

Bivariate Statistics

1 Bivariate Data and Representation

1.1 Basics of Bivariate Relationships

Bivariate data consists of two variables measured on the same unit of observation. The data is represented as pairs (x_i, y_i) .

- **Quantitative-Quantitative:** Both variables are numerical (e.g., Height and Weight). Analysis focuses on correlation and regression.
- **Categorical-Categorical:** Both variables are qualitative (e.g., Gender and Voting Preference). Analysis focuses on association.
- **Quantitative-Categorical:** One of each (e.g., Salary and Job Title). Analysis often involves comparing groups (e.g., side-by-side boxplots, ANOVA).

1.2 Graphical Display

- **Scatter Plot:** The primary tool for Quantitative-Quantitative data.
 - X-axis: Independent (or predictor) variable.
 - Y-axis: Dependent (or response) variable.
 - Used to visualize the **form** (linear, non-linear), **direction** (positive, negative), and **strength** (strong, weak) of the relationship.
- **Side-by-Side Boxplots:** Used for Quantitative-Categorical data to compare the distribution of the quantitative variable across different categories.
- **Stacked or Clustered Bar Charts:** Used for Categorical-Categorical data to show the joint frequencies or proportions.

1.3 Bivariate Frequency Distribution (Grouped Data)

For grouped or discrete data, a **contingency table** (or two-way frequency table) is used.

- f_{ij} : Joint frequency of the i^{th} category of X and j^{th} category of Y.
- $f_{i.} = \sum_j f_{ij}$: Marginal frequency of the i^{th} row (for X).
- $f_{.j} = \sum_i f_{ij}$: Marginal frequency of the j^{th} column (for Y).
- $n = \sum_i \sum_j f_{ij} = \sum_i f_{i.} = \sum_j f_{.j}$: Total number of observations.

| | | Y Variable | | | | Total |
|------------|----------|------------|----------|----------|----------|----------|
| | | Y_1 | Y_2 | \dots | Y_c | |
| X Variable | X_1 | f_{11} | f_{12} | \dots | f_{1c} | $f_{1.}$ |
| | X_2 | f_{21} | f_{22} | \dots | f_{2c} | $f_{2.}$ |
| | \vdots | \vdots | \vdots | \ddots | \vdots | \vdots |
| | X_r | f_{r1} | f_{r2} | \dots | f_{rc} | $f_{r.}$ |
| Total | | $f_{.1}$ | $f_{.2}$ | \dots | $f_{.c}$ | n |

2 Covariance and Pearson's Correlation

2.1 Covariance (s_{xy} or σ_{xy})

Measures the direction and extent of joint variability between two quantitative variables.

- **Ungrouped Data (Sample):**

- Mathematical: $s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
- Computational: $s_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \right]$

- **Grouped Data (Sample):** Let x_i, y_j be class marks.

- Mathematical: $s_{xy} = \frac{1}{n-1} \sum_{i=1}^r \sum_{j=1}^c f_{ij} (x_i - \bar{x})(y_j - \bar{y})$
- Computational: $s_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^r \sum_{j=1}^c f_{ij} x_i y_j - \frac{(\sum f_{i.} x_i)(\sum f_{.j} y_j)}{n} \right]$

- **Interpretation:** Positive value implies positive relationship; negative value implies negative relationship. Magnitude is hard to interpret as it depends on units.

2.2 Pearson's Coefficient of Correlation (r)

A standardized measure of the **linear** relationship between two quantitative variables.

- **Definition:** $r = \frac{Cov(X,Y)}{s_x s_y} = \frac{s_{xy}}{s_x s_y}$

- **Ungrouped Data:**

- Mathematical: $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$
- Computational: $r = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}}$

- **Grouped Data:** (Using $u_i = \frac{x_i - a}{h}$ and $v_j = \frac{y_j - b}{k}$ for simplification)

$$r = \frac{n \sum f_{ij} u_i v_j - (\sum f_{i.} u_i)(\sum f_{.j} v_j)}{\sqrt{[n \sum f_{i.} u_i^2 - (\sum f_{i.} u_i)^2][n \sum f_{.j} v_j^2 - (\sum f_{.j} v_j)^2]}}$$

(If no simplification, use the computational formula with f_{ij}, x_i, y_j).

- **Interpretation:**

- r ranges from -1 to +1.
- $r = +1$: Perfect positive linear relationship.
- $r = -1$: Perfect negative linear relationship.
- $r = 0$: No **linear** relationship. A strong non-linear relationship could still exist.

3 Properties & Other Correlation Types

3.1 Algebraic Properties of Correlation (r)

- **Symmetry:** $r_{xy} = r_{yx}$.
- **Scale and Origin Invariant:** r is a pure number, independent of the units of measurement.
 - If $u_i = a + bx_i$ and $v_i = c + dy_i$ (where $b, d \neq 0$).
 - $r_{uv} = r_{xy}$ if b and d have the **same** sign.
 - $r_{uv} = -r_{xy}$ if b and d have **opposite** signs.
- **Range:** $-1 \leq r \leq +1$.
- **Independence:** If X and Y are independent, $r = 0$. The converse is **not** true (e.g., $Y = X^2$ for $X \in [-1, 1]$ has $r = 0$ but is perfectly dependent).

3.2 Rank Correlation (for Ordinal Data)

- **Spearman's Rank Correlation (ρ or r_s):**
 - Concept: Pearson's r calculated on the *ranks* of the data.
 - Formula (no ties): $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$, where $d_i = R(x_i) - R(y_i)$.
 - Formula (with ties): Calculate Pearson's r directly on the (average) ranks.
- **Kendall's Tau (τ):**
 - Concept: Based on concordant and discordant pairs.
 - Concordant Pair (N_c): A pair $(x_i, y_i), (x_j, y_j)$ where ranks agree ($x_i > x_j$ and $y_i > y_j$, or $x_i < x_j$ and $y_i < y_j$).
 - Discordant Pair (N_d): A pair where ranks disagree.
 - Formula (τ - b , adjusts for ties): $\tau_b = \frac{N_c - N_d}{\sqrt{(N_c + N_d + T_x)(N_c + N_d + T_y)}}$, where T_x, T_y are pairs tied on X and Y respectively.

3.3 Correlation Ratio (η)

- Concept: Measures the strength of a *non-linear* relationship (Y on X , $\eta_{y \cdot x}$).
- Formula: $\eta_{y \cdot x}^2 = \frac{\text{Sum of Squares Between Groups (SSR)}}{\text{Total Sum of Squares (SST)}} = \frac{\sum_i n_i (\bar{y}_i - \bar{y})^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2}$
- \bar{y}_i is the mean of Y for the i^{th} category of X .
- Properties: $0 \leq r^2 \leq \eta_{y \cdot x}^2 \leq 1$. If $\eta^2 = r^2$, the relationship is perfectly linear.

3.4 Autocorrelation (ACF)

- Concept: Correlation of a time series with a lagged version of itself.
- Formula (lag k): $r_k = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}$

3.5 Intraclass Correlation (ICC)

- Concept: Measures the reliability or agreement of measurements within groups (e.g., rater reliability).
- Formula (One-way ANOVA model): $ICC = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W}$
- MS_B : Mean Square Between groups. MS_W : Mean Square Within groups. k : number of measurements per group.

4 Simple Linear Regression (LSR)

4.1 The Model

Describes a linear relationship between a dependent variable (Y) and an independent variable (X).

- **Population Model:** $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
 - β_0 : Population intercept (mean of Y when $X = 0$).
 - β_1 : Population slope (change in mean of Y for one-unit increase in X).
 - ϵ_i : Random error term, $\epsilon_i \sim N(0, \sigma^2)$.
- **Fitted (Sample) Model:** $\hat{y}_i = b_0 + b_1 x_i$
 - \hat{y}_i : Predicted value of Y for a given x_i .
 - b_0 : Sample intercept (estimate of β_0).
 - b_1 : Sample slope (estimate of β_1).
- **Residual:** $e_i = y_i - \hat{y}_i$ (Observed - Predicted).

4.2 Least Square Approximation (LSA)

The "best fitting" line is found by minimizing the sum of the squared residuals (SSE).

$$S = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2$$

This is minimized by finding b_0 and b_1 using calculus (partial derivatives).

4.3 Fitted Model Coefficients

- **Slope (b_1):**
 - Definition: $b_1 = \frac{s_{xy}}{s_x^2} = r \left(\frac{s_y}{s_x} \right)$
 - Computational: $b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
 - Computational (raw): $b_1 = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}$
- **Intercept (b_0):**
 - Definition: $b_0 = \bar{y} - b_1 \bar{x}$

4.4 Algebraic Properties of Regression

- The sum of the residuals is zero: $\sum e_i = \sum (y_i - \hat{y}_i) = 0$.
- The sum of observed y_i equals the sum of fitted \hat{y}_i : $\sum y_i = \sum \hat{y}_i$.
- The regression line **always** passes through the point of means (\bar{x}, \bar{y}) .
- The residuals e_i are uncorrelated with the predictor x_i : $\sum x_i e_i = 0$.
- The residuals e_i are uncorrelated with the fitted values \hat{y}_i : $\sum \hat{y}_i e_i = 0$.

5 Regression Analysis and Diagnostics

5.1 Explained and Unexplained Variation

The total variation in Y can be partitioned (Analysis of Variance - ANOVA).

- **Total Sum of Squares (SST):** Total variation in Y.

$$SST = \sum (y_i - \bar{y})^2$$

- **Regression Sum of Squares (SSR):** Variation in Y *explained* by the model.

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

- **Error Sum of Squares (SSE):** Variation in Y *unexplained* by the model (residuals).

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

- **Fundamental Partition:** $SST = SSR + SSE$

5.2 Coefficient of Determination (R^2)

The proportion of the total variance in the dependent variable (Y) that is explained by the independent variable (X).

- **Formula:** $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- **Range:** $0 \leq R^2 \leq 1$.
- **Interpretation:** An R^2 of 0.64 means 64% of the variability in Y is accounted for by the linear relationship with X.
- **Relationship to r :** In Simple Linear Regression, $R^2 = r^2$ (the square of the Pearson correlation coefficient).

5.3 Fitted Model for Extrapolation

- **Interpolation:** Using the model to predict \hat{y} for an x -value *within* the range of the original x data. This is generally safe and is the purpose of the model.
- **Extrapolation:** Using the model to predict \hat{y} for an x -value *outside* the range of the original x data.
- **Caution:** Extrapolation is highly risky and unreliable. It assumes the linear trend continues indefinitely, which is rarely true.
- **Causation Warning:** "Correlation does not imply causation." A strong r or R^2 does not prove that X *causes* Y. There may be a lurking variable, or the causal direction may be reversed.

5.4 Outliers and Influential Observations

- **Outlier:** A data point with a large residual ($|y_i - \hat{y}_i|$). It lies far from the regression line.
- **Leverage Point:** A data point with an x -value that is extreme (far from \bar{x}). It has the *potential* to influence the line.
- **Influential Observation:** A point that, if removed, would cause a significant change in the regression line (b_0 or b_1). A point with high leverage and a large residual is often highly influential. (Measured by metrics like Cook's Distance).

6 Analysis of Categorical Data (Association)

6.1 Contingency Table (2x2)

For measures of association between two dichotomous variables (attributes).

| | Y=1 | Y=0 | Total |
|-------|---------|---------|---------------------|
| X=1 | a | b | $a + b$ |
| X=0 | c | d | $c + d$ |
| Total | $a + c$ | $b + d$ | $n = a + b + c + d$ |

6.2 Chi-Square (χ^2) Statistic

Tests for independence between two categorical variables in an $R \times C$ table.

- **Observed Counts (O_{ij}):** The actual frequencies in the table (a, b, c, d, \dots).
- **Expected Counts (E_{ij}):** The frequency expected if the two variables were independent.

$$E_{ij} = \frac{(\text{Row } i \text{ Total}) \times (\text{Column } j \text{ Total})}{n}$$

- **χ^2 Statistic:**

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- **Degrees of Freedom:** $df = (R - 1)(C - 1)$.

6.3 Measures of Association Based on χ^2

These measures standardize χ^2 to a range (usually 0 to 1).

- **The Phi Statistic (ϕ):** For 2x2 tables.

$$\phi = \sqrt{\frac{\chi^2}{n}} = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

- **Coefficient of Contingency (Pearson's C):**

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

(Limitation: Max value is < 1 , e.g., 0.707 for a 2x2).

- **Tshuprow's T:**

$$T = \sqrt{\frac{\chi^2}{n \sqrt{(R - 1)(C - 1)}}$$

(Reaches 1 only for square tables $R = C$).

- **Cramér's V:** Most widely used χ^2 -based measure.

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(R - 1, C - 1)}}$$

(Ranges from 0 (no association) to 1 (perfect association)).

6.4 Measures for 2x2 Tables

- **Odds Ratio (OR):**

$$OR = \frac{\text{Odds of } Y=1 \text{ if } X=1}{\text{Odds of } Y=1 \text{ if } X=0} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

(Interpretation: $OR = 1$ (no assoc), $OR > 1$ (positive), $OR < 1$ (negative). Range 0 to ∞).

- **Yule's Q:**

$$Q = \frac{ad - bc}{ad + bc} = \frac{OR - 1}{OR + 1}$$

(Ranges from -1 to +1. $Q = 0$ when $OR = 1$).

- **Yule's Y (Coefficient of Colligation):**

$$Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

6.5 Somer's D

An asymmetric measure of association for *ordinal* variables.

- N_c : Number of concordant pairs.
- N_d : Number of discordant pairs.
- T_y : Number of pairs tied on the dependent variable Y (but not X).
- $D_{yx} = \frac{N_c - N_d}{N_c + N_d + T_y}$ (when Y is dependent).