Correlation - Measuring Relationships

Detailed Study Notes

Correlation measures the strength and direction of the linear relationship between two variables. Pearson's correlation coefficient is the most commonly used measure.

Definition:
Correlation coefficient (r) quantifies how closely two variables move together in a linear fashion.

Formula:
$r = \Sigma[(x - x■)(y - ■)] / \sqrt{[\Sigma(x - x■)^2 \times \Sigma(y - ■)^2]}$

Or equivalently:
$r = Cov(X,Y) / (\sigma■ \times \sigma■)$

Range: $-1 \leq r \leq 1$

Interpretation:

r = 1: Perfect positive correlation
- As X increases, Y increases proportionally

r = 0: No linear correlation
- X and Y have no linear relationship

r = -1: Perfect negative correlation
- As X increases, Y decreases proportionally

Strength Guidelines:
|r| = 0.0 to 0.3: Weak
|r| = 0.3 to 0.7: Moderate
|r| = 0.7 to 1.0: Strong

Detailed Example:
Study hours (X): 2, 4, 6, 8, 10
Test scores (Y): 60, 70, 80, 85, 95

Step 1: Calculate means
x■ = 6, ■ = 78

Step 2: Calculate deviations and products
(2-6)(60-78) = (-4)(-18) = 72
(4-6)(70-78) = (-2)(-8) = 16
(6-6)(80-78) = (0)(2) = 0
(8-6)(85-78) = (2)(7) = 14
(10-6)(95-78) = (4)(17) = 68
Sum = 170

Step 3: Calculate sum of squared deviations
$\Sigma(x - x■)^2 = 40$
$\Sigma(y - ■)^2 = 590$

Step 4: Calculate r
$r = 170 / \sqrt{(40 \times 590)} = 170 / 153.62 \approx 0.93$

Interpretation: Strong positive correlation

Real-World Examples:

Strong Positive ($r \approx 0.8$ to 0.9):
- Height and weight
- Study time and grades
- Education and income
- Temperature and ice cream sales

Moderate Positive ($r \approx 0.4$ to 0.6):
- Exercise and health
- Reading and vocabulary
- Marketing spend and sales

Weak or No Correlation ($r \approx 0$):
- Shoe size and intelligence
- Hair color and athletic ability
- Random stock movements

Strong Negative ($r \approx -0.7$ to -0.9):
- Price and demand
- Speed and travel time
- Altitude and temperature

Important Considerations:

1. Correlation ≠ Causation
- High correlation doesn't prove one causes the other
- May be due to confounding variables
- Example: Ice cream sales and drowning (both caused by summer)

2. Linearity Assumption
- Pearson's r only measures LINEAR relationships
- May miss non-linear relationships
- Example: U-shaped or inverted-U relationships

3. Outliers
- Can dramatically affect correlation
- Always visualize data with scatter plots
- Check for influential points

4. Range Restriction
- Limited range can reduce correlation
- Example: SAT scores only from admitted students

Applications:
- Finance: Portfolio diversification, risk management
- Medicine: Relationship between treatment and outcomes
- Social Science: Studying relationships between variables
- Quality Control: Identifying process relationships
- Marketing: Understanding factor relationships