**Correlation and Covariance**

**Correlation** is one of the most common statistical concepts. It is a statistical technique that determines how one variable changes with another variable. It gives us the degree of the relationship between the two variables.

Chart, scatter chart

Description automatically generated

**Covariance** provides similar information with correlation. However, correlation goes beyond covariance and gives information also about the strength of the relationship between two variables. **Covariance does not provide information about the strength of the relationship.**

Chart, scatter chart

Description automatically generated

Table

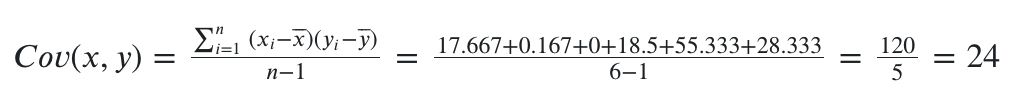
Description automatically generated

We can say correlation is calculated by the division of covariance by the standard deviation of variables. Because the correlation is a number between -1 and 1 it is often referred to as the **correlation coefficient**.

As we said earlier; covariance and correlation provide similar information while correlation goes beyond covariance and also gives us an idea about the **strength of the relationship**.

Table

Description automatically generated



The covariance between the temperature and the number of people is 24. Since the covariance is positive, the temperature and number of people have a positive relationship. As temperature rises, so does the number of people.

However there is no information about how strong the relationship is, and the answer to this problem is the correlation:

A picture containing text

Description automatically generated

The 0.909 which is called the **correlation coefficient shows that there is a very strong correlation between temperature and number of people**.

Pearson Correlation Coefficient

It is a number between -1 and 1 that indicates the strength of the relationship.

*-1 : Indicates complete negative correlation*

*+1 : Indicates complete correlation*

*0 : Indicates no correlation*

The Pearson correlation coefficient is calculated as follows:

Text

Description automatically generated

It captures not only the strength but also the direction of the linear association between two continuous variables. And **it tries to draw a line of best fit through the data points of two variables**. The picture below, in which the correlation coefficient is denoted by “r” illustrates this point very well. Blue arrows are the lines of best fit through the data points of two variables.

Chart

Description automatically generated

### **Calculate Correlation and Covariance with Python**

We can easily calculate covariance and correlation with numpy. The results of **np.cov()** and **np.corrcoef()** commands are displayed in matrix form. You remember the covariance between temperature and number of people was 24, and the correlation was 0,9. Try to find these values in the result matrices and ignore the other values. You can compare the values we get by numpy with the values we calculated manually.

import numpy as np

temp=[93,84,82,78,98,70]

number\_of\_people=[13,10, 11, 8, 15, 9]

print("covariance: ", np.cov(temp, number\_of\_people))

print("correlation: ", np.corrcoef(temp, number\_of\_people))

covariance: [[102.56666667 24. ]

[ 24. 6.8 ]]

correlation: [[1. 0.90876934]

[0.90876934 1. ]]