POLITEKNIK MALAYSIA NILAI

DBTpris Diploma Bioteknologi Politeknik Nilai

Biostatistics for Beginners

Review Notes

GRASPING CORRELATION

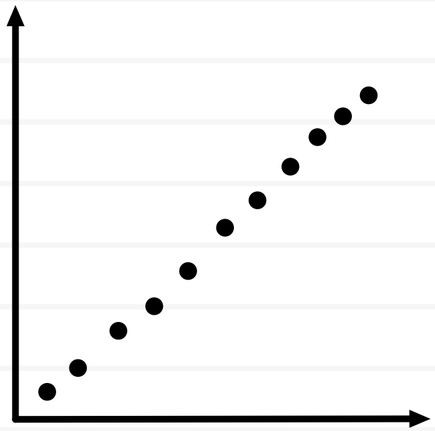
Scatter Plots - Definition

- A scatter plot is a graphical representation of the relationship between two quantitative variables.
- The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.

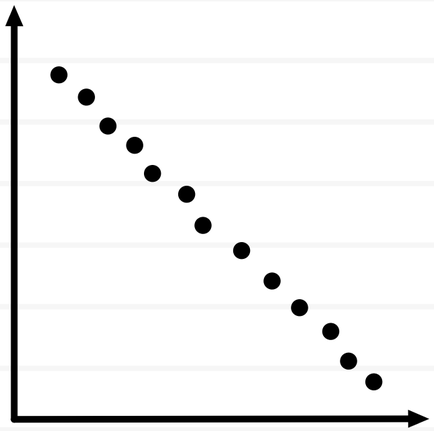
Scatter Plots - Usage

- It helps visualize the distribution, patterns, and possible correlations between the variables.
- To display paired numerical data.
- When multiple values of the dependent variable exist for each unique value of the independent variable.

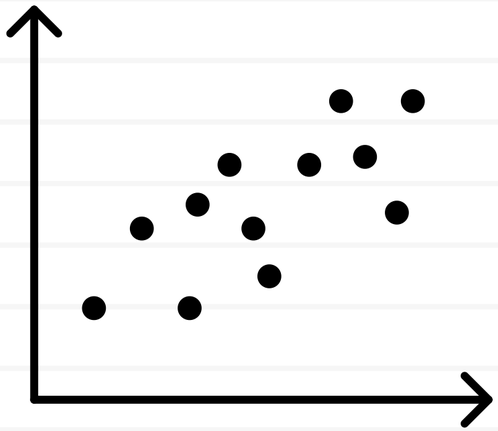
Scatter Plots - Interpretations



- **Positive Correlation:**
  Points trend upwards from left to right.



- **Negative Correlation:**
  Points trend downwards from left to right.



- **No Correlation:**
  Points are scattered without a discernible pattern.

Scatter Plots - Interpretations

Correlation Type	Description
Perfect Positive	Points form a perfectly straight upward line.
High Positive	Points are close to each other, trending upward.
Low Positive	Points are scattered but generally trend upward.
Perfect Negative	Points form a nearly straight downward line.
High Negative	Points are close to each other, trending downward.
Low Negative	Points are scattered but generally trend downward.
Outliers	Individual points far from the trend, which can affect correlation strength.

Pearson Correlation - Definition

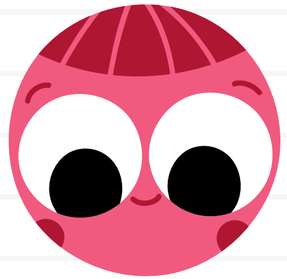
- **The Pearson correlation coefficient (r) quantifies the linear relationship between two continuous variables.**

Karl Pearson's coefficient of correlation is defined as a linear correlation coefficient that falls in the value range of -1 to +1.

Pearson Correlation - Range

- **1: Perfect positive linear correlation.**

  When the correlation coefficient is 1, it indicates that every increase in one variable corresponds to a proportional increase in the other. For example, shoe size tends to increase proportionally with foot length, representing an almost perfect correlation.



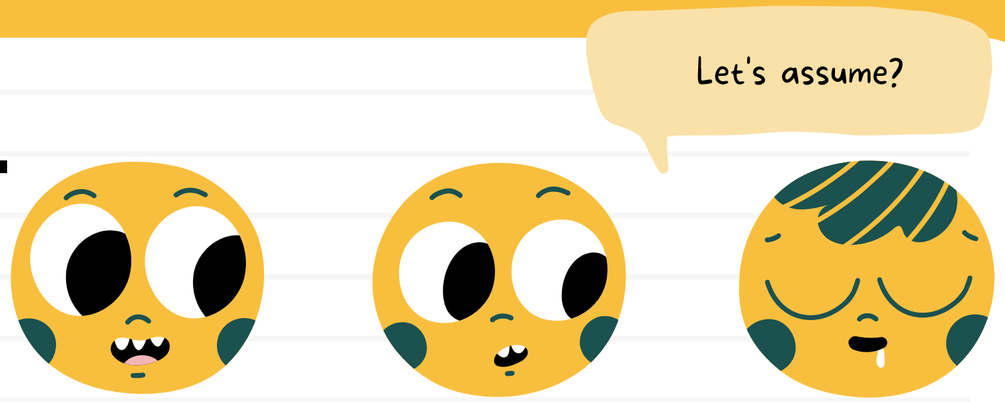
- **-1: Perfect negative linear correlation.**

  When the correlation coefficient is -1, it means that for every increase in one variable, there is a proportional decrease in the other. For example, the quantity of gas in a tank decreases in near-perfect inverse correlation with an increase in speed.

- **0: No linear correlation.**

  When the correlation coefficient is 0, it indicates that there is no consistent positive or negative relationship between the variables, meaning they are not related.

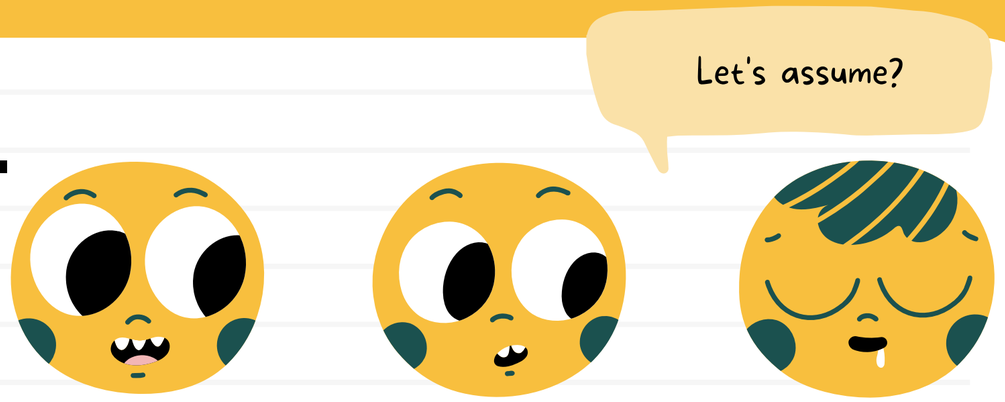
Pearson Correlation - Assumptions



- **Linearity: Relationship between variables is linear.**
  - This assumption means that the relationship between the two variables should follow a straight line. In other words, as one variable increases or decreases, the other variable changes in a predictable linear pattern.

    Example: A scatter plot of height and weight in a population generally shows a linear trend, where increases in height are often associated with increases in weight.

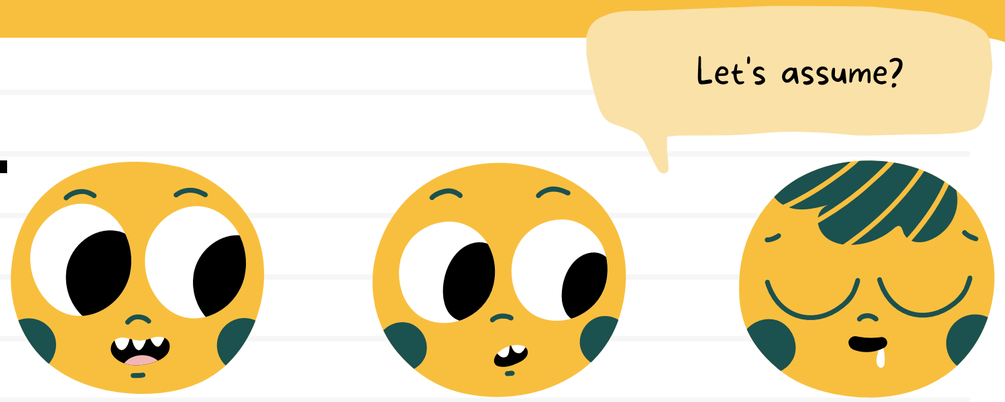
Pearson Correlation - Assumptions



- **Homoscedasticity: Constant variance of errors.**
  - Homoscedasticity refers to having a constant variance of errors across all levels of the independent variable. In other words, the spread or "scatter" of data points around the line of best fit should be roughly the same, regardless of whether values are high or low.

    Example: In a homoscedastic relationship between income and spending, individuals with low incomes and high incomes both show similar variability in spending patterns around the trend line.

Pearson Correlation - Assumptions



- **Normality: Data in each variable should be normally distributed (especially for small samples).**
  - This assumption means that the data for each variable should ideally follow a normal (bell-shaped) distribution, especially when working with small samples. For Pearson correlation, normally distributed data helps ensure that the correlation coefficient accurately represents the relationship without distortion.

Example: In analyzing test scores and study hours, a normal distribution for both variables implies that most values are near the mean, with fewer extreme values, giving a balanced view of their relationship.

Pearson Correlation - Example

Imagine you are studying the relationship between students' study hours and their exam scores.

Study Hours	Exam Score
1	55
2	60
3	65
4	70
5	80

Steps: Calculate the Pearson correlation coefficient to determine how closely related the study hours are to the exam scores. If you find a Pearson correlation of **0.95**, it indicates a strong positive correlation, meaning that as study hours increase, exam scores also tend to increase.

Spearman Correlation - Definition

- The Spearman rank correlation coefficient assesses the strength and direction of the association between two ranked variables.

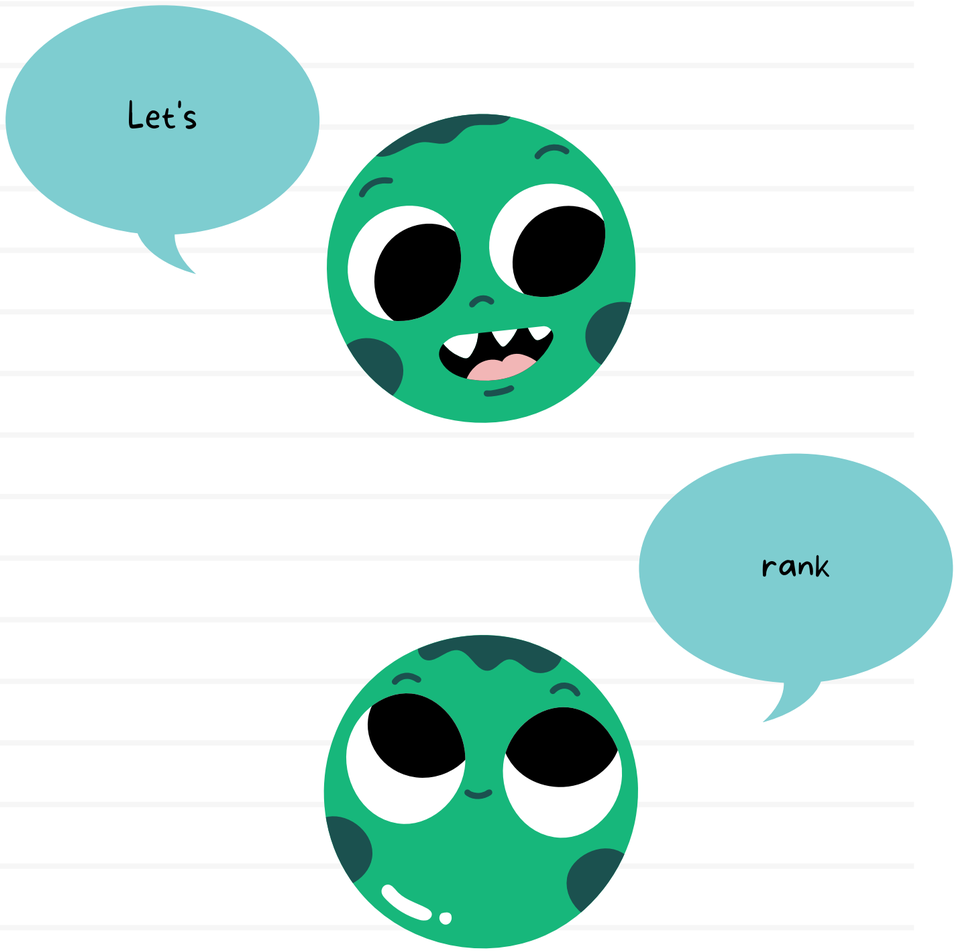
Spearman Correlation - Usage

- Suitable for non-parametric data or when assumptions of Pearson correlation are not met.
- Handles ordinal data or non-linear relationships effectively.

Spearman Correlation

- Range

- Similar to Pearson, r ranges from -1 to 1.



Spearman Correlation - Example

Consider you are ranking the same students based on their performance in a class activity and their exam scores.

Student	Activity Rank	Exam Score Rank
A	1	1
B	2	2
C	3	3
D	4	4
E	5	5

Steps: If you calculate the Spearman correlation coefficient and find it to be 1.0, this indicates a perfect positive monotonic correlation, meaning that higher activity ranks are associated with higher exam score ranks.

Summary

- **Scatter Plots** are essential for visualizing relationships between variables.
- **Pearson correlation** is used for continuous data with a linear relationship.
- **Spearman correlation** is used for ranked or ordinal data and does not assume a linear relationship.

Questions and Answers

Q1: What does a correlation of -0.45 mean?

A correlation of -0.45 means that 45% of the variance in one variable (e.g., x) is accounted for by the second variable (e.g., y), but in the opposite direction. This suggests a moderate negative relationship between the two variables, indicating that as one variable increases, the other variable tends to decrease.

Q2: What does a correlation of 0.45 mean?

A correlation of 0.45 means that 45% of the variance in one variable (e.g., x) is accounted for by the second variable (e.g., y). This suggests a moderate positive relationship between the two variables.