Covariance and Correlation

Lesson 7

Contents

Covariance

Definition: Suppose X and Y are random variables with means μ_X and μ_Y . The covariance of X and Y is defined as :

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

Properties of Covariance

1. $\mathrm{Cov}(aX + b, cY + d) = a\,c\,\mathrm{Cov}(X, Y)$; a, b, c, d are constants
2. $\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y)$
3. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
4. $\mathrm{Cov}(X, Y) = E(XY) - \mu_X \mu_Y$
5. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$ for any X and Y
6. If X and Y are independent then $\mathrm{Cov}(X, Y) = 0$ (Warning: The converse is false)

Computing Covariance - Discrete case

Since covariance is defined as an expected value we compute it in the usual way as a sum or integral.

If X and Y have joint pmf $p(x_i, y_j)$ then:

$$\mathrm{Cov}(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)(x_i - \mu_X)(y_j - \mu_Y) = \left(\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)x_i y_j\right) - \mu_X \mu_Y$$

Computing Covariance - Continuous case

If X and Y have joint pdf f (x, y) over range [a, b] × [c, d] then:

$$\mathrm{Cov}(X,Y) = \int_c^d \int_a^b (x-\mu_x)(y-\mu_y) f(x,y)\, dx\, dy = \left(\int_c^d \int_a^b xy f(x,y)\, dx\, dy\right) - \mu_x \mu_y$$

Example 1: Flip a fair coin 3 times. Let X be the number of heads in the first 2 flips and let Y be the number of heads on the last 2 flips. Compute Cov(X, Y).

solution:

- $E(X) = 1 = E(Y)$, $E(X) = \mu_X$, $E(Y) = \mu_Y$
- $Cov(X, Y) = E(XY) - \mu_X \mu_Y$
- $E(X) = p(x_1).\ x_1 + \dots + p(x_n).\ x_n$

Example 2: Let X be a random variable that takes values {-2; -1; 0; 1; 2} each with probability 1/5. Let $Y = X^2$. Compute Cov(X, Y).

Covariance without probabilities

For a data set with two columns:

$$Cov(X,Y) = \frac{\sum_{1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Example 3: Find the covariance between the given two sets of data

X = {2, 5, 8, 11},   Y = {5, 9, 1, 4}.

Solution:

```
import numpy as np
x = np.array([2, 5, 8, 11])
y = np.array([5, 9, 1, 4])
cov = np.cov(x,y)[0][1]
print(cov)
```

Correlation

Definition: The correlation coefficient between X and Y is defined by:

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \varrho$$

"Cor (X, Y) measures the linear relationship between X and Y."

Properties of Correlation

1. $\varrho \in [-1, 1]$

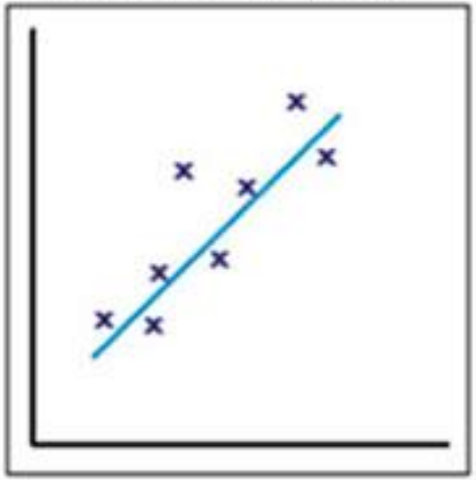
2. Además:

   $\varrho = +1$ if and only if $Y = aX + b$ with $a > 0$,

   $\varrho = -1$ if and only if $Y = aX + b$ with $a < 0$

3. Key components: magnitude and sign

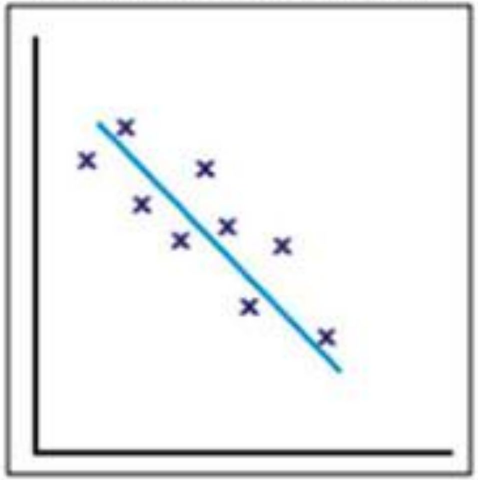
Positive correlation



The points lie close to a straight line, which has a positive gradient.

This shows that as one variable **increases** the other **increases**.

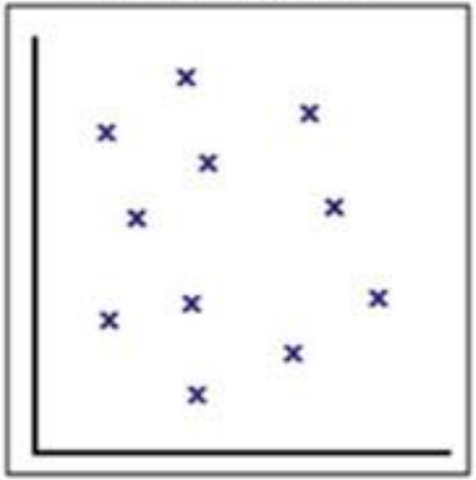
Negative correlation



The points lie close to a straight line, which has a negative gradient.

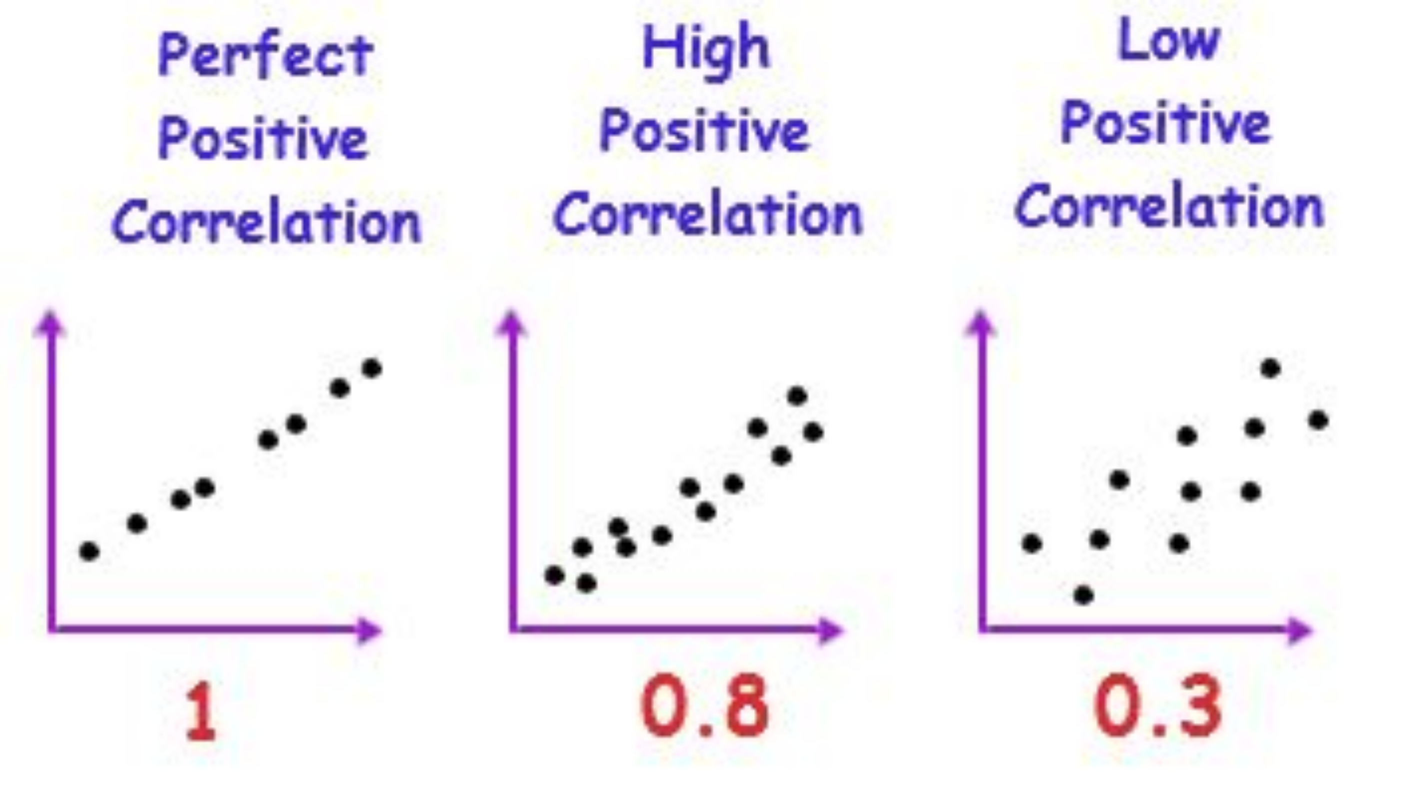
This shows that as one variable **increases**, the other **decreases**.

No correlation



There is no pattern to the points.

This shows that there is **no connection** between the two variables.



Example 4: Compute the correlation of Example 1.

Correlation without probabilities

For a data set with two columns:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Math Team

Christian Córdova

Bachelor in Mathematics

José Castro

Bachelor in Computer Science


Statistics
Density
population
Probability
experiment
intersection
Venn-Diagram
probability-tree
box-plot
observation
distribution
Standard
standard
randomness
histogram
bargraph
event
z-score
Median
union
Mode
Normal
independence
Mean
dotplot
pie-chart
conditional
stem-leaf
Random
chance
Deviation
Skewed
Sample