Optimization and Data Science

Lecture 14: Basic Stochastics and Statistics (1)

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

Stochastics and Statistics

- What is it?

  Important area in mathematics

  Provides theories and methods for data analysis
- Why are we studying this?

  Data can be regarded as results of random events
- How does it work?

  Using notions of random, random variables, probabilities

  Determining or estimating parameters of given data
- What if we can use it?

  Analyze data

  Detect "typical" behavior and outliers in data

  Provide tools to decide if data support a certain hypothesis

  Quantify uncertainties

Contents

Random events

- We consider a **random process**. Example:
  - Rolling the dice.
  - Randomly picking one picture out of a database of pictures.
  - Randomly choosing one index $j \in \{1, \ldots, m\}$.
- The result of a random process is called an **event**. Example:
  - Dice shows a "1".
  - One specific picture (e.g., number 27 in the database) is picked.
  - Index $j = 5$ is chosen.
- We denote by Ω the **set of all possible events**. Example:
  - $\Omega = \{1, 2, 3, 4, 5, 6\}$.
  - $\Omega = \{$ all pictures in the database $\} \triangleq \{1, \ldots, m\}$.
  - $\Omega = \{1, \ldots, m\}$.
- Ω can be discrete ($\Omega \subset \mathbb{Z}^m$, as in the above examples) ...
- ... or continuous ($\Omega \subset \mathbb{R}^m$).

Random variables and their realizations

Definition

- A **random variable** is a mapping

$$X : \{\text{set of all possible random events}\} =: \Omega \to \mathbb{R}.$$

- A **random vector** is a vector of random variables.
- If we consider one fixed random event $\omega \subset \Omega$, we call $X(\omega)$ the **realization** of the random variable. Often the notation

$$x := X(\omega), \quad \omega \in \Omega,$$

for the realization of a random variable is used.

- Throwing two dice is a **random event**.
- Mapping a throw onto the sum of the numbers the dice show, is a **random variable**.
- If we actually throw the dice and they show certain numbers, the sum of these numbers is the **realization** of the random value.

Random variables: Examples

- Rolling a dice: The value obtained is a random variable.
- Rolling two dice: The pair of both values is a two-dimensional random vector.
- Rolling two dice: The sum of both values is a random variable.
- Consider

$$\Omega = \begin{cases} \text{(all possible choices to take one picture out of)} \\ \text{a finite set of pictures with either cats or dogs.} \end{cases}$$

The mapping $X : \Omega \to \{0, 1\}$, defined as

$$X(\omega) := \begin{cases} 0, & \text{picture } \omega \in \Omega \text{ shows a cat} \\ 1, & \text{picture } \omega \in \Omega \text{ shows a dog} \end{cases}$$

is a random variable.

Discrete and continuous random variables

- A random variable is called **discrete** if it attains only countably many values $x_k \in \mathbb{R}, k \in \mathbb{N}$.
- If the set Ω is finite or countably infinite, then X is a discrete random variable.
  - The sets $\{0, 1\}, \{1, 2, 3, 4, 5, 6\}$ are finite and thus countable.
  - The sets $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ are countably infinite.
- All examples above were discrete random variables:
  - Rolling a dice: $X(\omega) \in \{1, 2, 3, 4, 5, 6\}$ for all $\omega \in \Omega$.
  - Rolling two dice: $X(\omega) \in \{(1, 1), (1, 2), \ldots, (6, 6)\}$ for all $\omega \in \Omega$.
  - Rolling two dice, sum: $X(\omega) \in \{2, \ldots, 12\}$ for all $\omega \in \Omega$.
  - Pictures of cats/dogs: $X(\omega) \in \{0, 1\}$ for all $\omega \in \Omega$.
- A random variable is called **continuous** if it may attain uncountably many values.
- If the set Ω is countably infinite, then X may be a discrete or a continuous random variable.
  - The sets $\mathbb{R}, \mathbb{R}^n, [0, 1], [0, 1]^n$ are uncountably infinite.

Random variable – Example: (Pseudo-) random number generator

- One way to generate (pseudo-)random numbers $x_k, k \in \mathbb{N}$, is:

$$a_0 \in \mathbb{N}, \quad a_{k+1} := ba_k + c \bmod m, k \in \mathbb{N}, \quad x_k := \frac{a_k}{m} \in [0,1], \tag{1}$$

  where b, c, m are fixed parameters.
- For fixed **random seed** a_0, the sequence is deterministic.
- If a_0 is taken randomly, each x_k in (1) is the realization of a random variable

$$X : \Omega \to [0,1].$$

- On a computer, a_0 is usually taken pseudo-randomly using the system clock etc.
- Thus, (1) actually does not realize a random variable.
- Even if a_0 is random, (1) can attain only countably many values.
- However, it is taken as an approximation of a **continuous** random variable, i.e., one that may attain all uncountably many values in $[0,1]$.

Random events and data science

- In data science, we often regard the given data as realizations of a random variable.
- Measurement results are often considered as random variables, since they usually contain measurement errors.
- Data from unknown or not exactly known processes are also considered as being random and are treated as random variables:
  - Consider data from social behavior (internet data)
  - data from social, medical, biological, pharmaceutical, ... experiments
  - natural phenomena, e.g. weather.
- Consider rolling dice: there are clear physical laws that determine this process, ...
- ... but the dependency on parameters (position, direction of throw, velocity) is too high.
- Thus, an exact computation is difficult ...
- ... and the modeling as random variable is used.
- Real random processes are rare in nature $\rightsquigarrow$ quantum physics, radioactive decay.

Contents

Mathematical assumption for the set of events

For the exact mathematical definition of probabilities, we need an abstract concept:

Definition (Sigma-Algebra)

Let $\Omega \neq \emptyset$. A system $\mathcal{A}$ of subsets of Ω, i.e., $\mathcal{A} \subset \mathcal{P}(\Omega)$ is called σ-**algebra** (on Ω) if

$$(1)\ \Omega \in \mathcal{A}, \qquad (2)\ A \in \mathcal{A} \Rightarrow A^c := \Omega \setminus A \in \mathcal{A}, \qquad (3)\ A_k \in \mathcal{A}, k \in \mathbb{N} \ \Rightarrow \bigcup_{k \in \mathbb{N}} A_k \in \mathcal{A}.$$

- The set $\{\{\}, \Omega\}$ is a trivial σ-algebra on Ω.
- The power set $\mathcal{P}(\Omega)$ (set of all subsets of Ω) is always a σ-algebra on Ω.
- Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ (rolling dice). The systems of subsets of Ω, given by

$$\mathcal{A} := \{\{\}, \Omega\} = \{\{\}, \{1, 2, 3, 4, 5, 6\}\},$$
$$\mathcal{A} := \mathcal{P}(\Omega) = \{\{\}, \{1\}, \ldots, \{1, 2\}, \ldots, \{3, 4, 5, 6\}, \ldots, \{2, 3, 4, 5, 6\}, \ldots, \{1, 2, 3, 4, 5, 6\}\}$$

are both σ-algebras on Ω. Clearly, the second one is more appropriate in this case.

Probability measures

Definition (Probability measure)

Let $\mathcal{A}$ be a σ-algebra on Ω. A mapping $P : \mathcal{A} \to [0, 1]$ is called **probability (measure)** if:

$$(1)\ P(\Omega) = 1, \qquad (2)\ P\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} P(A_i) \quad \text{for } A_i \in \mathcal{A}, A_i \cap A_j = \emptyset, i, j \in \mathbb{N}.$$

The triple $(\Omega, \mathcal{A}, P)$ is called **probability space**.

- Rolling dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$ with the σ-algebra $\mathcal{A} = \big\{\{\}, \{1, 2, 3, 4, 5, 6\}\big\}$. Then the setting $P(\{\}) = 0, P(\Omega) = 1$ defines a probability measure.
- Using $\mathcal{A} = \big\{\{\}, \{1\}, \ldots, \{6\}, \{1, 2\}, \ldots, \{1, 2, 3, 4, 5, 6\}\big\}$, the setting $P(\{1\}) = \ldots = P(\{6\}) = \frac{1}{6}$ and applying (2) from above defines a probability measure as well.

Probabilities associated with discrete random variables

A discrete random variable X can attain only countably many values $x_k, k \in \mathbb{N}$.

Definition

For a discrete random variable $X : \Omega \to \mathbb{R}$, the probability for X attaining a value $x \in \mathbb{R}$ is denoted by

$$P(X = x).$$

The function

$$F_X : \mathbb{R} \to [0,1], \quad F_X(x) := P(X \leq x) = \sum_{s \leq x} P(X = s)$$

is called **(cumulative) distribution function (cdf)** of X.

- Rolling a dice with $\mathcal{A} = \{\{\}, \{1\}, \ldots, \{6\}, \ldots, \{1,2,3,4,5,6\}\}$, taking the value of the dice as random variable. $P(X = 1) = \ldots P(X = 6) = \frac{1}{6}$. $F_X(2) = \sum_{s \leq 2} P(X \leq s) = P(X = 1) + P(X = 2) = \frac{1}{3}$.

Example: Probability of a discrete random variable, sum of two dice

- Set of events: $\Omega = \big\{\{1,1\},\{1,2\},\dots,\{6,6\}\big\}$, σ-algebra: $\mathcal{A} = \mathcal{P}(\Omega)$.
- Random variable: $X : \Omega \ni \omega = (\omega_1, \omega_2) \mapsto X(\omega) := \omega_1 + \omega_2$.
- Probabilities:

$$
\begin{aligned}
(\omega_1,\omega_2) = (1,1) \rightsquigarrow P(X=2) &= \frac{1}{36}, \quad (\omega_1,\omega_2) = (6,6) \rightsquigarrow P(X=12) = \frac{1}{36},\\
(\omega_1,\omega_2) \in \{(1,2),(2,1)\} \rightsquigarrow P(X=3) &= \frac{2}{36},\\
(\omega_1,\omega_2) \in \{(5,6),(6,5)\} \rightsquigarrow P(X=11) &= \frac{2}{36}, \dots\\
P(X=4) = P(X=10) &= \frac{3}{36} \quad P(X=5) = P(X=9) = \frac{4}{36}\\
P(X=6) = P(X=8) &= \frac{5}{36} \quad P(X=7) = \frac{6}{36}.
\end{aligned}
$$

Continuous random variables: probability density function

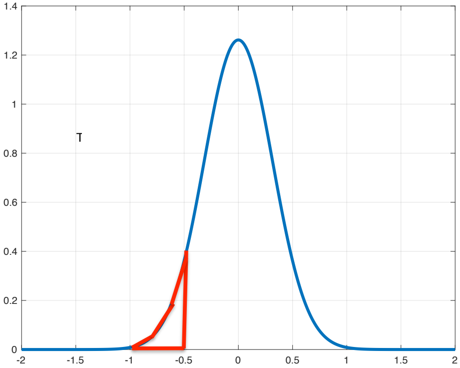
Definition

For a continuous random variable X, the probability for X attaining a value in $[a, b]$ is given by

$$P(a \leq X \leq b) := \int_a^b f_X(x)dx.$$

The function $f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is called the **probability density function (pdf)** of X. A probability density function is piecewise continuous and satisfies

$$\int_{-\infty}^{\infty} f_X(x)dx = 1.$$



Example: Gauss pdf,
area $= P(-1 \leq X \leq -0.5)$

Continuous random variables: cumulative distribution function

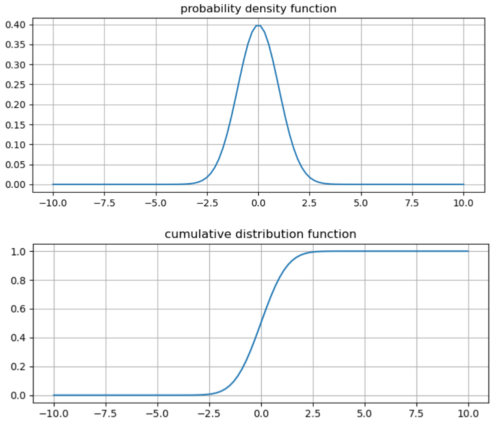
Definition

For a continuous random variable X, the function

$$F_X : \mathbb{R} \to [0,1],$$
$$F_X(x) := \int_{-\infty}^{x} f_X(s)\, ds = P(X \le x)$$

is called the **(cumulative) distribution function (cdf)**. A cumulative distribution function is non-decreasing and satisfies

$$\lim_{x\to-\infty} F_X(x) = 0, \qquad \lim_{x\to\infty} F_X(x) = 1.$$



Example: Gauss cdf and pdf

Contents

Expectation of a random variable

Definition (Expectation/Expected Value)

For a discrete random variable X, we define the **expectation** (or **expected value**) as

$$\mathbb{E}(X) := \sum_{k \in \mathbb{N}} x_k P(X = x_k),$$

and for a continuous random variable (with density f_X) as

$$\mathbb{E}(X) := \int_{\mathbb{R}} x \, f_X(x) dx.$$

For vectors and matrices of random variables, the expectation is defined component-wise.

- $\mathbb{E}(X)$ is the value that we "expect" as average of a high number of realizations of X.
- The expectation is linear: For random variables X_1, X_2 and $\alpha \in \mathbb{R}$, it holds

$$\mathbb{E}(\alpha X_1 + X_2) = \alpha \mathbb{E}(X_1) + \mathbb{E}(X_2).$$

Expectation of a random variable – Examples

- Rolling a dice:
$$\mathbb{E}(X) = (1+2+3+4+5+6)\frac{1}{6} = \frac{21}{6} = \frac{7}{2}.$$
- Sum of two dice, probabilities:
$$P(X=2) = P(12) = \frac{1}{36}, P(3) = \frac{2}{36} = P(11) = \frac{2}{36}, P(4) = P(10) = \frac{3}{36},$$
$$P(5) = P(9) = \frac{4}{36}, P(6) = P(8) = \frac{5}{36}, P(7) = \frac{6}{36}$$
$$\mathbb{E}(X) = \frac{2+12+2(3+11)+3(4+10)+4(5+9)+5(6+8)+6\cdot 7}{36} = \frac{18\cdot 14}{36} = 7.$$
Linearity of expectation: $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
- Uniform random number generator on $[0,1]$:

Expectation of a random variable – Examples

- Rolling a dice:
$$\mathbb{E}(X) = (1+2+3+4+5+6)\frac{1}{6} = \frac{21}{6} = \frac{7}{2}.$$
- Sum of two dice, probabilities:
$$P(X=2) = P(12) = \frac{1}{36}, P(3) = \frac{2}{36} = P(11) = \frac{2}{36}, P(4) = P(10) = \frac{3}{36},$$
$$P(5) = P(9) = \frac{4}{36}, P(6) = P(8) = \frac{5}{36}, P(7) = \frac{6}{36}$$
$$\mathbb{E}(X) = \frac{2+12+2(3+11)+3(4+10)+4(5+9)+5(6+8)+6\cdot 7}{36} = \frac{18\cdot 14}{36} = 7.$$
Linearity of expectation: $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
- Uniform random number generator on $[0,1]$:
$$\int_{\mathbb{R}} x\, f_X(x)dx = \int_0^1 x dx = \frac{1}{2}.$$

Variance of a random variable

Definition (Variance)

For a random variable X, the **variance** is defined as

$$\mathbb{V}(X) := \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$$

- $\mathbb{V}(X)$ is the expectation of the (squared) difference of the value of X and its expectation, i.e., the average deviation of X from its expectation for a high number of realizations.
- The variance is also called 2nd centered moment:

Definition (Moments)

For a random variable X, $\mathbb{E}(X^k)$ is called the k**-th moment**, and $\mathbb{E}\left((X - \mathbb{E}(X))^k\right)$ the k**-th centered moment**.

Variance of a random variable - Examples

- Rolling a dice: $\mathbb{E}(X) = \frac{7}{2}$:

$$\mathbb{V}(X) = \mathbb{E}\left(\left(X - \frac{7}{2}\right)^2\right) = \ldots$$

- gives:

Variance of a random variable - Examples

- Rolling a dice: $\mathbb{E}(X) = \frac{7}{2}$:

$$\mathbb{V}(X) = \mathbb{E}\left(\left(X - \frac{7}{2}\right)^2\right) = \dots$$

- gives:

$$\begin{aligned}
\mathbb{V}(X) &= \frac{1}{6}\sum_{x=1}^{6}\left(x - \frac{7}{2}\right)^2 = \frac{1}{6}\left(\left(1 - \frac{7}{2}\right)^2 + \cdots \left(6 - \frac{7}{2}\right)^2\right) \\
&= \frac{1}{6}\left(\left(-\frac{5}{2}\right)^2 + \left(-\frac{3}{2}\right)^2 + \left(-\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{3}{2}\right)^2 + \left(\frac{5}{2}\right)^2\right) \\
&= \frac{1}{3}\left(\frac{1}{4} + \frac{9}{4} + \frac{25}{4}\right) = \frac{35}{12}.
\end{aligned}$$

Variance of a random variable - Examples

- Sum of two dice: $\mathbb{E}(X) = 7$:

Variance of a random variable - Examples

- Sum of two dice: $\mathbb{E}(X) = 7$:

$$
\begin{aligned}
\mathbb{V}(X) &= \mathbb{E}\left((X-7)^2\right) \\
&= \frac{1}{36}\Big((2-7)^2 + 2(3-7)^2 + 3(4-7)^2 + 4(5-7)^2 + 5(6-7)^2 + 6(7-7)^2 \\
&\qquad + 5(8-7)^2 + 4(9-7)^2 + 3(10-7)^2 + 2(11-7)^2 + (12-7)^2\Big) \\
&= \frac{1}{36}(25 + 2\cdot 16 + 3\cdot 9 + 4\cdot 4 + 5 + 5 + 4\cdot 4 + 3\cdot 9 + 2\cdot 16 + 25) \\
&= \frac{1}{36}(50 + 64 + 54 + 32 + 10) \\
&= \frac{210}{36} = \frac{35}{6}
\end{aligned}
$$

Variance of a random variable - Examples

- Uniform random number generator on $[0,1]$: $\mathbb{E}(X) = \frac{1}{2}$

$$
\begin{aligned}
\mathbb{V}(X) &= \mathbb{E}\left(\left(X - \frac{1}{2}\right)^2\right) \\
&= \int_{\mathbb{R}} \left(x - \frac{1}{2}\right)^2 f_X(x)dx \\
&= \int_0^1 \left(x - \frac{1}{2}\right)^2 dx \\
&= \int_0^1 \left(x^2 - x + \frac{1}{4}\right) dx \\
&= \left[\frac{x^3}{3} - \frac{x^2}{2} + \frac{x}{4}\right]_{x=0}^{x=1} = \frac{1}{3} - \frac{1}{2} + \frac{1}{4} = \frac{4-6+3}{12} = \frac{1}{12}.
\end{aligned}
$$

Covariances and independence of random variables

Definition (Covariance (matrix))

For two random variables X, Y, the **covariance** is defined as

$$Cov(X, Y) := \mathbb{E}\big((X - \mathbb{E}(X))\,(Y - \mathbb{E}(Y)) \big)$$

For a random vector $(X_i)_{i=1}^n$, the **covariance matrix** is defined as

$$Cov(X) := \left(Cov(X_i, X_j)\right)_{i,j=1}^n .$$

- The covariance describes the interdependence of random variables.

Definition (Independence of random variables)

The random variables $X_i, i = 1, \dots, n$, are called **(mutually) independent**, if $Cov(X_i, X_j) = 0$ for all $i, j = 1, \dots, n, i \neq j$.

Correlation of random variables

Definition (Correlation (matrix))

For two random variables X, Y, the **correlation** is defined as

$$Cor(X,Y) := \frac{Cov(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$$

For a random vector $(X_i)_{i=1}^n$, the **correlation matrix** is defined as

$$Cor(X) := \left(\frac{Cov(X_i,X_j)}{\sqrt{\mathbb{V}(X_i)\mathbb{V}(X_j)}} \right)_{i,j=1}^n .$$

Random variables are called **uncorrelated (negative/positive correlated)** if their correlation is zero (lower/greater than zero).

- The correlation (matrix) is the normalized covariance (matrix). It has always values (elements) in $[-1, 1]$.

Rules for the Variance

- Linearity of $\mathbb{E}$ gives:

$$\mathbb{V}(X) = \mathbb{E}\left(X^2 - 2\mathbb{E}(X)X + \mathbb{E}(X)^2\right) = \mathbb{E}\left(X^2\right) - 2\mathbb{E}(X)^2 + \mathbb{E}(X)^2 = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2$$

$$\begin{aligned}
\mathbb{V}(\alpha X) &= \mathbb{E}\left((\alpha X - \mathbb{E}(\alpha X))^2\right) = \mathbb{E}\left((\alpha X - \alpha\mathbb{E}(X))^2\right) = \mathbb{E}\left(\alpha^2 (X - \mathbb{E}(X))^2\right) \\
&= \alpha^2 \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \alpha^2 \mathbb{V}(X), \quad \alpha \in \mathbb{R}, \\
\mathbb{V}(X+Y) &= \mathbb{E}\left((X + Y - \mathbb{E}(X+Y))^2\right) \\
&= \mathbb{E}\left((X - \mathbb{E}(X) + Y - \mathbb{E}(Y))^2\right) \\
&= \mathbb{E}\Big((X - \mathbb{E}(X))^2 + 2(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) + (Y - \mathbb{E}(Y))^2\Big) \\
&= \mathbb{E}\Big((X - \mathbb{E}(X))^2\Big) + 2\mathbb{E}\Big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\Big) + \mathbb{E}\Big((Y - \mathbb{E}(Y))^2\Big) \\
&= \mathbb{V}(X) + 2\,Cov(X,Y) + \mathbb{V}(Y).
\end{aligned}$$

- Rolling a dice: $\mathbb{V}(X) = \frac{35}{12} \leadsto$ Sum of two independent throws of a dice: $\mathbb{V}(X+Y) = \frac{35}{6}$.

Rule for Covariance

Lemma

Let $X = (X_i)_{i=1}^n$ be a random vector and $A \in \mathbb{R}^{m\times n}$. Then $Cov(AX) = A\,Cov(X)A^\top$.

Proof.

The covariance matrix can be written as

$$Cov(X) = \Big(\mathbb{E}((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j)))\Big)_{i,j=1}^n$$

$$= \mathbb{E}\left(\big((X_i - \mathbb{E}(X_i))\,(X_j - \mathbb{E}(X_j))\big)_{i,j=1}^n\right) = \mathbb{E}\Big(\overbrace{(X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top}^{\text{dyadic matrix product}}\Big)$$

This gives, using the linearity of the expectation:

$$Cov(AX) = \mathbb{E}\left((AX - \mathbb{E}(AX))\,(AX - \mathbb{E}(AX))^\top\right)$$

$$= \mathbb{E}\left((AX - A\,\mathbb{E}(X))(AX - A\,\mathbb{E}(X))^\top\right) = \mathbb{E}\left(A(X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top A^\top\right)$$

$$= A\,\mathbb{E}\left((X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top\right)A^\top = A\,Cov(X)A^\top.$$

Contents

Random samples

To perform a statistical analysis, a given set of data is regarded as a set of realizations of random variables with the same probability distribution.

Definition (Random sample)

A set of random variables $\{X_i, i = 1, \dots, n\}$ that

- are mutually independent
- and have the same cumulative distribution or (for a continuous random variable) density function

is called a **(random) sample** or **independent and identically distributed (iid)**.

→ Some authors call the **realizations of iid random variables** a random sample.

- Random variable: Age of the students in the course.
  Taking 10 students out of the group, we obtain a random sample.
- Random variable: Measurement of temperature at some location in the ocean at 0:00.
  Taking measurements at a number of days is a random sample.

What is important

- A random event is an event that is regarded as being not deterministic. The result of a random event can only be predicted with a probability.
- A random variable associates a real number with a random event.
- Data can be regarded as output of a random event, or as result of a random variable.
- We consider discrete and continuous random variables.
- For both, the distribution function describes the probability that the random variable attains a certain value or values in a certain range.
- For continuous random variables, the distribution can be (and is often) described by the probability density function (pdf).
- Important parameters of (the distribution of) a random variable are expectation, variance, and covariance.
- The covariance is used to define independence of random variables.
- A random sample is a set of independent and identically distributed random variables.