Univariate Statistics Problems

U.1. A dataset of $n = 12$ scores is $10, 15, 12, 18, 10, 20, 15, 12, 10, 15, 25, x$. If the **Arithmetic Mean** is 15, find the value of x. Then, calculate the **Median** of the complete dataset and interpret the difference between the Mean and Median.

U.2. **(Advanced Property Proof)** Prove that for any two positive numbers a and b, the relationship between the Arithmetic Mean (AM), Geometric Mean (GM), and Harmonic Mean (HM) is $GM^2 = AM \times HM$. Use this result to explain why, for a dataset with non-identical values, $AM > GM > HM$.

U.3. A variable X has observations $x_i > 0$. If $AM(X) = 10$ and $HM(X) = 6.4$. Calculate the Geometric Mean $GM(X)$. If a new variable $Y = \frac{1}{X}$ is created, calculate $HM(Y)$ and $AM(Y)$.

U.4. For a frequency distribution of income data, the mean is $\overline{x} = 1500$, the median is 1800, and the standard deviation is 300. Which skewness measure (Pearson's 1 or 2) is appropriate here? Calculate the measure and interpret the distribution's shape in terms of Mean-Median-Mode relationship.

U.5. The mode of a grouped distribution with equal class intervals is calculated using the formula. Explain the assumption that this formula makes about the **frequency distribution within the modal class**. Why can the median formula be considered more robust to deviations from this assumption?

U.6. Consider a dataset transformed by $Y_i = aX_i + b$. Prove that the Weighted Mean $\overline{y}_w$ is related to $\overline{x}_w$ by $\overline{y}_w = a\overline{x}_w + b$.

U.7. A dataset is highly negatively skewed. If the Median is 50 and the Standard Deviation is 10. Estimate the Mode using Pearson's second empirical relationship, assuming a Mean of 45.

U.8. Use **Sturges' Rule** to determine the optimal number of classes (k) for a dataset with $n = 2500$ observations. If the range of the data is $R = 120$, what is the approximate **class width** (c)?

U.9. **(Conceptual Synthesis)** Describe the geometric differences between a **Frequency Polygon** and an **Ogive** in terms of the axes used, the data points plotted, and the type of frequency (simple vs. cumulative) they represent.

U.10. In a portfolio of four assets, the returns over a year were $R_1 = 1.10$, $R_2 = 1.20$, $R_3 = 0.90$, $R_4 = 1.05$. Calculate the **Geometric Mean Return** and explain why the Arithmetic Mean would be an inappropriate measure of the average annual growth rate.

U.11. A grouped distribution has class intervals of size $c = 5$. The mean is calculated to be $\overline{x} = 45$ using an **Assumed Mean** $A = 40$. The calculated mean of the coded variable is $\overline{u} = 1$. Use the coding formula to verify the calculated mean $\overline{x}$.

U.12. **(Data Synthesis)** For the dataset $2, 4, 6, 8, 10, 12, 14, 16, 18, 20$: Calculate Q_1, Q_3, and the **Interquartile Range (IQR)**. If an observation $x_{outlier} = 50$ is added, how much does the Mean change, and how much does the IQR change?

U.13. **(Advanced Property Proof)** For any dataset x_i, prove the property $\sum_{i=1}^{n}(x_i - k)^2$ is minimized when $k = \overline{x}$. Use this result to explain why the Mean is considered the center of gravity of the distribution.

U.14. The Median of a dataset is 15. The **Median Absolute Deviation (MAD)** is 4. Construct a confidence interval using the approximate relationship: $P(\text{Median} - 1.4826 \times \text{MAD} \leq X \leq \text{Median} + 1.4826 \times \text{MAD}) \approx 0.50$.

U.15. A dataset on product life (in hours) is grouped with unequal intervals: $0 - 50 (f = 5), 50 - 100 (f = 10), 100 - 200 (f = 15)$. Calculate the **Frequency Density** for the $100 - 200$ class and explain why this is necessary for an accurate **Histogram**.

U.16. A dataset has $s^2 = 100$. If every observation is multiplied by -2 ($Y_i = -2X_i$), what is the new **Standard Deviation** (s_y)? Prove the general property relating s_y to s_x.

U.17. **(Conceptual Synthesis)** Explain the trade-off between the **Mean Deviation (MD)** and the **Standard Deviation (s)** as measures of dispersion. Specifically, detail the mathematical advantage of s (due to calculus) and the conceptual advantage of MD (due to its definition).

U.18. A financial analyst compares two stocks. Stock A: $\overline{x} = 200, s = 20$. Stock B: $\overline{x} = 50, s = 8$. Calculate the **Coefficient of Variation (CV)** for both and conclude which stock exhibits greater **relative risk** (variability).

U.19. **(Chebyshev's vs. Empirical Rule)** A population of test scores has $\mu = 60$ and $\sigma = 8$. a) Use **Chebyshev's Rule** to find the minimum percentage of scores between 44 and 76. b) If the distribution is known to be perfectly Normal, use the **Empirical Rule** to state the expected percentage in the same interval. c) Explain why the result from part (a) is always less than or equal to the result from part (b).

U.20. Prove the algebraic property that the **r^{th} Central Moment (μ_r)** is invariant to a change in origin (i.e., if $Y_i = X_i + a$, then $\mu_r(Y) = \mu_r(X)$).

U.21. A grouped distribution has the following central moments: $\mu_2 = 20$,

$\mu_4 = 780$, and class width $c = 10$. Apply **Sheppard's Correction** to find the corrected μ_2 and μ_4.

U.22. Given a dataset's raw moments about the origin: $\mu'_1 = 10$, $\mu'_2 = 110$, $\mu'_3 = 1250$. Calculate the first three **Central Moments** (μ_1, μ_2, μ_3).

U.23. **(Advanced Application)** Using the Central Moments from U.22, calculate the **Moment-based Coefficient of Skewness (γ_1)**. Is the distribution positively or negatively skewed?

U.24. Define a value that falls into the region between the **Upper Inner Fence (UIF)** and the **Upper Outer Fence (UOF)**. Explain why these two fences exist for classifying outliers.

U.25. The **3^{rd} Central Moment** of a variable X is $\mu_3(x) = 10$. If $Y = -2X$, calculate $\mu_3(y)$.

U.26. **(Z-score and Covariance)** A student transforms their dataset into **Z-scores**. What is the Mean and Standard Deviation of the resulting Z-score variable (Z)? If a linear relationship is found between Z_x and Z_y, how does the covariance $s_{z_x z_y}$ relate to the correlation coefficient r_{xy}?

U.27. The sample variance is $s^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$. Explain why the denominator $n-1$ (the **Degrees of Freedom**) is used, rather than n, in the context of estimating the population variance (σ^2).

U.28. Given the central moments: $\mu_2 = 9$ and $\mu_4 = 100$. Calculate the **Excess Kurtosis (γ_2)**. Classify the distribution as **Leptokurtic, Platykurtic, or Mesokurtic** and describe its shape relative to the normal distribution.

U.29. **(Combined Skewness Interpretation)** A distribution has $\overline{x} = 100$, Median$= 98$, $Q_1 = 80$, $Q_3 = 115$, and $s = 15$. Calculate **Pearson's Second** (SK_2) and **Bowley's** (SK) coefficients. Reconcile the results, noting why they might differ slightly.

U.30. A dataset is generated in a Software environment. Write a conceptual instruction for the Software package to calculate the **Quantile-based Coefficient of Kurtosis (K)** given the Quartile Deviation $(QD = 10)$ and the difference $P_{90} - P_{10} = 60$.

U.31. **(Combined Correction and Shape)** A large grouped dataset $(c = 5)$ yields the uncorrected central moments $\mu_2 = 25$, $\mu_4 = 1000$. After applying Sheppard's correction, calculate the final β_2 and state the shape (Leptokurtic/Platykurtic).

U.32. **(Raw Moment Calculation)** Calculate the 3^{rd} Raw Moment (μ'_3) about the origin for the sample data: $2, 3, 5, 8, 12$.

U.33. **(Conceptual Distinction)** Distinguish between the **Ratio** and

Interval scales of measurement by explaining the mathematical significance of the **zero point** in each. Give a statistical measure that is meaningful only for Ratio data.

U.34. A financial time series has $\overline{x} = 1.0$, $s = 0.05$, and $Q_1 = 0.95, Q_3 = 1.05$. Calculate the **Coefficient of Quartile Deviation** and compare it with the **Coefficient of Variation (CV)**.

U.35. **(Skewness Formula Equivalence)** Prove the algebraic equivalence of the two formulas for Pearson's second coefficient of skewness when the distribution is unimodal and moderately skewed: $SK_2 = \frac{3(\overline{x} - \text{Median})}{s}$.

U.36. If the mean of a grouped dataset is 35, and $n = 50$. What is the value of $\sum f_i x_i$? If the 4^{th} class is $30 - 40$ with $f_4 = 15$, what is the contribution of this class to the sum of squared deviations from the mean?

U.37. For a grouped frequency distribution, explain why the Median is derived by interpolation within the median class, whereas the Mean is calculated using the midpoint (class mark) approximation.

U.38. **(Moments to Properties)** Given $\mu'_1 = 5$, $\mu'_2 = 25$, $\mu'_3 = 140$. Calculate μ_3. Then, calculate γ_1. What is the expected Mean-Median-Mode ordering?

U.39. **(Outlier Detection Logic)** A dataset has $Q_1 = 50$ and $Q_3 = 80$. a) Calculate the **Upper Inner Fence (UIF)**. b) Calculate the **Lower Outer Fence (LOF)**. c) A value $x = 135$ is observed. Is it a mild or extreme outlier?

U.40. **(Data Density)** A dataset has $n = 1000$ and $R = 100$. An analyst uses the 2^k rule, resulting in $k = 10$ classes. Calculate the average number of observations per class interval.

U.41. **(Moment Definition)** Write the mathematical definition for the r^{th} **Raw Moment** (μ'_r) about the origin and the r^{th} **Central Moment** (μ_r) about the mean.

U.42. If a sample has a Mean $\overline{x} = 20$ and $s = 5$. What is the minimum and maximum Z-score of any observation that is *not* considered an outlier under the assumption that an outlier is beyond $3s$ from the mean?

U.43. **(Conceptual Application)** A dataset of stock prices shows an extremely high positive kurtosis ($\gamma_2 \gg 0$). Explain the implication for risk management compared to a normally distributed stock.

U.44. **(Chebyshev vs. Empirical - Conceptual)** What is the critical distinction between the two rules regarding the shape of the underlying probability distribution required for their application?

U.45. **(Data Scale Limitation)** Why is the Mode the *only* measure of

central tendency that is mathematically meaningful for **Nominal** data?

U.46. **(Algebraic Proof)** Prove the property $\mu_2 = \mu'_2 - (\mu'_1)^2$, which relates the Variance to the first two raw moments.

U.47. If a dataset is symmetric, the 1^{st} and 3^{rd} quartiles are equidistant from the median. If $Q_1 = 40$ and Median= 60, what must Q_3 be? What is the expected value of Bowley's coefficient?

U.48. **(Stem-and-Leaf vs. Histogram)** When is a **Stem-and-leaf plot** preferred over a **Histogram** for data visualization?

U.49. **(Advanced Transformation)** A variable X has $\overline{x} = 10$ and $s = 3$. If $Y = 5 - \frac{1}{2}X$, calculate $\overline{y}$, s_y^2, and CV_y.

U.50. **(Combined Indices)** Given a distribution with $Q_1 = 10, Q_3 = 40, P_{10} = 5, P_{90} = 50$. Calculate **Bowley's SK** and the **Quantile-based Kurtosis K**. Interpret both results.

U.51. **(Property Application)** Given $\sum(x_i - 10)^2 = 200$ and $\sum(x_i - 20)^2 = 300$ for $n = 10$. Find the Mean $\overline{x}$ and the Variance s^2.

U.52. **(Mixed Moments and Skewness)** Given $\mu_2 = 16$ and $\mu_3 = -128$. Calculate γ_1 and interpret the nature of the skewness. Which coefficient of skewness will be most reliable here?

U.53. **(Decile Interpolation)** For a grouped dataset of $N = 100$, the 7^{th} Decile (D_7) class is $50 - 60$. $L = 50, F = 60, f = 20, c = 10$. Calculate D_7.

U.54. **(Bimodal Interpretation)** A dataset of customer ages shows a bimodal distribution. Provide a plausible real-world explanation for this phenomenon and suggest which measure of central tendency (Mean, Median, or Mode) is least informative in this scenario.

U.55. **(Geometric Mean Limitation)** Why is the Geometric Mean mathematically undefined or highly problematic to calculate if the dataset contains zero or negative values?

U.56. **(Mean of Reciprocals)** Prove that the Harmonic Mean is the reciprocal of the Arithmetic Mean of the reciprocals of the observations: $HM = \frac{1}{\overline{x}_{1/x}}$.

U.57. **(Z-score Interpretation)** In a university class, scores are $\overline{x} = 75, s = 10$. A student scored 95. Another student is in a different class with $\overline{x} = 60, s = 5$. What score would the second student need to achieve the same relative performance (Z-score)?

U.58. **(Sheppard's Correction Logic)** Explain why μ_1 and μ_3 do not require Sheppard's correction, while μ_2 and μ_4 do.

U.59. **(Quantile Interpretation)** The P_{99} for house prices in a city is \$1.5

million. What does this indicate about the distribution of prices?

U.60. **(Outlier Removal Effect)** A dataset has $n = 10$ and $s = 5$. An extreme outlier ($4s$ above the mean) is removed. Qualitatively describe the expected change in the value of s and γ_1.

U.61. **(Conceptual) ** Explain the difference between **Discrete** and **Continuous** quantitative variables, and give an appropriate graphical display for each (excluding tables).

U.62. **(Kelley's Skewness)** Given $D_1 = 10$, $D_5 = 20$, $D_9 = 35$. Calculate Kelley's Coefficient of Skewness (SK) and interpret the result.

U.63. **(Median and Change of Scale)** If a dataset's median is $M_x = 40$. If the data is transformed by $Y_i = 10 - 5X_i$, what is the new median M_y?

U.64. **(Moment Generating Function - Conceptual)** State the definition of a moment generating function (MGF) $M_x(t)$. Explain how the r^{th} Raw Moment is derived from the MGF.

U.65. **(Raw Moments to Skewness)** Given $\mu'_1 = 0, \mu'_2 = 1, \mu'_3 = 0, \mu'_4 = 3$. What is the distribution? (Hint: Consider the Z-score transformation properties).

U.66. **(Kurtosis Application)** A hedge fund manager notes that the daily returns of his portfolio have $\gamma_2 > 0$. What is the practical implication of this **positive excess kurtosis** (Leptokurtosis) for setting capital reserve requirements?

U.67. **(Data Visualization Limitation)** Why is a **Pie Chart** considered a poor choice for visualizing frequency distribution when the number of categories exceeds five?

U.68. **(Composite Bar Chart)** Explain the difference between a **Stacked Bar Chart** and a **Clustered (or Multiple) Bar Chart** in representing multivariate data.

U.69. **(Algebraic Proof - Mean Invariance)** Given a dataset x_i. If a value c is added to every observation, prove that the Mean shifts by c.

U.70. **(Range and IQR)** Given an ordered dataset of n points. If the Range is $R = 100$ and $IQR = 10$. What does this imply about the data concentration in the central 50% versus the tails?

U.71. **(Five-Number Summary)** Construct a hypothetical dataset (N=10) that has a median of 15, $Q_1 = 10$, $Q_3 = 20$, and a range of 30.

U.72. **(Outlier Classification)** A value falls exactly on the **Upper Inner Fence (UIF)**. Is it classified as an outlier? Justify your answer based on

the strict inequality definitions of the fences.

U.73. **(Software Instruction)** Write a conceptual instruction for a Software package to calculate the 95^{th} percentile (P_{95}) and the **5^{th} Decile (D_5) for a dataset of $N = 5000$ non-negative observations.

U.74. **(Data Type and Scale)** Classify the variable "Time (in seconds) taken to complete a task" based on its data type (Quantitative/Qualitative) and measurement scale (Nominal/Ordinal/Interval/Ratio).

U.75. **(Moment Calculation)** Given a grouped distribution with $x_i : 5, 10, 15$ and $f_i : 3, 5, 2$. Calculate the 2^{nd} Central Moment (μ_2).

U.76. **(Change of Scale Effect on CV)** If a variable X has $CV_x = 20\%$. If $Y = 3X$. Calculate CV_y.

U.77. **(Bowley's vs. Pearson's)** Explain one circumstance (related to data shape) in which **Bowley's coefficient** would be a more reliable measure of skewness than **Pearson's coefficient**.

U.78. **(Property of AM, GM, HM)** Construct a simple dataset of $n = 3$ non-identical values where AM, GM, HM can be calculated. Verify the inequality $AM > GM > HM$.

U.79. **(Interpretation of μ_3)** If the 3^{rd} central moment μ_3 is a large positive value, what two key pieces of information does this provide about the distribution's shape?

U.80. **(Software Instruction - Outlier)** Write a conceptual instruction for a Software package to generate a **Box-Whisker Plot** and identify all observations that lie beyond the **Lower Inner Fence (LIF)**.

U.81. **(Moment Correction Logic)** In a grouped dataset with $c = 5$. If the uncorrected $\mu_2 = 20$, calculate the correction term for the variance.

U.82. **(Z-score Transformation Proof)** Prove the property that the standard deviation of the Z-score variable is $s_z = 1$.

U.83. **(Combined Skewness and Kurtosis)** A distribution has $SK_2 = 0$ and $\gamma_2 = 0.5$. Describe the overall shape (symmetry, peakedness, and tail weight).

U.84. **(Median Formula Breakdown)** The median formula for grouped data uses F (cumulative frequency before the median class). Explain why an extremely small f_{med} (frequency of the median class) can lead to a calculated median value far from the center of the median class.

U.85. **(Algebraic Proof - Sum of Squares)** Prove that $\sum(x_i - k)^2 = \sum(x_i - \overline{x})^2 + n(\overline{x} - k)^2$.

U.86. **(Moment Limitation)** Which moment is required to calculate the Standard Deviation? Which moment is required to calculate γ_1?

U.87. **(Decile Calculation)** A dataset has $N = 51$. Find the index position of the 20^{th} Percentile (P_{20}) and the 8^{th} Decile (D_8).

U.88. **(Conceptual Synthesis)** A dataset's mean is 10. $\sum x_i = 50$. $\sum (x_i - 10)^2 = 100$. Find n and s^2.

U.89. **(Coefficient of Range)** Define the **Coefficient of Range** and explain why it is a measure of relative dispersion.

U.90. **(Interpretation of $AM = GM = HM$)** What does the equality of the three means ($AM = GM = HM$) imply about the dataset?

U.91. **(Software Instruction - Moments)** Write a conceptual instruction for a Software package to calculate the **4^{th} Central Moment (μ_4)** of a variable X and test its significance against the normal distribution value.

U.92. **(Raw Moments and Mean)** Given $\mu'_1 = 5$, $\mu'_2 = 28$. Find the Mean and Standard Deviation.

U.93. **(Transformation Effect on Kurtosis)** If $Y = aX + b$. Prove that $\gamma_2(Y) = \gamma_2(X)$.

U.94. **(Unequal Class Intervals)** A grouped distribution has a class width ratio of $1 : 2 : 3$. If the first class frequency is $f_1 = 10$, what frequency densities must the second and third classes have to maintain the same vertical height on the histogram?

U.95. **(Conceptual - Data Ordering)** What property of the Median makes it a better measure of central tendency than the Mean in the presence of extreme outliers?

U.96. **(Kurtosis Interpretation)** Which distribution (Leptokurtic, Mesokurtic, Platykurtic) has thinner tails and a flatter peak than the normal distribution?

U.97. **(Weighted Mean Application)** A student scores 80% on a 5-credit course and 90% on a 3-credit course. Calculate the overall **Weighted Mean** grade.

U.98. **(Dispersion Measure Choice)** When comparing the variability of two datasets that are in different units (e.g., dollars and kilograms), which measure of dispersion is mandatory?

U.99. **(Skewness Interpretation)** If Bowley's coefficient is exactly -1, what must be true about the positions of Q_1, Q_2, and Q_3?

U.100. **(Software Instruction - Histogram)** Write a conceptual instruction for a

Software package to generate a **Histogram** using the optimal number of classes determined by **Sturges' Rule**.

U.101. **(Conceptual)** The sum of the absolute deviations from the Mean is 100. What can be concluded about the minimum possible sum of absolute deviations from any constant k?

U.102. **(Moment Property)** If a distribution is perfectly symmetric, what must be the value of all its **odd-ordered central moments** ($\mu_1, \mu_3, \mu_5, \dots$)?

U.103. **(Conceptual)** Define the $\frac{k(n+1)}{q}$ method for finding the position of quantiles (Q_k, D_k, P_k).

U.104. **(Kurtosis Range)** What is the minimum possible value for the **Moment-based Coefficient of Kurtosis (β_2)**?

U.105. **(Kurtosis Correction)** A grouped dataset with $c = 10$ has an uncorrected $\mu_2 = 50$ and $\mu_4 = 8000$. Apply Sheppard's correction to μ_4.

U.106. **(Property of Median)** Explain why the Median minimizes the sum of absolute deviations $\sum |x_i - k|$.

U.107. **(Conceptual)** A dataset of test scores has a range of 50. If the lowest score is 30, what is the highest possible score?

U.108. **(Change of Origin on Moments)** Explain how a change of origin ($Y = X + a$) affects μ'_2 and μ_2.

U.109. **(Data Synthesis)** Construct a simple dataset ($N = 5$) where the Mean and Median are equal, but the Mode is different.

U.110. **(Software Instruction - γ_2)** Write a conceptual instruction for a Software package to calculate the **Excess Kurtosis (γ_2)** and test the distribution against the null hypothesis of mesokurtosis.

U.111. **(Skewness vs. Kurtosis)** What is the conceptual difference between what Skewness (γ_1) measures and what Kurtosis (γ_2) measures?

U.112. **(Interpreting $Q.D.$)** If the Quartile Deviation (QD) of a distribution is 5, interpret what this means in terms of the data concentrated around the median.

U.113. **(Algebraic Proof)** Prove the property $\mu'_2 = \sigma^2 + \mu^2$ (for population parameters).

U.114. **(Conceptual)** When calculating the Median for a continuous variable from grouped data, why do we use **Upper Class Boundaries** instead of Upper Class Limits?

U.115. **(Moments to Mean)** Given the 1^{st} Raw Moment is $\mu'_1 = 15$. What does

this imply about the Arithmetic Mean?

U.116. **(Property of s^2)** Prove that $\sum(x_i - k)^2$ is minimized only when $k = \overline{x}$.

U.117. **(Software Instruction)** Write a conceptual instruction for a Software package to calculate the $\mathbf{90^{th}}$ Percentile($\mathbf{P_{90}}$) and the **Lower Inner Fence (LIF)** for a dataset of $N = 1000$.

Bivariate Statistics Problems

B.101. **(Combined Property Calculation)** Given two variables X and Y with $n = 10$. $\sum x = 20, \sum y = 50, \sum x^2 = 80, \sum y^2 = 300, \sum xy = 110$. a) Calculate the **Covariance (s_{xy})**. b) Calculate **Pearson's Correlation Coefficient (r)**. c) Interpret r in terms of strength and direction.

B.102. **(Advanced Transformation Property)** Variables X and Y have a correlation $r_{xy} = 0.8$. A transformation is applied: $U = 10 - 2X$ and $V = 5 + 3Y$. Calculate the correlation r_{uv} and justify the sign change based on the property of linear transformations.

B.103. **(Conceptual Limitation)** A researcher finds $r_{xy} = 0.95$ for the relationship between ice cream sales (X) and violent crime rates (Y). Explain why this strong correlation does not imply causation, and propose a specific **confounding variable** that could explain the relationship.

B.104. **(SLR Calculation Synthesis)** Using the summary statistics from B.101, calculate: a) The **slope (b_1)** of the regression of Y on X ($\hat{y} = b_0 + b_1 x$). b) The **intercept (b_0)**. c) The **Coefficient of Determination (R^2)**.

B.105. **(Property Proof)** Prove the algebraic property that the **Coefficient of Determination (R^2)** is the square of the Pearson's correlation coefficient (r) in the context of Simple Linear Regression.

B.106. **(Prediction and Residual)** The SLR equation is $\hat{y} = 10 + 5x$. An observation is $(x_i = 4, y_i = 35)$. Calculate the **predicted value ($\hat{y}_i$)** and the **residual (e_i)**. If the sum of residuals $\sum e_i$ is 10^{-6}, explain why this is acceptable in a computational setting.

B.107. **(Extrapolation Risk)** A regression model predicts annual sales based on GDP growth: $\hat{y} = 50 + 20x$. The observed GDP range (X) is $[1, 5]$. Explain the risk of **extrapolating** to predict sales when GDP growth is $X = 20$ in terms of model validity.

B.108. **(Algebraic Proof - Slope through Means)** Prove that the Least Squares Regression line, $\hat{y} = b_0 + b_1 x$, must pass through the point of means $(\overline{x}, \overline{y})$.

B.109. **(Conceptual Distinction)** Explain the difference between **Total Sum of Squares (SST)** and **Error Sum of Squares (SSE)** in terms of the variation they measure. Write the relationship between SST, SSE, and SSR.

B.110. **(Non-Linearity)** A scatter plot clearly shows a **parabolic relationship** ($Y = X^2$). What would be the expected approximate value of Pearson's r? Explain why the **Correlation Ratio ($\eta^2_{y \cdot x}$)** would be a more appropriate measure of association here.

B.111. **(Grouped Data Covariance)** Write the full mathematical formula for the sample **Covariance (s_{xy})** for grouped bivariate data using joint frequencies f_{ij} and marginal means $\overline{x}$ and $\overline{y}$.

B.112. **(Residual Property)** Prove the algebraic property that the sum of the residuals weighted by the predicted values is zero: $\sum \hat{y}_i e_i = 0$.

B.113. **(Heteroscedasticity - Conceptual)** One key assumption of the classical linear model is homoscedasticity. What does **Heteroscedasticity** mean in terms of the regression residuals, and how is it typically detected graphically?

B.114. **(Leverage and Influence)** Define the conceptual difference between a data point with high **leverage** and a data point with high **influence** in the context of SLR.

B.115. **(Conceptual)** If a dataset is used to calculate both the regression of Y on X ($\hat{y} = b_0 + b_1 x$) and the regression of X on Y ($\hat{x} = b_0' + b_1' y$), what is the relationship between the two slopes (b_1 and b_1') and the correlation coefficient (r)?

B.116. **(Software Instruction - LSA)** Write a conceptual instruction for a Software package to fit a Simple Linear Regression model to two variables X (Independent) and Y (Dependent) using the **Least Squares Approximation (LSA)** criterion.

B.117. **(Regression vs. Correlation)** Explain why regression analysis requires the designation of dependent and independent variables, while correlation analysis treats variables symmetrically.

B.118. **(Perfect Correlation)** If $r_{xy} = 1$, what must be the value of the slope b_1 of Y on X in relation to the standard deviations s_x and s_y?

B.119. **(Autocorrelation)** Define the concept of **Autocorrelation** (r_k). Write the formula for the autocorrelation coefficient at lag $k = 1$ for a time series X_t.

B.120. **(Conceptual)** What is the purpose of the **Intraclass Correlation (ICC)**, and how is it different from Pearson's r?

B.121. **(Spearman's ρ Calculation)** Ten students are ranked by two judges (R_1 and R_2). Calculate the rank difference d_i for each student.

DATASET/PARAMETERS:

Student	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
R_1	*1*	*3*	*5*	*7*	*9*	*2*	*4*	*6*	*8*	*10*
R_2	*2*	*1*	*6*	*8*	*10*	*3*	*5*	*7*	*9*	*4*

Calculate the sum of squared differences $\sum d_i^2$ and then calculate

Spearman's Rank Correlation (ρ).

B.122. **(Ties in Rank Correlation)** When ties occur in the data, the Pearson formula applied to ranks is generally preferred over the shortcut formula for ρ. Explain why the shortcut formula becomes inaccurate in the presence of ties.

B.123. **(Kendall's Tau - Conceptual)** Define the conceptual difference between a **Concordant Pair** and a **Discordant Pair** in the calculation of **Kendall's Tau (τ)**.

B.124. **(Contingency Table Degrees of Freedom)** A study on political affiliation (4 categories) and preferred news source (3 categories) results in a contingency table. Calculate the **Degrees of Freedom (df)** for the Chi-Square test of independence.

B.125. **(Chi-Square Calculation)** A 2×2 table has observed counts: $O_{11} = 20, O_{12} = 30, O_{21} = 10, O_{22} = 40$. Calculate the **Expected Count (E_{11}) ** and then the full χ^2 statistic.

B.126. **(Odds Ratio Interpretation)** A calculated **Odds Ratio (OR)** is $OR = 2.5$. Interpret this result in terms of the odds of the outcome occurring for the exposed group versus the unexposed group.

B.127. **(OR to Yule's Q)** Given $OR = 4$, calculate **Yule's Q**.

B.128. **(Chi-Square Based Measures)** A 3×3 contingency table has $\chi^2 = 50$ and $n = 200$. Calculate **Cramér's V** and **Pearson's Coefficient of Contingency (C)**.

B.129. **(Limitation of C)** Explain the main limitation of **Pearson's Coefficient of Contingency (C)** and how **Cramér's V** addresses this limitation.

B.130. **(Ordinal Association - Somer's D)** For two ordinal variables, X and Y, the number of concordant pairs is $N_c = 100$, discordant pairs is $N_d = 20$, and ties on the dependent variable Y is $T_y = 10$. Calculate **Somer's D_{yx}**.

B.131. **(Somer's D Symmetry)** Explain why **Somer's D_{yx}** is an asymmetric measure of association, while **Kendall's Tau (T)** is symmetric.

B.132. **(Tshuprow's T Limitation)** Explain why **Tshuprow's T** can only reach its maximum value of 1 when the contingency table is square ($R = C$).

B.133. **(Correlation Ratio η^2 Interpretation)** If the **Correlation Ratio $\eta^2_{y \cdot x}$** is 0.90 and r^2 is 0.40, what is the conclusion about the form of the relationship between X and Y?

B.134. **(Conceptual - Categorical)** Distinguish between the conceptual goals of

using the **Odds Ratio (OR)** versus the **Phi-statistic (ϕ)** for a 2×2 table.

B.135. **(Software Instruction - Rank)** Write a conceptual instruction for a Software package to calculate **Spearman's ρ** and **Kendall's τ** for a large dataset of $N = 1000$ paired ranks.

B.136. **(Autocorrelation vs. ρ)** Explain the conceptual difference between **Autocorrelation** and **Spearman's Rank Correlation** in terms of the variables they relate.

B.137. **(Regression Model Assumptions)** State the four core assumptions about the **error term (ϵ_i)** in the Simple Linear Regression model.

B.138. **(Prediction Interval Conceptual)** Distinguish between a **Confidence Interval for the Mean Response ($\mu_{y|x_0}$)** and a **Prediction Interval for a Single Response ($\hat{y}_{x_0}$)** at a specific value x_0. Which is wider, and why?

B.139. **(ANOVA to η^2)** The one-way ANOVA table for testing a quantitative variable (Y) against a categorical variable (X) yields $SS_{\text{Total}} = 500$ and $SS_{\text{Between}} = 400$. Calculate the **Correlation Ratio (η^2)** and interpret the result.

B.140. **(Software Instruction - Categorical)** Write a conceptual instruction for a Software package to test the independence of two categorical variables using the **Chi-Square test** and, if a significant association is found, calculate **Cramér's V**.

B.141. **(Combined Measure Interpretation)** Given $r = 0.75$, $s_x = 10, s_y = 12$. The SLR is $\hat{y} = 5 + b_1 x$. If $R^2 = 0.5625$, calculate the true value of r and the slope b_1. Reconcile the given r and R^2.

B.142. **(Residual Property Proof)** Prove the algebraic property that the correlation between the residuals (e_i) and the independent variable (x_i) is zero: $r_{xe} = 0$.

B.143. **(Extrapolation Check)** An SLR model for tree height (Y) vs. age (X) is highly linear up to $X = 100$ years. Predict the height at $X = 500$ and explain why this prediction is unreliable.

B.144. **(Odds Ratio to Probabilities)** Given an event prevalence of 20% in the unexposed group and an $OR = 3$. Estimate the probability of the event in the exposed group.

B.145. **(Software Instruction - R^2)** Write a conceptual instruction for a Software package to calculate the **Coefficient of Determination (R^2)** and the **Adjusted R^2** for a fitted SLR model.

B.146. **(Conceptual - R^2 vs η^2)** When a linear model is fitted, $R^2 = 0.8$. When

the Correlation Ratio is calculated, $\eta^2 = 0.95$. What is the strongest conclusion you can draw about the relationship?

B.147. **(Bivariate Data Types)** Provide one real-world example of each of the three types of bivariate relationships: Quantitative-Quantitative, Categorical-Categorical, and Quantitative-Categorical.

B.148. **(Autocorrelation Interpretation)** If the autocorrelation function (ACF) plot for a time series shows $r_1 = 0.9$ and $r_2 = 0.85$, what does this imply about the temporal dependence of the series?

B.149. **(Grouped Data Correlation)** Explain the primary challenge in calculating Pearson's r for a **grouped bivariate frequency distribution** compared to ungrouped data.

B.150. **(Regression Slope Sign)** If $r_{xy} = -0.6$, what must be the sign of the slope b_1 of Y on X? Justify your answer.

B.151. **(Property of Residuals)** Show that $SSE = \sum e_i^2$ is the minimized sum of squares in LSA.

B.152. **(Conditional Mean)** Define the **Conditional Mean** $E(Y|X = x_0)$ in the context of the population regression function.

B.153. **(Conceptual - Association)** Distinguish between the concepts of **Correlation** and **Association** as they apply to the types of data (Quantitative vs. Categorical).

B.154. **(Algebraic Proof - $\sum e_i = 0$)** Prove that the sum of the residuals from the Least Squares regression line is zero: $\sum(y_i - \hat{y}_i) = 0$.

B.155. **(Leverage Point Definition)** Write the definition of a leverage point using the diagonal elements of the hat matrix (h_{ii}).

B.156. **(Spearman's ρ Conceptual)** Explain why Spearman's rank correlation coefficient is a more robust measure of association than Pearson's r when the relationship is **monotonic but non-linear** or when the data contains extreme outliers.

B.157. **(Odds Ratio and Inverse)** If the Odds Ratio for X on Y is $OR = 4$, what is the Odds Ratio for the inverse relationship (i.e., Odds of $\sim Y$ for $X = 1$ vs. $X = 0$)?

B.158. **(Tied Ranks)** If three observations are tied at a rank that spans positions $5, 6, 7$, what is the single rank assigned to all three, and why?

B.159. **(Regression Line Intersection)** Given two SLR lines, $\hat{y} = b_0 + b_1 x$ and $\hat{x} = b_0' + b_1' y$. Under what condition (in terms of r) do these two lines become perpendicular?

B.160. **(Software Instruction - Heteroscedasticity)** Write a conceptual instruction for a Software package to diagnose **Heteroscedasticity** after fitting an SLR model.

B.161. **(Conceptual - R^2 Limitation)** Explain why R^2 must increase (or stay the same) when an additional predictor variable is added to a regression model.

B.162. **(Partition of Variation)** Given $SSR = 1000$ and $R^2 = 0.8$. Calculate the Total Sum of Squares (SST) and the Error Sum of Squares (SSE).

B.163. **(Phi Statistic Limitation)** A 2×2 table has $\chi^2 = 50$ and $n = 500$. Calculate the **Phi Statistic (ϕ)**. Explain why ϕ is only guaranteed to range from 0 to 1 when the marginal distributions are perfectly uniform.

B.164. **(Odds Ratio Equivalence)** Prove the equivalence of the two formulas for the Odds Ratio in a 2×2 table: $OR = \frac{a/b}{c/d}$ and $OR = \frac{ad}{bc}$.

B.165. **(Kendall's Tau Interpretation)** If Kendall's $\tau = -0.85$, what is the conclusion about the concordance/discordance of the paired ranks?

B.166. **(Conceptual)** What is the significance of the **Y-intercept (b_0)** in a regression model, and when is it often meaningless to interpret?

B.167. **(Software Instruction - D_{yx})** Write a conceptual instruction for a Software package to calculate **Somer's D_{yx}** for two ordinal variables, X and Y, specifying Y as the dependent variable.

B.168. **(Coefficient of Contingency Max Value)** Explain mathematically why the maximum value of C for a 3×3 table is $\sqrt{2/3} \approx 0.816$, thus confirming its primary limitation.

B.169. **(Correlation Ratio η^2)** Write the mathematical formula for the **Correlation Ratio ($\eta^2_{y \cdot x}$)** using the ANOVA sum of squares notation.

B.170. **(Time Series Lag)** A time series analysis requires examining the dependence between X_t and X_{t-k}. What is k called, and what value of r_k indicates a white noise process?

B.171. **(Algebraic Proof - SSE)** Prove that $SSE = (n-1)s_y^2(1-r^2)$.

B.172. **(Conceptual - Group Comparison)** When comparing the mean of a quantitative variable (Y) across five different categories of a categorical variable (X), which graphical tool is most appropriate, and what does the vertical size of the box indicate?

B.173. **(Regression vs. r)** If $r = 0$, what does this imply about the slope of the regression line b_1, and what is the resulting prediction $\hat{y}$ for any x_i?

B.174. **(Property of R^2)** Explain why R^2 is bounded between 0 and 1.

B.175. **(Residual Analysis)** A plot of residuals vs. predicted values shows a clear funnel shape (increasing variance). What assumption is violated, and what is the graphical term for this violation?

B.176. **(Conceptual - χ^2)** The value of the χ^2 statistic is heavily dependent on the sample size n. Explain how ϕ and V correct for this dependency.

B.177. **(Spearman's ρ vs. Pearson's r)** Calculate Pearson's r for the paired data: $(1,1),(2,2),(10,10)$. Calculate Spearman's ρ for the same data. Explain why they are equal in this case.

B.178. **(Extrapolation Limits)** Explain the danger of extrapolation when the relationship is known to be **non-monotonic** (e.g., cubic) outside the observed range.

B.179. **(OR vs. Relative Risk)** Distinguish between the **Odds Ratio (OR)** and the **Relative Risk (RR)** for a 2×2 table. When is $OR \approx RR$?

B.180. **(Conceptual - Influence)** Define a data point that is a high leverage point but has low influence.

B.181. **(Software Instruction - ρ)** Write a conceptual instruction for a Software package to test the association of two ordinal variables using **Spearman's ρ** and state the null and alternative hypotheses.

B.182. **(ANOVA Partition)** Given $SST = 100$, $SS_{\text{Between}} = 70$. Calculate the η^2 value and interpret it.

B.183. **(Correlation Ratio vs. R^2)** If a non-linear relationship exists and $\eta^2_{y \cdot x} = 0.7$, what is the range of possible values for r^2?

B.184. **(Conceptual)** Explain the difference between **Autocorrelation** and **Cross-correlation**.

B.185. **(Regression to the Mean)** Explain the concept of "regression to the mean" and how it relates to the magnitude of the slope b_1 when $r < 1$.

B.186. **(Conceptual)** What condition must be met for a fitted SLR model to be classified as a **Perfect Fit**?

B.187. **(Odds Ratio - Zero Cell)** Explain the computational problem that arises in the calculation of the Odds Ratio (OR) if any of the cell frequencies in a 2×2 table is zero.

B.188. **(Conceptual)** What is the significance of the **Eigenvalues** in Principal Components Analysis (PCA) when dealing with highly correlated variables? (Hint: Link to Variance).

B.189. **(Degrees of Freedom - χ^2)** For a 4×5 contingency table, what is the required χ^2 degrees of freedom?

B.190. **(Final Synthesis)** Given $r = -0.8$, $s_x = 5$, $s_y = 10$, $\overline{x} = 20$, $\overline{y} = 50$. Calculate the slope b_1, the R^2, and the predicted value of Y when $X = 25$.