# Unit 10.2. MCQs Set 1

## Results

**#1. Q1. In descriptive statistics, which of the following is a measure of central tendency that splits the dataset into two equal parts?**

☐
(A) Mean
☐
(B) Mode
☐
(C) Median
☐
(D) Variance

The median is the middle value when the data is arranged in order.

**#2. Q2. The arithmetic mean is generally suitable for numerical data, but can be misleading if the data:**

☐
(A) Is normally distributed
☐
(B) Has no outliers
☐
(C) Is highly skewed or has extreme outliers
☐
(D) Consists only of positive integers

**#3. Q3. Which of these is calculated by summing all values and dividing by the number of observations?**

☐
(A) Median
☐
(B) Mode
☐
(C) Mean
☐
(D) Standard error

The arithmetic mean is computed by dividing the total sum by the number of data points.

### #4. Q4. (Fill in the blank) The _____ is defined as the most frequently occurring value in a dataset.

☐
(A) Mean
☐
(B) Median
☐
(C) Mode
☐
(D) Range

The mode is the value that appears most frequently in the dataset.

### #5. Q5. Descriptive statistics typically do not include:

☐
(A) Calculating measures like mean, median, and mode
☐
(B) Drawing inferences about a population from a sample
☐
(C) Displaying data in histograms or boxplots
☐
(D) Summarizing variability with standard deviation or range

Descriptive statistics summarize the data at hand; inferential statistics are used to draw conclusions about a broader population.

### #6. Q6. Which statement about Analysis of Variance (ANOVA) is correct?

☐
(A) It’s used only for two–sample comparisons
☐
(B) It tests if more than two group means differ significantly
☐
(C) It discards within–group variance entirely
☐
(D) It is a commonly used tool in biostatistics

ANOVA is used to test for significant differences among the means of three or more groups.

### #7. Q7. A t-test for independent samples is typically used when:

☐
(A) The same group is measured before and after an intervention
☐
(B) You have two unrelated groups to compare on a continuous outcome
☐
(C) You have more than two groups to compare
☐
(D) The data is strictly categorical

An independent t-test compares the means of two distinct groups on a continuous variable.

### #8. Q8. (Fill in the blank) In hypothesis testing, “power” is the probability of correctly

**rejecting the null hypothesis when it is _____.**

☐
(A) True
☐
(B) Valid
☐
(C) False
☐
(D) Undefined

Power is defined as 1 – β, the probability of detecting an effect when the null hypothesis is false.

### #9. Q9. Sample size calculation often requires:

☐
(A) Desired power, significance level (alpha), and an estimate of the effect size
☐
(B) A fixed sample size regardless of effect size
☐
(C) Ignoring effect size in favor of random sampling
☐
(D) Only the significance level

Accurate sample size calculation depends on power, alpha, and effect size to detect a true effect.

### #10. Q10. The p–value in a statistical test typically indicates:

☐
(A) The probability that the null hypothesis is true
☐
(B) The probability of observing a result as extreme or more extreme, assuming the null hypothesis is true
☐
(C) The guaranteed correctness of the alternative hypothesis
☐
(D) The margin of error in the measurement

The p–value indicates the probability of obtaining results at least as extreme as the observed ones if the null hypothesis is true.

### #11. Q11. Which of the following is not an assumption of a one–way ANOVA?

☐
(A) Each group sample is independent
☐
(B) The data in each group is roughly normally distributed
☐
(C) Homogeneity of variances across groups
☐
(D) Each group must contain exactly 100 observations

ANOVA does not require equal sample sizes; each group does not have to have exactly 100 observations.

### #12. Q12. (Fill in the blank) A “Type I error” is also known as a _____ in hypothesis testing.

☐
(A) False positive
☐
(B) False negative
☐
(C) True positive
☐

(D) True negative

A Type I error occurs when a true null hypothesis is incorrectly rejected—a false positive.

### #13. Q13. The correlation coefficient (Pearson's r) measures:

☐
(A) Causal relationships
☐
(B) The strength and direction of the linear association between two continuous variables
☐
(C) Differences among group means
☐
(D) The proportion of variance explained by a regression model

Pearson's r quantifies the linear association between two continuous variables, including direction and strength.

### #14. Q14. If the correlation coefficient r = 0, it implies:

☐
(A) No linear relationship between the variables
☐
(B) A perfect negative correlation
☐
(C) A perfect positive correlation
☐
(D) An undefined relationship

An r value of 0 indicates no linear association between the variables.

### #15. Q15. In linear regression, the slope coefficient indicates:

☐
(A) The predicted change in the outcome for a one–unit change in the predictor
☐
(B) The proportion of variance in the outcome explained by the predictor
☐
(C) The value of the intercept
☐
(D) The error in predicting the outcome

The slope coefficient quantifies the expected change in the dependent variable for a one unit change in the independent variable.

### #16. Q16. Multiple regression allows for:

☐
(A) Only one predictor variable
☐
(B) Testing several predictor variables simultaneously to see how they collectively relate to the outcome
☐
(C) Ignoring the effects of confounding variables
☐
(D) Using only categorical predictors

Multiple regression models use more than one predictor to explain the variability in the outcome variable.

### #17. Q17. Survival analysis is used to:

☐

(A) Analyze time–to–event data, such as time until recovery or relapse in clinical studies
☐
(B) Describe the frequency distribution of categorical data
☐
(C) Calculate the mean of continuous data
☐
(D) Compare group medians only

Survival analysis techniques, like Kaplan–Meier and Cox regression, are used to analyze the time until an event occurs, including censoring.

### #18. Q18. (Fill in the blank) The _____ method estimates and plots survival probabilities over time for different groups.

☐
(A) Log-rank
☐
(B) Cox proportional hazards
☐
(C) Kaplan–Meier
☐
(D) Weibull

The Kaplan–Meier method is used to estimate survival probabilities and generate survival curves.

### #19. Q19. Genome mapping statistics might involve:

☐
(A) Calculating linkage disequilibrium, LOD scores, or association p–values for genetic variants
☐
(B) Only describing phenotypic data
☐
(C) Ignoring genetic markers entirely
☐
(D) Aggregating all data into a single average value

Genome mapping employs statistical measures (e.g., LOD scores) to associate genetic variants with traits.

### #20. Q20. Bioinformatics often deals with ‘big data’ from sequencing. The process typically includes:

☐
(A) Data encryption only
☐
(B) Data cleaning, alignment, variant calling, and statistical association tests
☐
(C) Manual entry of sequence data
☐
(D) Discarding data that are not perfectly clean

Bioinformatics workflows involve multiple steps from data cleaning to statistical analysis.

### #21. Q21. Which of these is a nominal type of data?

☐
(A) Heights in centimeters
☐
(B) Temperature in Celsius
☐
(C) Blood group categories like A, B, AB, O
☐

(D) (None)

Nominal data classify observations into distinct categories with no inherent order.

### #22. Q22. Ordinal data differs from nominal data in that:

☐
(A) They are measured on an interval scale
☐
(B) Ordinal categories have a ranked order, though differences between ranks are not necessarily equal
☐
(C) They are always numerical
☐
(D) They require no categorization

Ordinal data have an inherent order, unlike nominal data.

### #23. Q23. In big data contexts, ‘metadata’ is:

☐
(A) Data about the analysis methods
☐
(B) Data describing other data, such as file name, creation date, and experimental conditions
☐
(C) Irrelevant information
☐
(D) The raw experimental measurements only

Metadata provides essential contextual details that help organize and interpret the primary data.

### #24. Q24. ‘Multi–dimensional data’ might refer to:

☐
(A) Data collected over time only
☐
(B) Datasets with multiple features or variables per observation
☐
(C) Data with only one variable
☐
(D) Data that have no variability

Multi-dimensional data include multiple variables for each observation and require specialized analysis techniques.

### #25. Q25. (Fill in the blank) A ‘linear algebraic treatment’ of data might involve _____ to reduce dimensions or solve large systems.

☐
(A) Differential equations
☐
(B) Graph theory
☐
(C) Matrix operations
☐
(D) Logical reasoning

Matrix operations, such as eigenvalue decomposition, are key in reducing dimensions in data analysis.

### #26. Q26. Eigenvalues and eigenvectors are key in:

☐

(A) Calculating simple averages
☐
(B) Principal Component Analysis (PCA) and dimensionality reduction
☐
(C) Determining the median of a dataset
☐
(D) Creating pie charts

PCA uses eigen decomposition to identify the principal components that capture the most variance.

### #27. Q27. (Fill in the blank) _____ decomposition is a factorization approach used in big data to break a matrix into singular values and orthonormal vectors.

☐
(A) QR
☐
(B) LU
☐
(C) SVD (Singular Value Decomposition)
☐
(D) Cholesky

Singular Value Decomposition (SVD) decomposes a matrix into singular values and vectors, widely used for dimensionality reduction.

### #28. Q28. Exploratory data analysis (EDA) typically includes:

☐
(A) Ignoring data summaries
☐
(B) Generating summary statistics, histograms, box plots, and scatter plots to detect patterns and outliers
☐
(C) Only applying advanced machine learning models
☐
(D) Discarding any data that do not fit the hypothesis

EDA involves initial data visualization and summarization to understand structure and detect anomalies.

### #29. Q29. Descriptive statistics differ from inferential statistics by focusing on:

☐
(A) Drawing conclusions about a population
☐
(B) Summarizing datasets using measures like mean and standard deviation without making population inferences
☐
(C) Making predictions for future samples
☐
(D) Testing hypotheses

Descriptive statistics summarize the current dataset without extrapolating to a broader population.

### #30. Q30. The ‘mean absolute deviation’ around the mean is:

☐
(A) The average of the squared differences between each data point and the mean
☐
(B) The average of the absolute differences between each data point and the mean
☐
(C) The difference between the maximum and minimum values
☐
(D) The standard deviation divided by the mean

Mean absolute deviation is calculated as the average of the absolute deviations from the mean, providing a measure of dispersion.

### #31. Q31. Match the following data types with their descriptions:

**1. Nominal data;**
**2. Ordinal data;**
**3. Interval data;**
**4. Ratio data**

**(a) Rank-ordered categories (e.g., mild, moderate, severe)**
**(b) Named categories with no inherent order (e.g., hair color)**
**(c) Numeric scale with a meaningful absolute zero (e.g., weight, height)**
**(d) Numeric scale with equal intervals but an arbitrary zero (e.g., temperature in °C)**

☐
(A) 1–b, 2–a, 3–d, 4–c
☐
(B) 1–a, 2–b, 3–c, 4–d
☐
(C) 1–c, 2–d, 3–a, 4–b
☐
(D) 1–d, 2–c, 3–b, 4–a

Nominal data are categories with no order (b), ordinal data are rank-ordered (a), interval data have equal spacing (d), and ratio data have a true zero (c).

### #32. Q32. Survival analysis methods may incorporate ‘censoring,’ which arises when:

☐
(A) Data collection is halted abruptly
☐
(B) A participant does not experience the event by the study end or is lost to follow-up
☐
(C) The data are missing completely at random
☐
(D) None of the above

Censoring occurs when the event of interest is not observed in some participants, due to dropout or study end.

### #33. Q33. The standard error of the mean (SEM) indicates:

☐
(A) The median’s variability
☐
(B) How far the sample mean is likely to be from the population mean
☐
(C) The range of the data
☐
(D) The overall sample size

SEM estimates the precision of the sample mean as an estimate of the population mean.

### #34. Q34. In correlation analysis, $r^2$ (the coefficient of determination) represents:

☐
(A) The strength of the linear relationship only
☐
(B) The proportion of the variance in one variable that is explained by another

☐
(C) The p-value of the correlation
☐
(D) The slope of the regression line

The coefficient of determination ($r^2$) quantifies the proportion of variance in one variable accounted for by the other.

### #35. Q35. (Fill in the blank) The _____ distribution is often used to model count data, such as the number of events in a fixed time interval.

☐
(A) Exponential
☐
(B) Normal
☐
(C) Poisson
☐
(D) Uniform

The Poisson distribution is commonly used for modeling the number of occurrences of an event in a fixed interval of time or space.

### #36. Q36. The Wilcoxon Rank–Sum (Mann–Whitney) test compares:

☐
(A) Mean values of two groups assuming normality
☐
(B) Two independent groups on ordinal or non–normally distributed continuous data using ranks
☐
(C) Variances across multiple groups
☐
(D) Paired data from the same group

This non-parametric test compares two independent groups based on rank-order, suitable when data do not follow a normal distribution.

### #37. Q37. Which statement is incorrect about standard deviation (SD)?

☐
(A) It measures data spread around the mean
☐
(B) A larger SD indicates greater variability
☐
(C) SD is robust to outliers
☐
(D) SD is the square root of the variance

SD is sensitive to outliers, which can significantly affect its value.

### #38. Q38. Chi–square tests are commonly used for:

☐
(A) Comparing means of continuous variables
☐
(B) Examining associations between two categorical variables in contingency tables
☐
(C) Analyzing correlation coefficients
☐
(D) Predicting continuous outcomes

Chi–square tests analyze categorical data to determine if there is a significant association between variables.

### #39. Q39. Degrees of freedom in statistical tests generally indicate:

☐
(A) The number of values in the final calculation that are free to vary
☐
(B) The total sample size
☐
(C) The number of missing data points
☐
(D) The reliability of the data

Degrees of freedom represent the number of independent values that can vary in a calculation without breaking any constraints.

### #40. Q40. If the data is heavily skewed to the right, which measure of central tendency is typically preferred?

☐
(A) Mean
☐
(B) Median
☐
(C) Mode
☐
(D) Range

The median is less affected by extreme values and is preferred in skewed distributions.

### #41. Q41. Match the following:
**1. Parametric test;**
**2. Non-parametric test;**
**3. One-sample t-test;**
**4. Kruskal-Wallis test**

**(a) Compares the median among more than two groups (a rank-based alternative to ANOVA)**
**(b) Assumes normal distribution, using means and standard deviations**
**(c) Compares a single sample mean against a hypothesized value**
**(d) Distribution-free test often using ranks**

☐
(A) 1–b, 2–d, 3–c, 4–a
☐
(B) 1–d, 2–b, 3–a, 4–c
☐
(C) 1–c, 2–b, 3–d, 4–a
☐
(D) 1–a, 2–c, 3–b, 4–d

Parametric tests assume a normal distribution (b), non-parametric tests are distribution-free (d), the one-sample t-test compares a sample mean with a hypothesized value (c), and the Kruskal-Wallis test compares medians across groups (a).

### #42. Q42. (Fill in the blank with multiple-choice) A _____ distribution typically describes the sum of independent Bernoulli trials—for example, the number of successes in n trials.

☐
(A) Poisson
☐
(B) Exponential

☐
(C) Normal
☐
(D) Binomial

The binomial distribution models the number of successes in a fixed number of independent Bernoulli trials.

### #43. Q43. “Metadata” in big data analysis commonly includes:

☐
(A) Detailed raw measurements only
☐
(B) Contextual details such as creation date, experimental conditions, and sample descriptions
☐
(C) Irrelevant information
☐
(D) The raw experimental measurements only

Metadata comprises additional information that provides context for the primary data.

### #44. Q44. In matrix algebra, ‘eigenvalues’ are best described as:

☐
(A) Scalars λ for which there is a non–zero vector v satisfying Av = λv
☐
(B) The sum of the elements in a matrix
☐
(C) The diagonal elements of the identity matrix
☐
(D) The inverse of a matrix

Eigenvalues are defined by the equation Av = λv, where v is a non-zero eigenvector.

### #45. Q45. The primary purpose of PCA (Principal Component Analysis) in multi–dimensional data is:

☐
(A) To increase the dimensionality of the dataset
☐
(B) Reducing the dimensionality of the dataset while retaining as much variance as possible
☐
(C) To standardize data to have zero mean
☐
(D) To calculate simple averages of multiple variables

PCA reduces the number of variables while preserving the most variance in the data.

### #46. Q46. (Fill in the blank) _____ is the method used to factor a rectangular matrix M into U·∑·V$^T$, which is widely employed for dimensionality reduction in big data.

☐
(A) QR decomposition
☐
(B) LU decomposition
☐
(C) Cholesky decomposition
☐
(D) SVD (Singular Value Decomposition)

SVD decomposes a matrix into singular values and vectors, facilitating dimensionality reduction.

### #47. Q47. Exploratory data analysis (EDA) typically includes:

☐
(A) Ignoring data summaries completely
☐
(B) Generating summary statistics, histograms, box plots, and scatter plots to identify patterns and outliers
☐
(C) Immediately building predictive models
☐
(D) Discarding any data that do not match expectations

EDA involves a preliminary analysis of data through visualizations and summary statistics to understand its structure.

### #48. Q48. A ‘z-score’ is used to:

☐
(A) Standardize an individual value by subtracting the mean and dividing by the standard deviation
☐
(B) Calculate the range of the dataset
☐
(C) Determine the median position
☐
(D) Estimate the mode from the data

A z-score standardizes data, showing how many standard deviations an observation is from the mean.

### #49. Q49. Which statement is incorrect about inferential statistics?

☐
(A) They allow conclusions about a population based on sample data
☐
(B) Confidence intervals and hypothesis tests are commonly used
☐
(C) They always produce a zero margin of error
☐
(D) They often require assumptions about the data distribution

Inferential statistics always involve some level of uncertainty; a zero margin of error is not possible.

### #50. Q50. If the correlation coefficient between two continuous variables is r = –0.85, it implies:

☐
(A) No linear relationship
☐
(B) A strong negative linear relationship, indicating that as one variable increases, the other tends to decrease
☐
(C) A strong positive linear relationship
☐
(D) A weak relationship with high variability

A correlation coefficient of –0.85 indicates a strong inverse linear relationship between the variables.

Previous
Submit