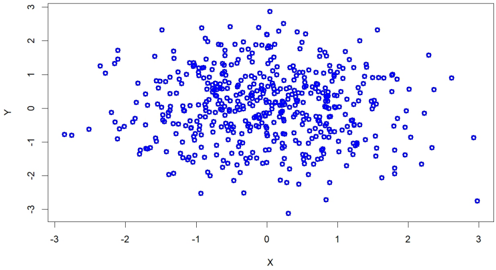
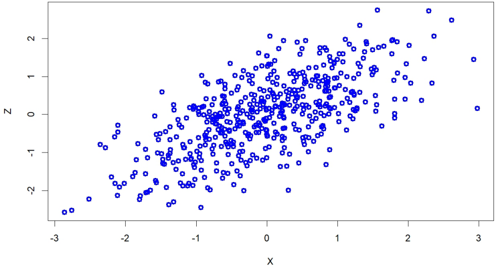
Independent random variables (Lecture 8)



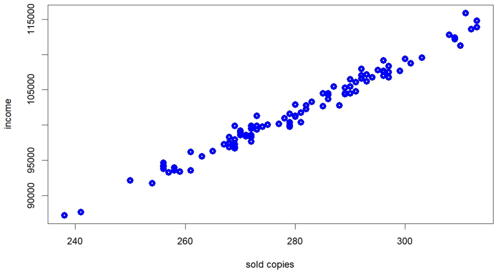
500 random points on the plane, whose coordinates follow **independent** standard normal distribution. Both the covariance and correlation coefficient will be equal to 0.

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.

Independence: example

Reminder: events A and B are **independent**, if

$$\mathbb{P}(\mathrm{A} \cap \mathrm{B}) = \mathbb{P}(\mathrm{A}) \cdot \mathbb{P}(\mathrm{B}).$$

If X is the quantity of rain in Buda tomorrow (in mm), and Y is the same in New York, then for events

$$\mathrm{A} : X \leq 5; \qquad \mathrm{B} : Y \leq 5$$

this condition means that

$$\mathbb{P}(X \leq 5, Y \leq 5) = \mathbb{P}(X \leq 5) \cdot \mathbb{P}(Y \leq 5).$$

That is, by assuming that the weather of the two cities are independent, the probability that **there will be at most 5 mm rain in both cities**, is the **product of the probabilities**.

Independence of random variables

- **for two random variables**: random variables $X, Y : \Omega \to \mathbb{R}$ are **independent**, if
$$\mathbb{P}(X \leq t_1, Y \leq t_2) = \mathbb{P}(X \leq t_1) \cdot \mathbb{P}(Y \leq t_2)$$
holds for every real numbers $t_1, t_2 \in \mathbb{R}$.
- **for finitely many random variables**: random variables $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ are **independent**, if
$$\begin{aligned}\mathbb{P}(X_1 \leq t_1, X_2 \leq t_2, \ldots, X_n \leq t_n) = \\ = \mathbb{P}(X_1 \leq t_1) \cdot \mathbb{P}(X_2 \leq t_2) \ldots \mathbb{P}(X_n \leq t_n)\end{aligned}$$
holds for every real numbers $t_1, t_2, \ldots, t_n$.
- **for countably many random variables**: the random variables $X_1, X_2, X_3 \ldots$ are **independent**, if we get independent random variables with every choice of finitely many from $X_1, X_2, \ldots$.

Independence in the discrete case

If the random variables are **discrete**, that is, their range is finite or countable infinite, then independence is equivalent to the following condition.

The **discrete** random variables X and Y are **independent** if and only if

for **every possible value** x_k **of** X and

for **every possible value** y_l **of** Y the following holds:

$$\mathbb{P}(X = x_k, Y = y_l) = \mathbb{P}(X = x_k) \cdot \mathbb{P}(Y = y_l) \qquad (k, l = 1, 2, \ldots).$$

That is, the probability that **the value of** X **is** x_k **and the value of** Y **is** y_l is equal to the **product of the correponding probabilities**.

Covariance

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

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

Covariance

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

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

- **Calculating covaiance:**

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

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

Covariance

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

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

- **Calculating covaiance:**

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

- **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.

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** $\Rightarrow$ / $\not\Leftarrow$ 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**

$\mathrm{cov}(U, V) = \mathrm{cov}(X + Y, X - Y) = D^2(X) - \mathrm{cov}(X, Y) +$
$+\mathrm{cov}(X, Y) - D^2(Y) = 0 \Rightarrow X$ and Y **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** $\Rightarrow$ / $\not\Leftarrow$ X and Y **uncorrelated**

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

$U = X + Y$ **is the sum**

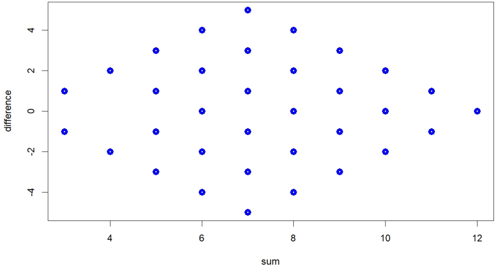
$V = X - Y$ **is the difference**

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

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

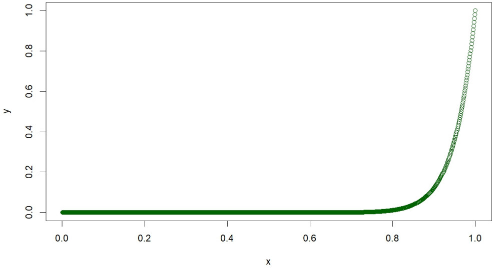
$$0 = \mathbb{P}(U = 11, V = 0) \neq \mathbb{P}(U = 11) \cdot \mathbb{P}(V = 0) = \frac{2}{36} \cdot \frac{1}{6}$$

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 $X_1, X_2, \ldots, X_n, Y_1, \ldots, Y_n$ be observations.

Rank correlation

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

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

For example:

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

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

Rank correlation

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

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

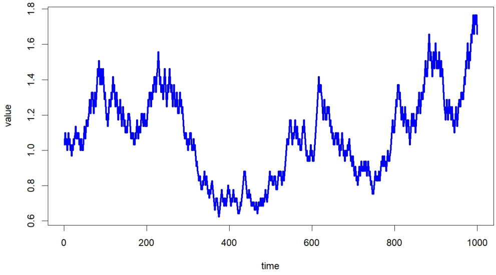
For example:

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

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

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

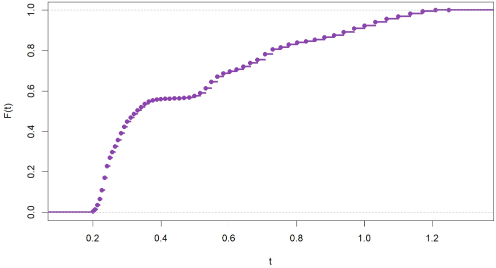
- random variable X: value of a random experiment
- before: X **discrete**, and probabilities $\mathbb{P}(X = x)$ provide the distribution
- if the set of possible values is "too large", or probabilities are "too small", this is not informative
- for example: X is the price of this stock tomorrow, $\mathbb{P}(X = 784) = 0.0038$, $\mathbb{P}(X = 785) = 0.004$, 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 $X : \Omega \to \mathbb{R}$ be a random variable. Then the **cumulative distribution function** of X is the function $F : \mathbb{R} \to [0, 1]$ defined as follows:

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

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.

Cumulative distribution function: example

We toss a fair coin three times. We have eight equally likely cases (elementary events):

$$\{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$$

Let X be the number of heads. Then X is a discrete random variable with range $\{0, 1, 2, 3\}$, and

$$\mathbb{P}(X = 0) = \frac{1}{8}, \quad \mathbb{P}(X = 1) = \frac{3}{8}, \quad \mathbb{P}(X = 2) = \frac{3}{8}, \quad \mathbb{P}(X = 3) = \frac{1}{8}.$$

Cumulative distribution function: example

We toss a fair coin three times. We have eight equally likely cases (elementary events):

$$\{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$$

Let X be the number of heads. Then X is a discrete random variable with range $\{0, 1, 2, 3\}$, and

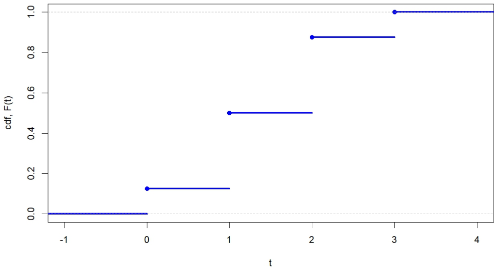
$$\mathbb{P}(X = 0) = \frac{1}{8}, \quad \mathbb{P}(X = 1) = \frac{3}{8}, \quad \mathbb{P}(X = 2) = \frac{3}{8}, \quad \mathbb{P}(X = 3) = \frac{1}{8}.$$

Let F be the **cumulative distribution function** of X. The value of F at a few points:

$$F(0) = \mathbb{P}(X \leq 0) = \frac{1}{8}; \qquad F(1) = \mathbb{P}(X \leq 1) = \frac{1}{2};$$

$$F(2,4) = \mathbb{P}(X \leq 2,4) = \frac{7}{8}; \qquad F(4) = \mathbb{P}(X \leq 4) = 1.$$

Cumulative distribution function: example



Cumulative distribution function of the number of heads out of three fair coin tosses horizontal: t, vertical: $F(t) = \mathbb{P}(X \leq t)$.

Cumulative distribution function

Definition (Cumulative distribution function)

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

$$F(t) = \mathbb{P}(X \leq t) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq t\}) \qquad \text{for every real number } t \in \mathbb{R}.$$

To every real number t, function F maps the probability that the value of the random variable is at most t. For example, if X is the number of heads out of three fair coin tosses:

$$F(1) = \mathbb{P}(X \leq 1) = \mathbb{P}(\text{at most 1 heads}) = 1/2;$$

$$F(2) = \mathbb{P}(X \leq 2) = \mathbb{P}(\text{at most 2 heads}) = 7/8;$$

$$F(2.3) = \mathbb{P}(X \leq 2.3) = \mathbb{P}(\text{at most 2.3 heads}) = 7/8.$$

Cumulative distribution function

Definition (Cumulative distribution function)

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

$$F(t) = \mathbb{P}(X \leq t) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq t\}) \qquad \textit{for every real number } t \in \mathbb{R}.$$

To every real number t, function F maps the probability that the value of the random variable is at most t. For example, if X is the number of heads out of three fair coin tosses:

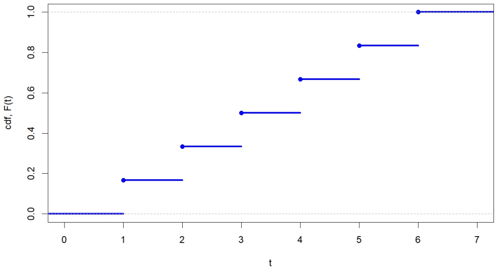
$$F(1) = \mathbb{P}(X \leq 1) = \mathbb{P}(\text{at most 1 heads}) = 1/2;$$

$$F(2) = \mathbb{P}(X \leq 2) = \mathbb{P}(\text{at most 2 heads}) = 7/8;$$

$$F(2.3) = \mathbb{P}(X \leq 2.3) = \mathbb{P}(\text{at most 2.3 heads}) = 7/8.$$

If X has a finite range, then its cumulative distribution function is a monotone step function (its range is finite), and the size of the "steps" are given by the probabilities of the possible values.

Cumulative distribution function: example



Cumulative distribution function of the value of a roll with a fair die
horizontal: t, vertical: $F(t) = \mathbb{P}(X \leq t)$.

Properties of cumulative distribution functions

If $a, b \in \mathbb{R}$ are real numbers, and F is the cumulative distribution function of X, then we have

$$\mathbb{P}(a < X \leq b) = F(b) - F(a).$$

Indeed, the probability that X is larger than a, but smaller than b, can be obtained by taking $\mathbb{P}(X \leq b)$, and subtracting $\mathbb{P}(X \leq a)$.

Let F be the cumulative distribution function of an arbitrary random variable. Then the following properties hold:

1. F is monotone increasing: if $a < b$, then $F(a) \leq F(b)$.
2. $\lim_{t\to-\infty} F(t) = 0; \qquad \lim_{t\to\infty} F(t) = 1.$
3. F is continuous from the right, that is, for every real number $t \in \mathbb{R}$ we have $\lim_{s\to t+} F(s) = F(t)$.

Characterization: if F satisfies the properties above, then there exists X, whose cumulative distribution function is F.

Uniform distribution

- We are waiting for a package, which is delivered at a random time Y.
- Suppose that Y has uniform distribution on the interval $[8, 12]$ (in hours).
- What is the probability that the package is delivered until 11?
- **Given that we did not receive the package until** 10, **what is the probability that we get it until** 11?

Uniform distribution

- We are waiting for a package, which is delivered at a random time Y.
- Suppose that Y has uniform distribution on the interval $[8, 12]$ (in hours).
- What is the probability that the package is delivered until 11?
- **Given that we did not receive the package until** 10, **what is the probability that we get it until** 11?

$$\mathbb{P}(X \leq 11)$$

Uniform distribution

- We are waiting for a package, which is delivered at a random time Y.
- Suppose that Y has uniform distribution on the interval $[8, 12]$ (in hours).
- What is the probability that the package is delivered until 11?
- **Given that we did not receive the package until** 10, **what is the probability that we get it until** 11?

$$\mathbb{P}(X \leq 11) = \frac{11-8}{12-8} = \frac{3}{4} = 75\%.$$

Uniform distribution

- We are waiting for a package, which is delivered at a random time Y.
- Suppose that Y has uniform distribution on the interval $[8, 12]$ (in hours).
- What is the probability that the package is delivered until 11?
- **Given that we did not receive the package until** 10, **what is the probability that we get it until** 11?

$$\mathbb{P}(X \leq 11) = \frac{11-8}{12-8} = \frac{3}{4} = 75\%.$$

$$\mathbb{P}(X \leq 11 | X > 10)$$

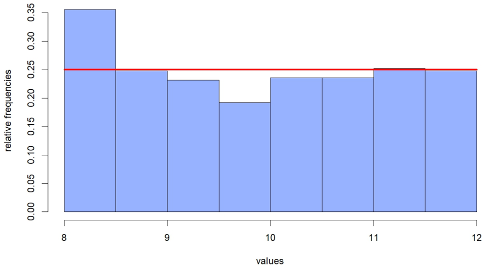
Uniform distribution

- We are waiting for a package, which is delivered at a random time Y.
- Suppose that Y has uniform distribution on the interval $[8, 12]$ (in hours).
- What is the probability that the package is delivered until 11?
- **Given that we did not receive the package until** 10, **what is the probability that we get it until** 11?

$$\mathbb{P}(X \leq 11) = \frac{11-8}{12-8} = \frac{3}{4} = 75\%.$$

$$\mathbb{P}(X \leq 11 | X > 10) = \frac{\mathbb{P}(\{X \leq 11\} \cap \{X > 10\})}{\mathbb{P}(X > 10)} = \frac{1/4}{2/4} = \frac{1}{2}.$$

Uniform distribution



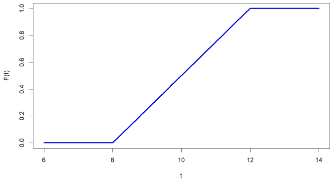
Histogram of a sample of size 500 from uniform distribution on the interval $[8, 12]$

Uniform distribution

Definition (Uniform distribution)

The random variable X has **uniform distribution** *on the interval $[a, b]$, if its distribution function is:*

$$F(t) = \mathbb{P}(X \leq t) = \begin{cases} 0, & \text{if } t \leq a; \\ \frac{t-a}{b-a}, & \text{if } a < t < b; \\ 1, & \text{if } t \geq b. \end{cases}$$



Uniform distribution

A package is delivered at a random time, which has uniform distribution between 10 and 12 (hours). We suppose that the time of delivery is uniformly distributed on the interval $[10, 12]$. Then the following hold ($a = 10, b = 12$):

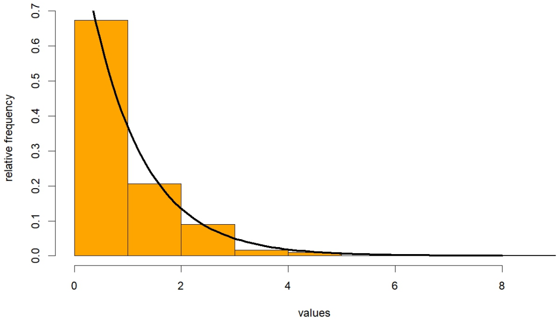
- The probability that the delivery is between 10 and 11: $(11 - 10)/(12 - 10) = 1/2$.
- The probability that the delivery is between $10:15$ and $10:30$: $1/8 = 12.5\%$.
- The probability that the delivery is after $10:30$: $3/4 = 75\%$.

Exponential distribution

Exponential distribution is often used to model random time intervals, for example,

- the time needed for a certain operation: service time of a customer, or an operation on a computer
- reaction time of a person
- time between the two occurences of an event, e.g. arrivals of two customers in a shop
- epidemic spread: time until recovery or infection
- radioactive decay: time until the decay of a particle

Exponential distribution



Density function of the exponential distribution with parameter $\lambda = 1$ and histogram of a sample of size 500 from the same distribution

Exponential distribution

Definition

Let $\lambda > 0$ be real number. Random variable X has **exponential distribution** *with parameter λ, if its cumulative distribution function is:*

$$F(t) = \mathbb{P}(X \leq t) = \begin{cases} 1 - e^{-\lambda t}, & \text{if } t > 0; \\ 0 & \text{otherwise.} \end{cases}$$