis?

- **Definition** :
  Analyzes **time-to-event** data (e.g., death, failure).
- **Key Metric** : Survival function $S(t)=P(T>t)$.

---

46. What is time-to-event data?

- **Definition** :
  Data where the outcome is the **time until an event occurs** .
- **Example** : Time until a customer churns.

---

47. What is the hazard function?

- **Definition** :
  The **instantaneous risk** of an event at time t, given survival until t.
- **Formula** :$h(t)=\lim_{\Delta t \to 0} \Delta t P(t \leq T < t+\Delta t \mid T \geq t)$

---

48. What is censoring in survival analysis?

- **Definition** :
  When the event is not observed for some subjects.
- **Types** :
  - **Right-censoring** : Event not observed (e.g., study ends).
  - **Left-censoring** : Event occurred before observation.

---

49. What is the Kaplan-Meier estimator?

- **Definition** :
  Estimates the survival function non-parametrically.
- **Formula** :$S(t)=\prod_{t_i \le t}(1-\frac{d_i}{n_i})$
  - d_i: Events at time t_i.
  - n_i: Subjects at risk at t_i.

---

50. What is the Cox proportional hazards model?

- **Definition** :
  A regression model for survival data.
- **Formula** :$h(t)=h_0(t)\exp(\beta_1 X_1+\cdots+\beta_p X_p)$
  - $h_0(t)$: Baseline hazard.
  - β: Coefficients for predictors.