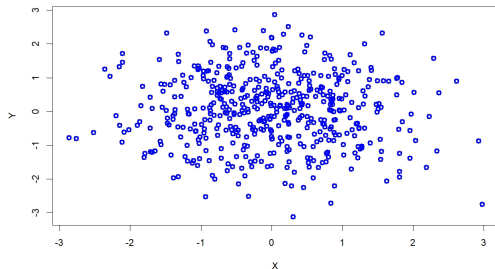


Covariance and correlation coefficient (Lecture 7)

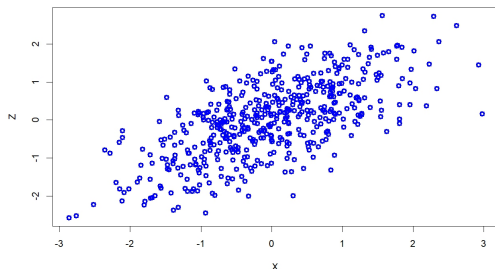
- two random variables can be
 - ▶ **independent**: the income of two randomly chosen people, or
 - ▶ **not independent**: the monthly income of a given person now, and in next April
- the **strength of the connection** can be different:
 - ▶ the age and monthly income of a randomly chosen person has a "strong connection", young and elderly people often have significantly less income;
 - ▶ the age and the height of an adult can have a "weak connection".
- **covariance** and **correlation coefficient** measure the strength of the connection (among other possibilities)

Independent random variables



500 random points on the plane, whose coordinates follow **independent** standard normal distribution. Both the covariance and correlation coefficient will be equal to **0**.

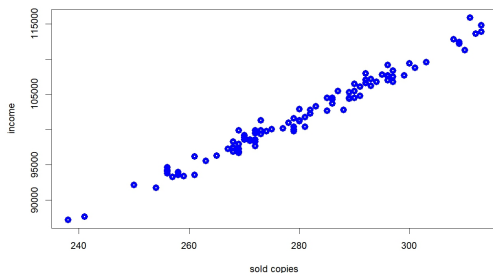
Positive correlation



A sample of size 500 from the following two-dimensional distribution: $(X, \frac{X+Z}{\sqrt{2}})$, where $X, Z \sim N(0, 1)$ are independent.

The larger X is, „probably” the larger $(X+Z)/\sqrt{2}$ is \rightarrow both the **covariance** and the **correlation coefficient** is **positive**.

Strong positive correlation



Sample of size 100 from distribution $(X + Y, 300X + 400Y)$, where $X \sim \text{Poisson}(100)$ and $Y \sim \text{Poisson}(180)$ are independent. The points fit very well to a line with positive slope \rightarrow the **correlation coefficient** is **positive** and **close to 1**, which is the largest possible value.

Covariance

Let X and Y be random variables whose standard deviation exist. Then the **covariance** of X and Y is defined by

$$\text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))).$$

Covariance

Let X and Y be random variables whose standard deviation exist. Then the **covariance** of X and Y is defined by

$$\text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))).$$

- **Calculating covariance:**

$$\text{cov}(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X)\mathbb{E}(Y).$$

- **Symmetry.** $\text{cov}(X, Y) = \text{cov}(Y, X)$.

- **Relationship with variance.** $\text{cov}(X, X) = D^2(X)$.

Covariance

Let X and Y be random variables whose standard deviation exist. Then the **covariance** of X and Y is defined by

$$\text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))).$$

- **Calculating covariance:**

$$\text{cov}(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X)\mathbb{E}(Y).$$

- **Symmetry.** $\text{cov}(X, Y) = \text{cov}(Y, X)$.
- **Relationship with variance.** $\text{cov}(X, X) = D^2(X)$.
- **Relationship with independence.** If random variables X and Y are **independent**, then $\text{cov}(X, Y) = 0$.

The other direction is not true: $\text{cov}(X, Y) = 0$ does not imply that X and Y are independent.

Properties of covariance

- Covariance with a constant. $\text{cov}(X, c) = 0$, if c is a real number.
- **Linearity.** We have

$$\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z),$$

and furthermore, for every $c \in \mathbb{R}$ we have

$$\text{cov}(c \cdot X, Y) = c \cdot \text{cov}(X, Y).$$

- **Variance of a sum.** $D^2(X + Y) = D^2(X) + D^2(Y) + 2\text{cov}(X, Y)$. In addition, we have

$$D^2\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n D^2(X_i) + 2 \sum_{i < j} \text{cov}(X_i, X_j).$$

- Variance of a difference. $D^2(X - Y) = D^2(X) + D^2(Y) - 2\text{cov}(X, Y)$.

Covariance: example

A store sells two products, A and B .

- Let X be the number of products A sold in a day;
- Let Y be the number of products B sold in a day.
- Suppose that X and Y are **independent**, have **Poisson distribution**, X has parameter 100, Y has parameter 180.
- Suppose that the price of product A is 300 forints, the price of product B is 400.

What is the covariance of the **total number of sold products** and the corresponding **income**?

Covariance: example

A store sells two products, A and B .

- Let X be the number of products A sold in a day;
- Let Y be the number of products B sold in a day.
- Suppose that X and Y are **independent**, have **Poisson distribution**, X has parameter 100, Y has parameter 180.
- Suppose that the price of product A is 300 forints, the price of product B is 400.

What is the covariance of the **total number of sold products** and the corresponding **income**? That is, what is $\text{cov}(X + Y, 300X + 400Y)$ equal to?

Is this value **positive**, **negative** or **0**?

Covariance: example

A store sells two products, A and B .

- Let X be the number of products A sold in a day;
- Let Y be the number of products B sold in a day.
- Suppose that X and Y are **independent**, have **Poisson distribution**, X has parameter 100, Y has parameter 180.
- Suppose that the price of product A is 300 forints, the price of product B is 400.

What is the covariance of the **total number of sold products** and the corresponding **income**? That is, what is $\text{cov}(X + Y, 300X + 400Y)$ equal to?

Is this value **positive**, **negative** or **0**?

The more products are sold, "probably" the larger the income is, hence we expect **positive covariance.**

Covariance: example

X and Y are **independent**, have **Poisson distribution**, X has parameter 100, Y has parameter 180. Then the covariance of **the number of sold products** and the **income**:

$$\begin{aligned}\text{cov}(X + Y, 300X + 400Y) &\stackrel{(a)}{=} \text{cov}(X, 300X) + \text{cov}(X, 400Y) + \\ &\quad + \text{cov}(Y, 300X) + \text{cov}(Y, 400Y) = \\ &\stackrel{(a,b)}{=} 300 \cdot \text{cov}(X, X) + 400 \cdot \text{cov}(Y, Y) = \\ &\stackrel{(b)}{=} 300D^2(X) + 400D^2(Y) = \\ &\stackrel{(c)}{=} 300 \cdot 100 + 400 \cdot 180 = 102000,\end{aligned}$$

where we used that

Covariance: example

X and Y are **independent**, have **Poisson distribution**, X has parameter 100, Y has parameter 180. Then the covariance of **the number of sold products** and the **income**:

$$\begin{aligned}\text{cov}(X + Y, 300X + 400Y) &\stackrel{(a)}{=} \text{cov}(X, 300X) + \text{cov}(X, 400Y) + \\ &\quad + \text{cov}(Y, 300X) + \text{cov}(Y, 400Y) = \\ &\stackrel{(a,b)}{=} 300 \cdot \text{cov}(X, X) + 400 \cdot \text{cov}(Y, Y) = \\ &\stackrel{(b)}{=} 300D^2(X) + 400D^2(Y) = \\ &\stackrel{(c)}{=} 300 \cdot 100 + 400 \cdot 180 = 102000,\end{aligned}$$

where we used that (a) covariance is **linear**;

Covariance: example

X and Y are **independent**, have **Poisson distribution**, X has parameter 100, Y has parameter 180. Then the covariance of **the number of sold products** and the **income**:

$$\begin{aligned}\text{cov}(X + Y, 300X + 400Y) &\stackrel{(a)}{=} \text{cov}(X, 300X) + \text{cov}(X, 400Y) + \\ &\quad + \text{cov}(Y, 300X) + \text{cov}(Y, 400Y) = \\ &\stackrel{(a,b)}{=} 300 \cdot \text{cov}(X, X) + 400 \cdot \text{cov}(Y, Y) = \\ &\stackrel{(b)}{=} 300D^2(X) + 400D^2(Y) = \\ &\stackrel{(c)}{=} 300 \cdot 100 + 400 \cdot 180 = 102000,\end{aligned}$$

where we used that (a) covariance is **linear**;

(b) **independent** random variables have covariance **0**, and the covariance of a random variable with itself is the same as its variance;

Covariance: example

X and Y are **independent**, have **Poisson distribution**, X has parameter 100, Y has parameter 180. Then the covariance of **the number of sold products** and the **income**:

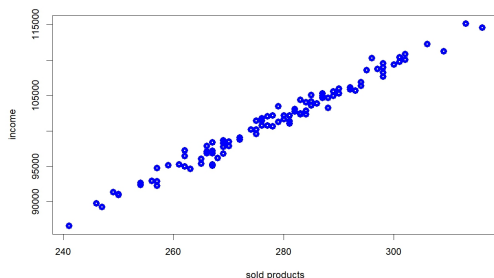
$$\begin{aligned}\text{cov}(X + Y, 300X + 400Y) &\stackrel{(a)}{=} \text{cov}(X, 300X) + \text{cov}(X, 400Y) + \\ &\quad + \text{cov}(Y, 300X) + \text{cov}(Y, 400Y) = \\ &\stackrel{(a,b)}{=} 300 \cdot \text{cov}(X, X) + 400 \cdot \text{cov}(Y, Y) = \\ &\stackrel{(b)}{=} 300D^2(X) + 400D^2(Y) = \\ &\stackrel{(c)}{=} 300 \cdot 100 + 400 \cdot 180 = 102000,\end{aligned}$$

where we used that (a) covariance is **linear**;

(b) **independent** random variables have covariance **0**, and the covariance of a random variable with itself is the same as its variance;

(c) a random variable with **Poisson distribution** and parameter λ has **variance** λ .

Covariance: example



Joint distribution of the income ($300X + 400Y$) and the number of sold copies ($X + Y$) in $n = 100$ observations.

Covariance: $\text{cov}(X + Y, 300X + 400Y) = 102000$.

Correlation coefficient: introduction

- **Covariance** aims to measure the **strength of dependence** for two random variables.
- In the above example, the covariance of the number of sold products and the income is **102000**.
- If income is counted in 1000 forints:

X : number of sold products Y : income in forints Z : income in 1000 forints

then

$$\text{cov}(X, Z) = \text{cov}\left(X, \frac{Y}{1000}\right) = \frac{\text{cov}(X, Y)}{1000} = 102.$$

That is, covariance **depends on the unit** \Rightarrow we introduce a similar quantity, which does not depend on the choice of the units.

- This is **correlation coefficient**.

Correlation coefficient

Let X and Y be random variables whose standard deviation exist. Then the (Pearson) **correlation coefficient** of X and Y is defined by

$$R(X, Y) = \begin{cases} \frac{\text{cov}(X, Y)}{D(X)D(Y)}, & \text{if } D(X) > 0, D(Y) > 0; \\ 0, & \text{if } D(X) = 0 \text{ or } D(Y) = 0. \end{cases}$$

Correlation coefficient

Let X and Y be random variables whose standard deviation exist. Then the (Pearson) **correlation coefficient** of X and Y is defined by

$$R(X, Y) = \begin{cases} \frac{\text{cov}(X, Y)}{D(X)D(Y)}, & \text{if } D(X) > 0, D(Y) > 0; \\ 0, & \text{if } D(X) = 0 \text{ or } D(Y) = 0. \end{cases}$$

- **Possible values.** The value of the correlation coefficient is always between -1 and 1 :

$$|R(X, Y)| \leq 1.$$

- **Linear dependence.** Let $a > 0$ and b be real numbers. Then we have

$$R(X, aX + b) = 1 \quad \text{and} \quad R(X, -aX + b) = -1.$$

- Suppose that $|R(X, Y)| = 1$. Then there exist real numbers a and b such that $Y = aX + b$ holds with probability 1. That is, the extreme values of R can be achieved in the case of linear dependence.

Covariance: example

A store sells two products, A and B .

- Let X be the number of products A sold in a day;
- Let Y be the number of products B sold in a day.
- Suppose that X and Y are **independent**, have **Poisson distribution**, X has parameter 100, Y has parameter 180.
- Suppose that the price of product A is 300 forints, the price of product B is 400.

What is the correlation coefficient of the **total number of sold products** and the corresponding **income**?

Covariance: example

A store sells two products, A and B .

- Let X be the number of products A sold in a day;
- Let Y be the number of products B sold in a day.
- Suppose that X and Y are **independent**, have **Poisson distribution**, X has parameter 100, Y has parameter 180.
- Suppose that the price of product A is 300 forints, the price of product B is 400.

What is the correlation coefficient of the **total number of sold products** and the corresponding **income**? That is, what is $\text{cov}(X + Y, 300X + 400Y)$ equal to?

$$\begin{aligned} R(X + Y, 300X + 400Y) &= \frac{\text{cov}(X + Y, 300X + 400Y)}{D(X + Y)D(300X + 400Y)} = \\ &= \frac{102000}{D(X + Y)D(300X + 400Y)} \end{aligned}$$

based on the previous calculation, hence we have to find the standard deviations.

Correlation coefficient: example

X and Y are **independent**, have **Poisson distribution**, X with parameter 100, Y with parameter 180. Then the standard deviation of the **number of sold products**:

Correlation coefficient: example

X and Y are **independent**, have **Poisson distribution**, X with parameter 100, Y with parameter 180. Then the standard deviation of the **number of sold products**:

$$D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{100 + 180} = 16.73.$$

The standard deviation of the income:

Correlation coefficient: example

X and Y are **independent**, have **Poisson distribution**, X with parameter 100, Y with parameter 180. Then the standard deviation of the **number of sold products**:

$$D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{100 + 180} = 16.73.$$

The standard deviation of the income:

$$\begin{aligned} D(300X + 400Y) &= \sqrt{300^2 D^2(X) + 400^2 D^2(Y)} = \\ &= \sqrt{300^2 \cdot 100 + 400^2 \cdot 180} = 6148.17. \end{aligned}$$

We get the correlation coefficient:

Correlation coefficient: example

X and Y are **independent**, have **Poisson distribution**, X with parameter 100, Y with parameter 180. Then the standard deviation of the **number of sold products**:

$$D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{100 + 180} = 16.73.$$

The standard deviation of the income:

$$\begin{aligned} D(300X + 400Y) &= \sqrt{300^2 D^2(X) + 400^2 D^2(Y)} = \\ &= \sqrt{300^2 \cdot 100 + 400^2 \cdot 180} = 6148.17. \end{aligned}$$

We get the correlation coefficient:

$$\begin{aligned} R(X + Y, 300X + 400Y) &= \frac{\text{cov}(X + Y, 300X + 400Y)}{D(X + Y)D(300X + 400Y)} = \\ &= \frac{102000}{16.73 \cdot 6148.17} = 0.9915. \end{aligned}$$

Correlation coefficient: example

X and Y are **independent**, have **Poisson distribution**, X with parameter 100, Y with parameter 180. Then the standard deviation of the **number of sold products**:

$$D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{100 + 180} = 16.73.$$

The standard deviation of the income:

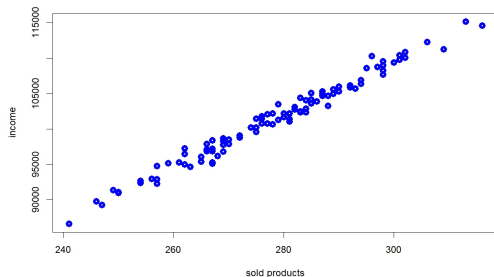
$$\begin{aligned} D(300X + 400Y) &= \sqrt{300^2 D^2(X) + 400^2 D^2(Y)} = \\ &= \sqrt{300^2 \cdot 100 + 400^2 \cdot 180} = 6148.17. \end{aligned}$$

We get the correlation coefficient:

$$\begin{aligned} R(X + Y, 300X + 400Y) &= \frac{\text{cov}(X + Y, 300X + 400Y)}{D(X + Y)D(300X + 400Y)} = \\ &= \frac{102000}{16.73 \cdot 6148.17} = 0.9915. \end{aligned}$$

The maximal value of R is **1**, hence this means **a strong positive correlation**.

Correlation coefficient: example



Joint distribution of the income ($300X + 400Y$) and the number of sold products ($X + Y$) in $n = 100$ independent observations. Covariance: **102000**, correlation coefficient: **0.9915**.

Correlation coefficient: example.

Example. As before, suppose that X and Y (the number of sold products of type A and B) are independent, have Poisson distribution, X with parameter 100, Y with parameter 180. Product A has price 300 forints, product B has price 4000. What is the correlation coefficient of the total number of sold products and the corresponding income?

Correlation coefficient: example.

Example. As before, suppose that X and Y (the number of sold products of type A and B) are independent, have Poisson distribution, X with parameter 100, Y with parameter 180. Product A has price 300 forints, product B has price **4000**. What is the correlation coefficient of the total number of sold products and the corresponding income?

$$\text{cov}(X + Y, 300X + 4000Y) = 300 \cdot 100 + 4000 \cdot 180 = 750000;$$

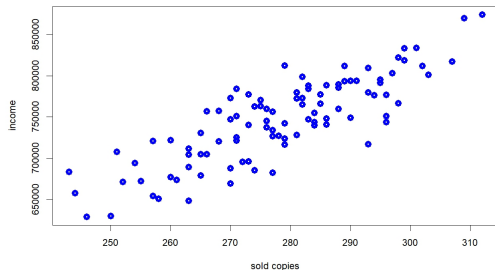
$$D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{100 + 180} = 16,73;$$

$$\begin{aligned} D(300X + 4000Y) &= \sqrt{300^2 D^2(X) + 4000^2 D^2(Y)} = \\ &= \sqrt{300^2 \cdot 100 + 4000^2 \cdot 180} = 53749,42; \end{aligned}$$

$$\begin{aligned} R(X + Y, 300X + 4000Y) &= \frac{\text{cov}(X + Y, 300X + 4000Y)}{D(X + Y)D(300X + 4000Y)} = \\ &= \frac{750000}{16.73 \cdot 53749.42} = 0.83. \end{aligned}$$

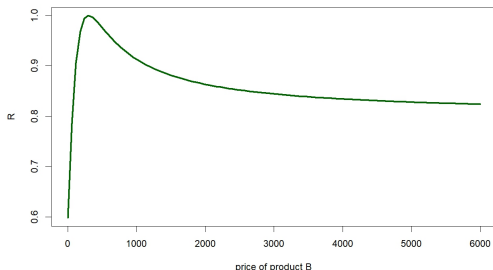
This is smaller than in the previous case, when the prices were closer to each other.

Correlation coefficient: example



The joint distribution of the income ($300X + 4000Y$) and the total number of sold products ($X + Y$) in $n = 100$ independent observations. Covariance: 750000, correlation coefficient: 0.83.

Correlation coefficient: example



Correlation coefficient of the income $(300X + cY)$ and the total number of sold products $(X + Y)$ as a function of c , where product A has price 300, product B has price c

Correlation and causality

- number of sunny hours and temperature:

Correlation and causality

- number of sunny hours and temperature: positive correlation, **causality** in one direction
- number of sunny hours and quantity of rain:

Correlation and causality

- number of sunny hours and temperature: positive correlation, **causality** in one direction
- number of sunny hours and quantity of rain: negative correlation, causality
- monthly income and highest educational degree:

Correlation and causality

- number of sunny hours and temperature: positive correlation, **causality** in one direction
- number of sunny hours and quantity of rain: negative correlation, causality
- monthly income and highest educational degree: positive correlation, causality in both directions
- time spent at the sea and health:

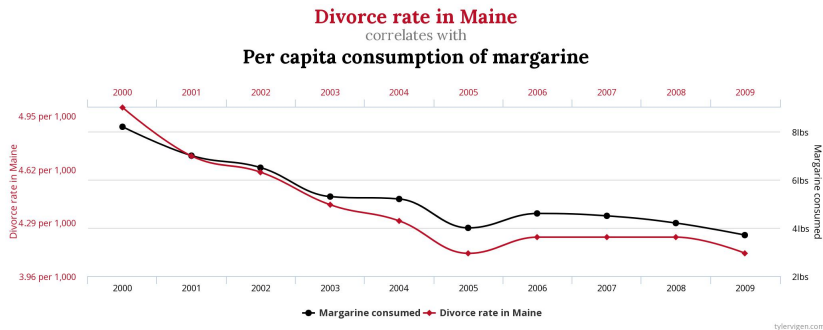
Correlation and causality

- number of sunny hours and temperature: positive correlation, **causality** in one direction
- number of sunny hours and quantity of rain: negative correlation, causality
- monthly income and highest educational degree: positive correlation, causality in both directions
- time spent at the sea and health:
there is probably a positive correlation, **but not with a strong causality**: both depend on the wealth status, we cannot say that time spent at the sea improves health directly (also ill people cannot go)
- number of divorces and consumption of margarine in Maine, USA:

Correlation and causality

- number of sunny hours and temperature: positive correlation, **causality** in one direction
- number of sunny hours and quantity of rain: negative correlation, causality
- monthly income and highest educational degree: positive correlation, causality in both directions
- time spent at the sea and health:
there is probably a positive correlation, **but not with a strong causality**: both depend on the wealth status, we cannot say that time spent at the sea improves health directly (also ill people cannot go)
- number of divorces and consumption of margarine in Maine, USA: strong positive correlation ($R = 0.9926$), but **no causality** (source and further examples:
<http://tylervigen.com/spurious-correlations>)

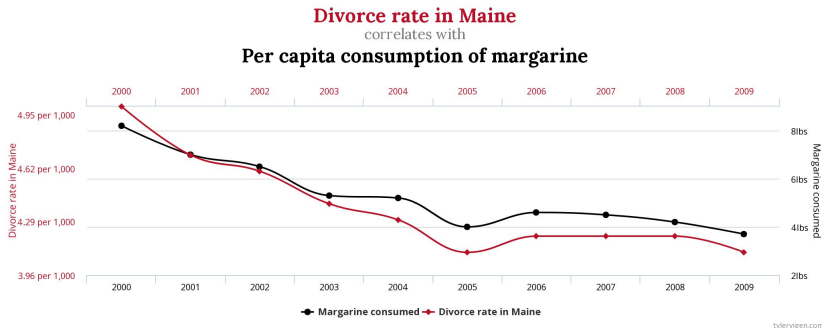
Correlation and causality



Number of divorces and consumption of margarine in Maine, USA; correlation coefficient: 0.9926

<http://tylervigen.com/spurious-correlations>

Correlation and causality



Number of divorces and consumption of margarine in Maine, USA; correlation coefficient: 0.9926

<http://tylervigen.com/spurious-correlations>

„Big data” analysis: among 200-300 quantities, it is easy to find a few which show very strong positive correlation without causality or independence; there might be other pairs with real connections → we need deeper analysis.

Uncorrelated random variables

If random variables X, Y have **covariance** 0, then we say that X and Y **uncorrelated**. What is the connection of this to **independence**?

X and Y **independent**

X and Y **uncorrelated**

Uncorrelated random variables

If random variables X, Y have **covariance** 0, then we say that X and Y **uncorrelated**. What is the connection of this to **independence**?

X and Y **independent**



X and Y **uncorrelated**

Let X and Y be the result of two independent fair dice rolls.

$U = X + Y$ is the sum

$V = X - Y$ is the difference

$$\text{cov}(U, V) = \text{cov}(X + Y, X - Y) = D^2(X) - \text{cov}(X, Y) + \text{cov}(X, Y) - D^2(Y) = 0 \Rightarrow X \text{ and } Y \text{ uncorrelated}$$

On the other hand, U and V are **not independent**:

Uncorrelated random variables

If random variables X, Y have **covariance** 0, then we say that X and Y **uncorrelated**. What is the connection of this to **independence**?

X and Y **independent**



X and Y **uncorrelated**

Let X and Y be the result of two independent fair dice rolls.

$U = X + Y$ is the sum

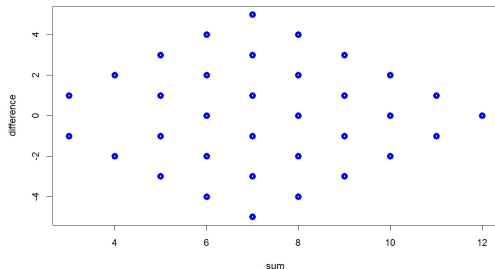
$V = X - Y$ is the difference

$$\text{cov}(U, V) = \text{cov}(X + Y, X - Y) = D^2(X) - \text{cov}(X, Y) + \text{cov}(X, Y) - D^2(Y) = 0 \Rightarrow X \text{ and } Y \text{ uncorrelated}$$

On the other hand, U and V are **not independent**:

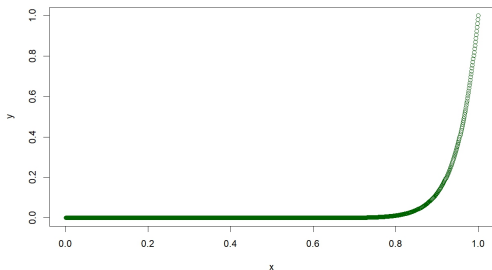
$$0 = \mathbb{P}(U = 11, V = 0) \neq \mathbb{P}(U = 11) \cdot \mathbb{P}(V = 0) = \frac{2}{36} \cdot \frac{1}{6}$$

Uncorrelated random variables



Joint distribution of the **difference** ($X - Y$) and **sum** ($X + Y$) of the values in 100 experiments. **Covariance: 0**, but $X + Y$ and $X - Y$ are **not independent**.

Rank correlation



Let X have uniform distribution on interval $[0, 1]$, and $Y = X^{20}$. The correlation coefficient is around 0.5, although there is a very strong deterministic dependence.

Rank correlation

Let $X_1, X_2, \dots, X_n, Y_1, \dots, Y_n$ be observations.

Rank correlation

Let $X_1, X_2, \dots, X_n, Y_1, \dots, Y_n$ be observations.

To each observation, let us consider its position in the sample if it was in decreasing order. More precisely, in sample X_1, X_2, \dots, X_n , the largest observation has rank 1, the second largest has rank 2, and so on. Similarly for the other sample.

For example:

$$X_1 = 650, X_2 = 870, X_3 = 720 \quad \Rightarrow \quad (3, 1, 2)$$

$$Y_1 = 18, Y_2 = 15, Y_3 = 17 \quad \Rightarrow \quad (1, 3, 2)$$

Rank correlation

Let $X_1, X_2, \dots, X_n, Y_1, \dots, Y_n$ be observations.

To each observation, let us consider its position in the sample if it was in decreasing order. More precisely, in sample X_1, X_2, \dots, X_n , the largest observation has rank 1, the second largest has rank 2, and so on. Similarly for the other sample.

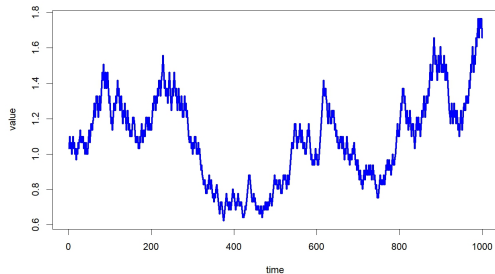
For example:

$$X_1 = 650, X_2 = 870, X_3 = 720 \quad \Rightarrow \quad (3, 1, 2)$$

$$Y_1 = 18, Y_2 = 15, Y_3 = 17 \quad \Rightarrow \quad (1, 3, 2)$$

Let us calculate the correlation coefficient of the two sequences of ranks. This is **rank correlation** (Spearman correlation).

Cumulative distribution function



Value of an imaginary stock in a period of 1000 days

Cumulative distribution function

- random variable X : value of a random experiment
- before: X **discrete**, and probabilities $\mathbb{P}(X = x)$ provide the distribution
- if the set of possible values is "too large", or probabilities are "too small", this is not informative
- for example: X is the price of this stock tomorrow, $\mathbb{P}(X = 784) = 0.0038$, $\mathbb{P}(X = 785) = 0.004$, etc. \rightarrow it can be more useful to say that

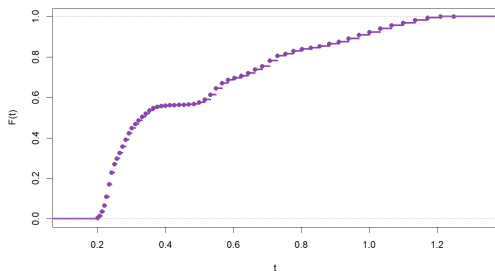
$$\mathbb{P}(X \leq 800) = 0.5,$$

that is, with probability 50%, the price is under 800

- **cumulative distribution function**: $F(t)$ is the probability that **the value of the random variable is at most t** , that is,

$$F(t) = \mathbb{P}(X \leq t).$$

Cumulative distribution function



the proportion of days with value at most t , as a function of t , in the previous example

Cumulative distribution function

Let $X : \Omega \rightarrow \mathbb{R}$ be a random variable. Then the **cumulative distribution function** of X is the function $F : \mathbb{R} \rightarrow [0, 1]$ defined as follows:

$$F(t) = \mathbb{P}(X \leq t) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq t\})$$

for every real number $t \in \mathbb{R}$.

This is well-defined for **every random variable** and every real number $t \in \mathbb{R}$: in the definition of a random variable, we supposed that $\{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{A}$, this set is an event, its probability is well-defined.