Chapter 5

Correlation and Regression Study

Contents

Double Statistical Series

Definition 1: When two statistical variables are studied on a given population, the resulting data is called a **double statistical series**.

Scatter Plot

Definition 2: The set of points M_i with coordinates (x_i, y_i).

5.1 Case 1: Two-line Table

x_i	x_1	x_2	$\dots$	x_n
y_i	y_1	y_2	$\dots$	y_n

Definition 3: Marginal Means

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} x_i, \quad \bar{Y} = \frac{1}{N}\sum_{i=1}^{n} y_i$$

Marginal Variances

$$V(X) = \left(\frac{1}{N}\sum_{i=1}^{n} x_i^2\right) - \bar{X}^2, \quad V(Y) = \left(\frac{1}{N}\sum_{i=1}^{n} y_i^2\right) - \bar{Y}^2$$

Standard Deviations

$$\delta_X = \sqrt{V(X)}, \quad \delta_Y = \sqrt{V(Y)}$$

Covariance

$$\operatorname{cov}(X, Y) = \left(\frac{1}{N}\sum_{i=1}^{n} x_i y_i\right) - \bar{X}\bar{Y}$$

5.1.1 5.1.1 Regression Line (Least Squares Method)

Theorem 5.1.1. *The regression line of Y in terms of X, denoted $D_Y(X)$, is:*

$$Y = aX + b \quad \textit{where} \quad a = \frac{cov(X, Y)}{V(X)}, \quad b = \bar{Y} - a\bar{X}$$

Properties:

- It is a unique line.
- It always passes through the point $(\bar{X}, \bar{Y})$.

5.1.2 5.1.2 Linear Correlation Coefficient

Definition 4:

$$r = \frac{\text{cov}(X, Y)}{\delta_X \delta_Y}$$

Remark 5.1.1.
- $-1 \leq r \leq 1$
- *If* $r = 0$*: no correlation (independent variables)*
- *If* $0 < r < 1$*: weak to strong positive correlation*
- *If* $-1 < r < 0$*: weak to strong negative correlation*

Example 5.1

x_i	2	5	6	10	12
y_i	83	70	70	54	49

$$\bar{X} = 7, \quad \bar{Y} = 65.2, \quad V(X) = 12.8, \quad V(Y) = 150.16$$

$$\delta_X = 3.578, \quad \delta_Y = 12.25, \quad \text{cov}(X, Y) = -43.6$$

$$a = \frac{-43.6}{12.8} = -3.4, \quad b = 65.2 + 23.8 = 89 \Rightarrow Y = -3.4X + 89$$

$$r = \frac{-43.6}{3.578 \cdot 12.25} = -0.99$$

Conclusion: There is a strong negative linear correlation between X and Y.

5.2 5.2 Case 2: Three-line Table

x_i	x_1	x_2	$\dots$	x_k
y_i	y_1	y_2	$\dots$	y_k
n_i	n_1	n_2	$\dots$	n_k

Marginal Means

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{k} n_i x_i, \quad \bar{Y} = \frac{1}{N} \sum_{i=1}^{k} n_i y_i$$

Marginal Variances

$$V(X) = \left(\frac{1}{N} \sum_{i=1}^{k} n_i x_i^2 \right) - \bar{X}^2, \quad V(Y) = \left(\frac{1}{N} \sum_{i=1}^{k} n_i y_i^2 \right) - \bar{Y}^2$$

Standard Deviations

$$\delta_X = \sqrt{V(X)}, \quad \delta_Y = \sqrt{V(Y)}$$

Covariance

$$\operatorname{cov}(X,Y) = \left(\frac{1}{N}\sum_{i=1}^{k} n_i x_i y_i\right) - \bar{X}\bar{Y}$$

5.3 5.3 Case 3: Contingency Table

XY	y_1	y_2	$\dots$	y_l
x_1	n_{11}	n_{12}	$\dots$	n_{1l}
x_2	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{kl}

Marginal Means

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i, \quad \bar{Y} = \frac{1}{N}\sum_{j=1}^{l} n_j y_j$$

Marginal Variances

$$V(X) = \left(\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2\right) - \bar{X}^2, \quad V(Y) = \left(\frac{1}{N}\sum_{j=1}^{l} n_j y_j^2\right) - \bar{Y}^2$$

Covariance

$$\operatorname{cov}(X,Y) = \left(\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} x_i y_j\right) - \bar{X}\bar{Y}$$

Example 5.2. We consider the following double-entry table:

XY	1	2	4	n_i
3	2	0	3	5
5	4	6	1	11
6	5	1	7	13
n_j	11	7	11	$N = 29$

Marginal distributions

$$x_i = 3,\ 5,\ 6 \quad \text{with} \quad n_i = 5,\ 11,\ 13$$

$$y_j = 1,\ 2,\ 4 \quad \text{with} \quad n_j = 11,\ 7,\ 11$$

Marginal means

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i = \frac{5\cdot 3 + 11\cdot 5 + 13\cdot 6}{29} = 5.10$$

$$\bar{Y} = \frac{1}{N}\sum_{j=1}^{l} n_j y_j = \frac{11\cdot 1 + 7\cdot 2 + 11\cdot 4}{29} = 2.38$$

Marginal variances

$$V(X) = \left(\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2\right) - \bar{X}^2 = \left(\frac{5\cdot 9 + 11\cdot 25 + 13\cdot 36}{29}\right) - (5.10)^2 = 1.13$$

$$V(Y) = \left(\frac{1}{N}\sum_{j=1}^{l} n_j y_j^2\right) - \bar{Y}^2 = \left(\frac{11\cdot 1^2 + 7\cdot 4 + 11\cdot 16}{29}\right) - (2.38)^2 = 1.75$$

Marginal standard deviations

$$\delta_X = \sqrt{V(X)} = \sqrt{1.13} = 1.06, \quad \delta_Y = \sqrt{V(Y)} = \sqrt{1.75} = 1.32$$

Covariance of X and Y

$$\mathrm{cov}(X, Y) = \left(\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} x_i y_j\right) - \bar{X}\bar{Y} = 0$$