Course: Mathematical statistics

Week 11: Correlation Coefficient

Lecturer: Nagulama Moses

Kumi University

May 24, 2025

Outline

Intended learning outcomes

- Define the correlation coefficient and explain its purpose in statistics.
- Differentiate between positive, negative, and zero correlation using real-life examples.
- Interpret the value of the Pearson correlation coefficient (r) within the range $-1 \leq r \leq 1$ explaining strength and direction.

Bivariate Distribution and Correlation Coefficient

A bivariate distribution refers to the joint distribution of two random variables, say X and Y.

It describes how the values of these two variables are distributed together. When analyzing a bivariate distribution, we often ask the following questions:

Are the two variables related?

This question explores whether there is any association or dependency between the two variables. For example, do changes in one variable correspond to changes in the other?

What is the strength of the relationship?

If the variables are related, we want to quantify the degree of their association. A strong relationship implies that changes in one variable are closely tied to changes in the other.

What type of relationship exists between the two variables?

Relationships can take various forms: linear, non-linear, positive (direct), or negative (inverse). Identifying the nature of the relationship helps us understand the underlying dynamics.

What kind of predictions can you make from the relationship?

Once the relationship is established, we may use it to predict the value of one variable based on the value of the other.

Role of Correlation Coefficient in Answering These Questions

The correlation coefficient (commonly Pearson's r) provides answers to the first two questions:

Are the two variables related?

The correlation coefficient measures the extent to which two variables are linearly related.
A value of $r = 0$ indicates no linear relationship, while $r \neq 0$ suggests some degree of linear association. For example, if $r = 0.8$, it implies a

strong positive linear relationship, whereas $r = -0.5$ suggests a moderate negative linear relationship.

What is the strength of the relationship?

The magnitude of r quantifies the strength of the linear relationship:

$|r| = 1$: Perfect linear relationship.

$0.7 \leq |r| < 1$: Strong linear relationship.

$0.3 \leq |r| < 0.7$: Moderate linear relationship.

$0 < |r| < 0.3$: Weak linear relationship.

$|r| = 0$: No linear relationship.

Role of Regression in Answering the Last Two Questions

While the correlation coefficient addresses the first two questions, regression analysis helps answer the last two:

What type of relationship exists between the two variables?

Regression analysis allows us to model the relationship between X and Y. For instance:

A linear regression model assumes a straight-line relationship ($Y = a + bX$).

Non-linear regression models can capture more complex relationships (e.g., quadratic, exponential).

What kind of predictions can you make from the relationship?

Regression provides a predictive equation that estimates the value of Y for a given value of X. For example, in a linear regression model:

$$Y = a + bX$$

where:

a: Intercept (value of Y when X=0).

b: Slope (rate of change of Y with respect to X).

Using this equation, we can predict Y for any new value of X.

Definition of correlation and regression

Correlation is a statistical technique used to assess whether a relationship exists between two variables and, if so, to quantify the strength and direction of that relationship.

It provides a numerical measure that ranges from -1 to +1, indicating the degree of linear association between the variables. A value of r=0 implies no linear relationship, while values closer to -1 or +1 indicate stronger relationships.

Regression is a statistical technique used to model and describe the nature of the relationship between variables.

It establishes a mathematical equation that captures how one variable (the dependent variable) changes as a function of another (the independent variable).

correlation

Correlation describes how changes in one variable correspond to changes in another. The nature of this relationship can be categorized as follows:

Positive Correlation (Direct Relationship)

When two variables deviate in the same direction, meaning an increase in one variable corresponds to an increase in the other, or a decrease in one corresponds to a decrease in the other, the correlation is said to be positive or direct .

Examples of positively correlated variables include: Heights and weights of

a group of people where taller individuals tend to weigh more.

Negative Correlation (Inverse Relationship)
When an increase in one variable corresponds to a decrease in the other, or vice versa, the correlation is said to be negative or inverse .

Examples of negatively correlated variables include:

Price and demand of a commodity: As the price of a product increases, its demand typically decreases.

Perfect Correlation

Examples of Perfect Correlation.

The relationship between the radius of a circle and its circumference ($C = 2\pi r$). As the radius increases, the circumference increases proportionally.

Methods for Measuring Correlation

- Scatter Plot; which is a visual method to assess the direction and general pattern of the relationship between two variables.
- Pearson Correlation Coefficient (r); Measures the strength and direction of a linear relationship.
- Spearman's Rank Correlation (ρ or r_s)for Non-parametric method based on ranked values.
- Kendall's τ; for Non-parametric method to assess ordinal association between two quantities.

Correlation Coefficients

- Pearson's r; Measures linear correlation.
- Spearman's Rank Correlation Coefficient (r_s)

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

where d is the difference between ranks.

- Kendall's τ; Based on concordant and discordant pairs:

$$\tau = \frac{(C-D)}{\frac{1}{2}n(n-1)}$$

where C = number of concordant pairs, D = discordant pairs.

Definition of Pearson's Correlation Coefficient

Karl Pearson, a biometrician (1867–1936), developed the *product moment correlation coefficient* to measure the strength and direction of the linear relationship between two variables.

Definition: The correlation coefficient between two random variables X and Y, denoted by $\rho(X, Y)$ or ρ_{XY}, is given by:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Where:

- $\text{Cov}(X, Y)$ is the covariance of X and Y
- σ_X, σ_Y are the standard deviations of X and Y

Covariance and Variance Expressions

Covariance:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Variance:

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad \sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Expanded Covariance Formulation

$$\begin{aligned}\text{Cov}(X,Y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\bar{y}) \\ &= \frac{1}{n}\left(\sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + n\bar{x}\bar{y}\right) \\ &= \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}\end{aligned}$$

Or equivalently:

$$\text{Cov}(X,Y) = \frac{1}{n}\sum x_i y_i - \frac{\sum x_i \sum y_i}{n^2}$$

$$\Rightarrow \quad \text{Numerator: } n\sum x_i y_i - \sum x_i \sum y_i = S_{XY}$$

Covariance and Variance Shortcuts

Define:

$$S_{XY} = n\sum x_i y_i - \sum x_i \sum y_i$$

$$S_{XX} = n\sum x_i^2 - (\sum x_i)^2, \quad S_{YY} = n\sum y_i^2 - (\sum y_i)^2$$

$$\sigma_X = \sqrt{S_{XX}}, \quad \sigma_Y = \sqrt{S_{YY}}$$

Then:

$$\rho(X, Y) = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}}$$

Correlation Coefficient: Invariance under Linear Transformation

Let:

$$u = \frac{x-a}{h}, \quad v = \frac{y-b}{k} \quad \text{(for constants } a, b, h, k)$$

Then:

$$x = a + hu, \quad y = b + kv$$

Expectation

$$E[x] = a + hE[u], \quad E[y] = b + kE[v]$$

$$x - E[x] = h(u - E[u]), \quad y - E[y] = k(v - E[v])$$

Covariance

$$\text{Cov}(x, y) = E[(x - E[x])(y - E[y])] = hk \cdot \text{Cov}(u, v)$$

Variance

$$\sigma_X^2 = h^2\sigma_u^2, \quad \sigma_Y^2 = k^2\sigma_v^2$$

Therefore:

$$\rho(x,y) = \frac{hk \cdot \mathrm{Cov}(u,v)}{h\sigma_u \cdot k\sigma_v} = \frac{\mathrm{Cov}(u,v)}{\sigma_u\sigma_v} = \rho(u,v)$$

Pearson's correlation coefficient is invariant under change of origin and scale.

Example: Heights of Fathers and Sons

The heights (in kg) of 8 fathers X and their sons Y are given:

X (Father)	65	66	67	67	68	69	70	72
Y (Son)	67	68	65	68	72	72	69	71

calculate the correlation coefficient of x and y and comment on your results.

We use the formula:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

x	y	x^2	y^2	xy
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112

$$\sum x = 544 \qquad \sum y = 552$$
$$\sum x^2 = 37028 \qquad \sum y^2 = 39132$$
$$\sum xy = 37560 \qquad n = 8$$

Substitute into the formula:

$$r = \frac{8(37560) - (544)(552)}{\sqrt{[8(37028) - (544)^2][8(39132) - (552)^2]}}$$

$$\text{Numerator: } 8(37560) - 544 \times 552 = 300480 - 300288 = 192$$

$$\text{Denominator: } \sqrt{(296224 - 295936)(313056 - 304704)}$$

$$= \sqrt{288 \times 8352} = \sqrt{2405376} = 1550.96$$

Thus,

$$r = \frac{192}{1550.96} \approx 0.124$$

The Pearson correlation coefficient is approximately:

$$r \approx 0.124$$

- This indicates a very weak positive linear relationship.
- The heights of fathers and their sons are not strongly correlated in this data.
- Further investigation with a larger sample might provide more insight.

Example 2: Correlation Between Age and Blood Pressure

Problem Statement

The blood pressure (mmHg) and age (years) of 6 randomly selected subjects are shown below:

Age (x)	Pressure (y)
43	128
48	120
56	135
61	143
67	141
70	152

Calculate the product moment correlation coefficient between age and blood pressure, and comment on your result.

x	y	x^2	y^2	xy
43	128	1849	16384	5504
48	120	2304	14400	5760
56	135	3136	18225	7560
61	143	3721	20449	8723
67	141	4489	19881	9447
70	152	4900	23104	10640
Sum $\Sigma x = 345$	$\Sigma y = 819$	**20399**	**112443**	**47634**

Pearson's correlation coefficient:

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$r = \frac{6(47634) - (345)(819)}{\sqrt{6(20399) - 345^2} \cdot \sqrt{6(112443) - 819^2}}$$

$$r = \frac{285804 - 282555}{\sqrt{122394 - 119025} \cdot \sqrt{674658 - 670761}}$$

$$r = \frac{3249}{\sqrt{3369} \cdot \sqrt{3897}} = \frac{3249}{58.04 \times 62.42} \approx \frac{3249}{3624.9}$$

$$\boxed{r \approx 0.896}$$

Interpretation of Result

- The computed correlation coefficient is approximately $r = 0.896$.
- This indicates a strong positive linear relationship between age and blood pressure.
- As age increases, blood pressure also tends to increase.

Theorem

If two random variables X and Y are independent, then they are uncorrelated. That is:

$$X \perp Y \quad \Rightarrow \quad \text{Cov}(X, Y) = 0 \quad \Rightarrow \quad \rho_{XY} = 0$$

Proof

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

If X and Y are independent, then:

$$E[XY] = E[X]E[Y]$$

$$\Rightarrow \text{Cov}(X, Y) = E[X]E[Y] - E[X]E[Y] = 0$$

$$\Rightarrow \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

Example: Given the following data:

Observation	X	Y
1	-1	2
2	1	2
3	-1	0
4	1	0

- Determine whether X and Y are independent.
- Compute the correlation coefficient $\rho(X, Y)$.

We construct the joint and marginal distributions:

$X \backslash Y$	0	2	Total
-1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
Total	$\frac{1}{2}$	$\frac{1}{2}$	1

We construct the joint and marginal distributions:

$X \backslash Y$	0	2	Total
-1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
Total	$\frac{1}{2}$	$\frac{1}{2}$	1

Check:

$$P(X = -1, Y = 2) = \frac{1}{4}, \quad P(X = -1)P(Y = 2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Therefore, X and Y are independent.

$$\bar{x} = \frac{-1+1-1+1}{4} = 0, \quad \bar{y} = \frac{2+2+0+0}{4} = 1$$

$$\sigma_x^2 = \frac{1}{4}[(-1)^2 + 1^2 + (-1)^2 + 1^2] = 1, \quad \sigma_x = 1$$

$$\sigma_y^2 = \frac{1}{4}[(2-1)^2 + (2-1)^2 + (0-1)^2 + (0-1)^2] = 1, \quad \sigma_y = 1$$

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} = \frac{0}{1 \cdot 1} = 0$$

Conclusion:

- X and Y are **independent**.
- The correlation coefficient is **zero**, i.e., they are **uncorrelated**.

References

- Hogg,R;Mckean,J;Craig,A(2012).Introduction to mathematical statistics, 7th edition, pearson Prentice Hall, 2012.

- Hastings K.J,(1997) Probability and statistics, Addison Wesley reading,massachusetts.

Thank You!

Next Lecture: The least square method