EndSem Exam: 2024
Course: Biostatistics
Instructor: Gaurav Ahuja
Duration: 1 hour and 30 minutes
Special Instruction: Don't use computers/calculators
Maximum Marks: 50

Perform any two questions from question 1 to question 5 [Each question 5 marks]

Question 1: Explain the concept of skewness in data distribution and how it influences measures of central tendency. Provide an example to illustrate your explanation.

Solution : Skewness in data distribution refers to the asymmetry of the distribution around its mean.

- Positively skewed distributions have a longer right tail, pulling the mean to the right of the median.
- Negatively skewed distributions have a longer left tail, pulling the mean to the left of the median.

It influences measures of central tendency:

Mean is affected by extreme values, being pulled towards the skew.

Median is less affected, providing a better representation of the central value.

Example: In a dataset of ages {18, 19, 20, 21, 22, 23, 24, 25, 60}, the mean is influenced by the outlier 60, making it higher, indicating positive skewness

Question 2: How does the geometric mean differ from the arithmetic mean, and in what scenarios would you choose one over the other for summarizing data?

Answer : The arithmetic mean is calculated by summing up all values and dividing by the number of values, suitable for additive data like heights or scores. The geometric mean, calculated by taking the nth root of the product of all values, is preferred for multiplicative data like growth rates or investment returns. Use arithmetic mean for additive data and geometric mean for multiplicative data. Arithmetic mean when dealing with data that are additive in nature, such as heights, weights, temperatures, and scores.

Question 3: Briefly explain how you would interpret a QQ plot to identify departures from a theoretical distribution and discuss the implications of such findings on data analysis strategies.

Answer : A QQ plot (Quantile-Quantile plot) is used to assess whether a dataset follows a particular theoretical distribution, like the normal distribution. In a QQ plot, if the data points fall approximately along a diagonal line, it suggests that the dataset is normally distributed. Departures from this line indicate deviations from the theoretical distribution.

If the points deviate upwards, it suggests heavier tails than expected.

If the points deviate downwards, it suggests lighter tails than expected.

Identifying departures from the theoretical distribution can impact data analysis strategies:

- **Adjusting statistical tests**: If the data deviates significantly, it may warrant the use of non-parametric tests or transformations to meet the assumptions of the statistical test.
- **Choosing appropriate models**: If the data departs from normality, it might be necessary to use models better suited for non-normal distributions.
- **Understanding underlying phenomena**: Departures can indicate underlying processes in the data that require further investigation or different modeling approaches.

Question 4: Explain the nuanced difference between standard deviation and standard error of the mean, highlighting their respective roles in statistical inference.

Answer : The standard deviation (SD) measures the dispersion or spread of individual data points in a dataset around the mean. It quantifies the variability within the dataset itself.

On the other hand, the standard error of the mean (SEM) measures the variability of the sample mean from multiple samples of the same population. It indicates how much the sample mean is likely to vary from the true population mean.

In statistical inference:

Standard deviation (SD) is used to describe the variability within a single sample or population.

Standard error of the mean (SEM) is used to estimate the variability of the sample mean and to calculate confidence intervals or conduct hypothesis testing about the population mean. It accounts for the uncertainty in the sample mean due to sampling variability.

In summary, while standard deviation describes the spread of data within a single sample, standard error of the mean estimates the variability of the sample mean across multiple samples and aids in making inferences about the population mean.

Question 5: What unique insights can be gleaned from interpreting a box plot compared to other descriptive statistical techniques?

Answer : A box plot provides several unique insights compared to other descriptive statistical techniques:

- **Visualizing distribution:** It gives a clear visual representation of the distribution of the data, including the median, quartiles, and potential outliers.
- **Identification of outliers:** Outliers are easily identified as individual data points beyond the whiskers of the box plot.

- **Assessment of skewness and symmetry:** The length and direction of the whiskers provide information about the symmetry and skewness of the data distribution.
- **Comparison of multiple groups:** Box plots allow for easy comparison of multiple groups or datasets, showing differences in central tendency, spread, and variability.

Perform any two questions from <u>question 6 to question 10 [Each question 5 marks]</u>

Question 6: Discuss a scenario where using a simple random sample may lead to different conclusions compared to using a stratified or cluster sampling approach.

Answer : In a scenario where a population consists of distinct subgroups with varying characteristics, using a simple random sample may lead to different conclusions compared to using a stratified or cluster sampling approach.For example, consider a university with undergraduate students categorized into different majors such as engineering, humanities, and natural sciences. If we use a simple random sample to select students for a survey on academic performance, there's a chance that certain majors might be overrepresented or underrepresented in the sample. This could skew the results and lead to biased conclusions, especially if one major has significantly different academic performance compared to others.

However, by using a stratified sampling approach, where students are sampled proportionally from each major, or a cluster sampling approach, where entire majors are sampled as clusters, we can ensure that each subgroup is adequately represented in the sample. This allows for more accurate comparisons between groups and reduces the risk of drawing misleading conclusions based on the sampled data.

Question 7: Mention and briefly discuss the types of data transformations.

Answer : **1. Constructive:** The data transformation process adds, copies, or replicates data.

Subtypes

1. Log Transformation:
   - Parameter: Base of the logarithm (e.g., natural logarithm, base-10 logarithm).
   - Purpose: Reduces the impact of extreme values and stabilizes variance.

2. Square Root Transformation:
   - Parameter: None.
   - Purpose: Stabilizes variance and reduces the impact of extreme values.

3. Box-Cox Transformation:
   - Parameter: Lambda (λ) parameter.
   - Purpose: Generalization of power transformations, stabilizes variance, and handles different types of transformations.

4. Reciprocal Transformation:
   - Parameter: None.
   - Purpose: Inverts the values, useful for data with reciprocal relationships.
5. Exponential Transformation:
   - Parameter: Power or rate of growth.
   - Purpose: Amplifies the differences between small values, useful for data with exponential relationships.
6. Square Transformation:
   - Parameter: None.
   - Purpose: Amplifies the differences between large values, useful for data with quadratic relationships.
7. Inverse Hyperbolic Sine (ASinh) Transformation:
   - Parameter: None.
   - Purpose: Stabilizes variance, particularly for count data or percentages.
8. Rank Transformation:
   - Parameter: None.
   - Purpose: Transforms the data into ranks, useful for non-parametric analyses and handling non-normal distributions.
9. Winsorizing:
   - Parameter: Trimming percentage or threshold values.
   - Purpose: Reduces the impact of outliers by capping extreme values at a specified threshold.
10. Centering and Scaling:
    - Parameter: Mean, standard deviation, or other scaling factors.
    - Purpose: Centers the data around a specific value and scales it to a desired range.
11. Winsorized Standardization:
    - Parameter: Trimming percentage or threshold values.
    - Purpose: Combines winsorizing and standardization to handle outliers.
12. Truncate Transformation:
    - Parameter: Truncation points.
    - Purpose: Cuts off extreme values beyond specified thresholds.

2. Destructive: The system deletes fields or records.

3. Aesthetic: The transformation standardizes the data to meet requirements or parameters.

4.Structural: The database is reorganized by renaming, moving, or combining columns.

Question 8: Explain the paradoxical nature of the Central Limit Theorem (CLT) and its implications for real-world data analysis.

Answer : The Central Limit Theorem (CLT) presents a paradoxical yet powerful concept in statistics by stating that the distribution of sample means tends towards normality as the sample size increases, irrespective of the shape of the population distribution. This implies that even if the underlying population is not normally distributed, with a sufficiently large sample size, the distribution of sample means will approximate a normal distribution. In practical terms, this theorem offers several implications for real-world data analysis. Firstly, it ensures the robustness of statistical methods reliant on the normal distribution assumption, such as hypothesis testing and confidence intervals, even when dealing with non-normally distributed populations. Additionally, understanding the CLT aids in determining appropriate sample sizes for studies, as larger samples lead to sample means that closely follow a normal distribution. Moreover, in cases where the population distribution is highly skewed or non-normal, data transformation techniques can be employed to make the distribution of sample means more normal, thus enhancing the validity of statistical analyses. Overall, the CLT serves as a cornerstone of statistical inference, allowing researchers to generalize findings from sample data to the population, even under non-ideal conditions.

Question 9: Provide a concise example where understanding conditional probability alters the perception of likelihood, highlighting its importance in decision-making and risk assessment.

Answer : Imagine a medical test for a rare disease that is only 1% prevalent in the population. The test has a sensitivity of 95% and a specificity of 90%. At first glance, one might assume that if the test comes back positive, there's a 95% chance of having the disease. However, understanding conditional probability alters this perception.

If we calculate the probability of having the disease given a positive test result using Bayes' theorem, we find that it's about 9%, much lower than the initial intuition. This shift occurs because we are not just considering the probability of a positive test result but also taking into account the low prevalence of the disease in the population.

This example highlights the importance of understanding conditional probability in decision-making and risk assessment. Without considering the conditional probabilities, one might overestimate the likelihood of an event, leading to inappropriate actions or unnecessary anxiety.

Question 10: Discuss a scenario where a model with higher accuracy achieves a lower AUC-ROC score than a model with lower accuracy, and explain the underlying reasons for this discrepancy in performance evaluation.

Answer : Consider a scenario in which you're developing models to predict whether customers will churn (cancel their subscription) based on various features such as age, usage patterns, and satisfaction scores. You've developed two models: Model A and Model B.

Model A has a higher accuracy of 85% compared to Model B, which has an accuracy of 80%. However, when you evaluate the models using the Area Under the Receiver Operating Characteristic curve (AUC-ROC), you find that Model B has a higher AUC-ROC score of 0.85, while Model A has a lower AUC-ROC score of 0.80.

The underlying reasons for this discrepancy lie in the nature of the evaluation metrics:

- **Accuracy:** Accuracy measures the overall correctness of the model's predictions, i.e., the ratio of correctly predicted instances to the total number of instances. However, accuracy alone doesn't consider the balance between true positives, true negatives, false positives, and false negatives.
- **AUC-ROC:** The AUC-ROC score measures the model's ability to discriminate between positive and negative classes across different thresholds. It plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings. A higher AUC-ROC score indicates better discrimination performance, regardless of the threshold chosen.

In this scenario, Model B may have a lower accuracy but achieves a higher AUC-ROC score because it can better rank the positive and negative instances correctly, leading to a higher true positive rate and a lower false positive rate across different threshold settings. On the other hand, Model A, despite having a higher accuracy, may misclassify some instances, leading to a suboptimal discrimination performance when considering the trade-off between sensitivity and specificity.

This highlights the importance of using appropriate evaluation metrics based on the specific goals and requirements of the problem. While accuracy is a commonly used metric, it may not always provide a complete picture of the model's performance, especially in imbalanced or skewed datasets. AUC-ROC, on the other hand, provides a comprehensive assessment of the model's ability to distinguish between classes and is particularly useful when the class distribution is imbalanced or when the costs of false positives and false negatives are different.

Perform any two questions from <u>question 11 to question 13 [Each question 5 marks]</u>

Question 11: In what ways do Type I and Type II error trade-offs challenge the interpretation of hypothesis testing results?

Answer : Type I and Type II errors represent the trade-offs in hypothesis testing:

Type I error: This occurs when we reject a true null hypothesis. It represents a false positive, indicating that we concluded there is an effect or difference when there isn't one.

Type II error: This happens when we fail to reject a false null hypothesis. It represents a false negative, indicating that we missed detecting an effect or difference when there actually is one.

These errors challenge the interpretation of hypothesis testing results because they involve balancing the risks. Minimizing one type of error often increases the risk of the other. For instance, lowering the threshold for rejecting the null hypothesis to reduce the likelihood of Type II errors increases the risk of Type I errors, and vice versa. Thus, interpreting hypothesis testing results requires careful consideration of the potential consequences of both types of errors and the relative costs associated with them.

Question 12: In what ways does power analysis influence sample size estimation, and how can a small sample size yield misleading results despite achieving statistical significance?

Answer : Power analysis influences sample size estimation by determining the minimum sample size required to detect a meaningful effect or difference with a specified level of statistical power. A higher statistical power indicates a greater ability to detect true effects, while a lower power increases the likelihood of Type II errors (false negatives). Therefore, power analysis helps researchers determine the sample size needed to achieve a desired balance between Type I and Type II errors.

Despite achieving statistical significance, a small sample size can yield misleading results due to several factors:

Limited generalizability: Small samples may not adequately represent the population, leading to results that cannot be generalized beyond the sample.

Increased variability: With fewer data points, variability within the sample can have a larger impact on results, potentially leading to inflated effect sizes or misleading estimates of population parameters.

Increased susceptibility to outliers: Outliers can disproportionately influence results in small samples, leading to skewed findings.

Reduced statistical power: Small samples may lack the power to detect true effects, increasing the risk of false negatives (Type II errors) despite achieving statistical significance.

Question 13: In what scenarios would you prefer using an F-test over a T-test, and why?

Answer : An F-test is preferred over a T-test in scenarios involving the comparison of variances or means across more than two groups. Specifically:

Comparison of variances: When comparing the variances of two or more independent samples, such as in analysis of variance (ANOVA), an F-test is used. This allows for assessing whether the variability within groups is significantly different from each other.

Regression analysis: In regression analysis, the F-test is used to assess the overall significance of the model by comparing the variability explained by the regression model to the variability not explained.

Experimental designs with multiple factors or treatments: When conducting experiments with more than two treatment groups or factors, such as in factorial designs, the F-test is used to assess the significance of main effects and interactions.

Compulsory questions: 14 and 15 [Each question 5 marks]

Question 14: How does Kaplan-Meier survival analysis accommodate censored data and provide insights into time-to-event outcomes?

Answer: Kaplan-Meier survival analysis is a statistical method used to analyze time-to-event data, particularly in medical and biological studies where individuals may not experience the event of interest (e.g., death, disease progression) during the study period. This method accommodates censored data by incorporating information from both individuals who experience the event and those who are censored (i.e., the event does not occur within the study period).

The Kaplan-Meier estimator calculates the probability of survival (or the probability of not experiencing the event) at each time point throughout the study. It does this by considering the proportion of individuals still at risk of experiencing the event at each time point, taking into account both the observed events and the censored observations. The survival function is estimated as the product of the conditional probabilities of survival at each event time.

Insights provided by Kaplan-Meier survival analysis include:

Survival probabilities: It estimates the probability of survival over time, allowing for the visualization of survival curves that depict the probability of survival as a function of time.

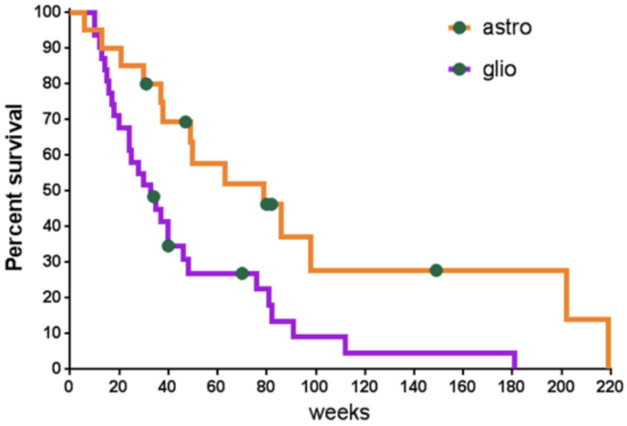
Median survival time: The median survival time represents the time point at which 50% of individuals have experienced the event of interest. It provides a summary measure of the central tendency of time-to-event outcomes.

Comparison between groups: Kaplan-Meier curves can be stratified by different groups (e.g., treatment vs. control) to compare survival probabilities between groups and assess the impact of covariates on time-to-event outcomes.

Hazard rates: While Kaplan-Meier analysis estimates survival probabilities, it does not directly estimate hazard rates (i.e., the instantaneous rate of experiencing the event at a given time). However, hazard rates can be estimated using more advanced survival analysis techniques, such as Cox proportional hazards regression.

Overall, Kaplan-Meier survival analysis provides valuable insights into time-to-event outcomes by accommodating censored data and estimating survival probabilities over time.

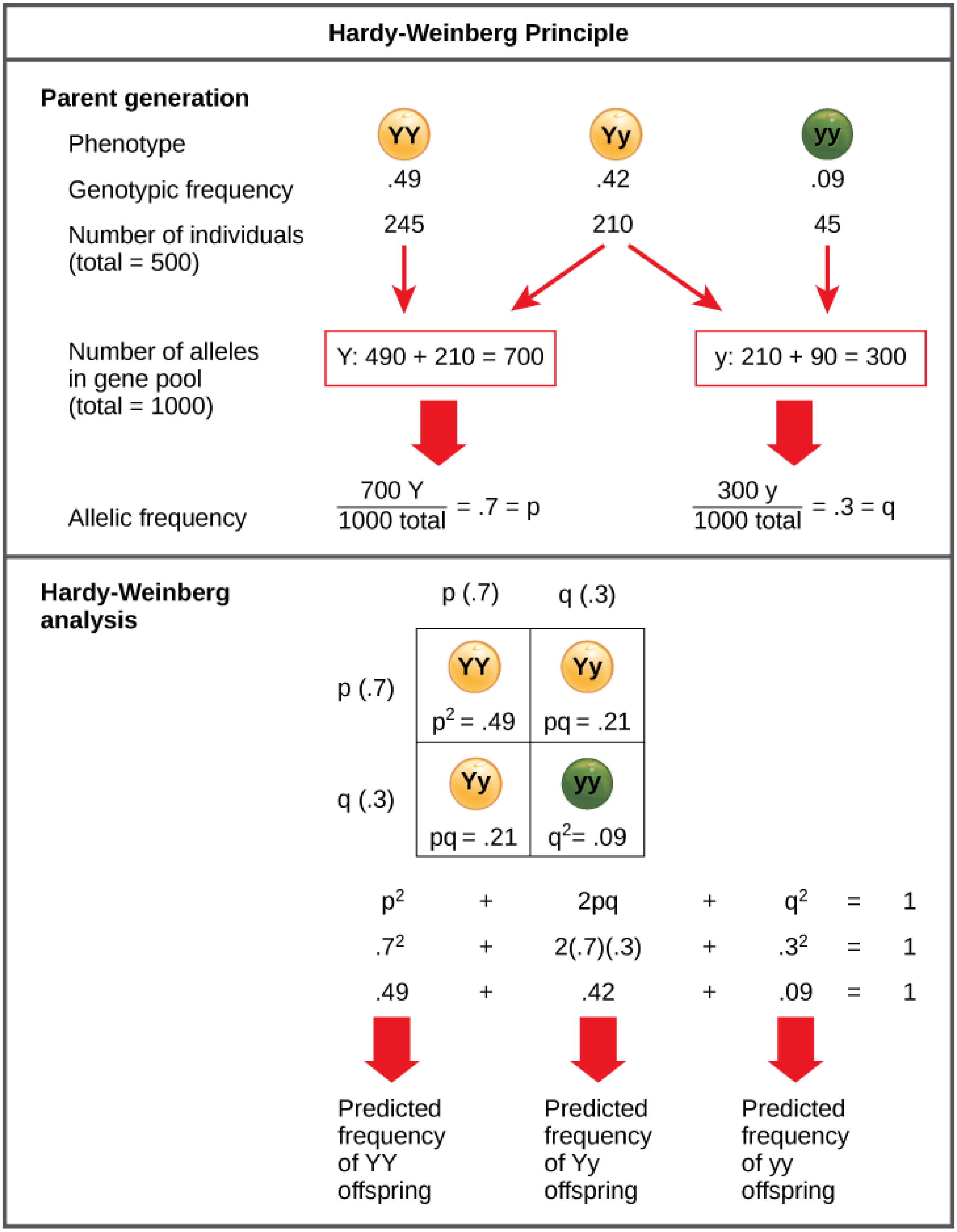
K-M plot of survivor function by tumour type



Question 15: How does Hardy-Weinberg equilibrium serve as a fundamental concept in biostatistics, particularly in population genetics and disease association studies?

Answer : Hardy-Weinberg equilibrium (HWE) is a fundamental concept in biostatistics, particularly in population genetics and disease association studies, because it provides a null hypothesis against which observed genotype frequencies can be compared.

In population genetics, HWE predicts the expected genotype frequencies in a population under specific conditions: random mating, large population size, no migration, no mutation, and no selection pressure. Deviations from HWE indicate factors such as non-random mating, genetic drift, mutation, migration, or natural selection, providing insights into the evolutionary forces shaping genetic variation within populations.In disease association studies, deviations from HWE can indicate potential genotyping errors, population stratification, or genetic association with the disease. Departures from HWE can lead to spurious associations or obscure true associations between genetic variants and diseases. Therefore, checking for HWE is an essential quality control step in genetic association studies to ensure the reliability of the results.Overall, Hardy-Weinberg equilibrium serves as a critical concept in biostatistics, providing a null model for genetic variation within populations and serving as a reference point for assessing genetic associations with diseases.


Hardy-Weinberg Principle
Parent generation
Phenotype
YY
Yy
yy
Genotypic frequency
.49
.42
.09
Number of individuals (total = 500)
245
210
45
Number of alleles in gene pool (total = 1000)
Y: 490 + 210 = 700
y: 210 + 90 = 300
Allelic frequency
$\frac{700\ Y}{1000\ total} = .7 = p$
$\frac{300\ y}{1000\ total} = .3 = q$
Hardy-Weinberg analysis
p (.7)
q (.3)
p (.7)
q (.3)
YY
$p^2 = .49$
Yy
$pq = .21$
Yy
$pq = .21$
yy
$q^2 = .09$
$p^2 + 2pq + q^2 = 1$
$.7^2 + 2(.7)(.3) + .3^2 = 1$
$.49 + .42 + .09 = 1$
Predicted frequency of YY offspring
Predicted frequency of Yy offspring
Predicted frequency of yy offspring

Compulsory MCQs [Each question 5 marks]

MCQ1: Which of the following statements best describes the Shapiro-Wilk test?

- It is used to assess the normality of data distribution by comparing sample skewness and kurtosis to theoretical values.
- It is a non-parametric test used to compare the means of two independent groups with unequal variances.
- **It evaluates whether a dataset follows a normal distribution by testing the null hypothesis that the sample comes from a normally distributed population.**
- It is primarily employed to determine the homogeneity of variances in multiple groups before conducting an analysis of variance (ANOVA).

MCQ2: Which of the following statements best captures the difference between correlation and regression analysis?

- **Correlation measures the strength and direction of the relationship between two variables, while regression predicts the value of one variable based on the value of another.**
- Correlation determines causation between variables, whereas regression assesses the degree of association between them.
- Correlation is a parametric statistical method, whereas regression is non-parametric.
- Correlation analysis only applies to categorical variables, while regression analysis is suited for continuous variables.