NMAI059 Probability and statistics
Lecture notes.

15. listopadu 2018

Aim of the notes. The aim of the lecture notes is not to replace the attendance of the lectures. The main goal is to summarize all definitions and theorems in correct form while the complementary text (remarks, examples, detailed discussion) is rather limited.

The definitions and theorems should be numbered as they are during the lecture. I hope this makes the study easier. Some parts of the text are typed using small size font. These parts are mostly complementary parts to the inevitably simplified text. Their role is to show mathematical background and context to those who are interested. These parts are not necessary for the exam.

1. Axioms; Probability space; Random events

1.1. **Axioms of probability.** Our everyday experience is that the results of our actions or experiments cannot be completely predicted. The result depends on many factors which cannot be completely measured or even observed. Therefore the result is considered to be *random*. Probability of the result is some measure of „frequency“: intuitively we believe that *repeating the same experiment under the same conditions (if the result of one experiment doesn't change the results of the other experiments) the relative ratio of the results is getting close to some value if the number of experiments is increasing to infinity.* This limiting ratio is (intuitively) the probability.

Such intuitive frequentist definition of probability is, however, hardly useful for mathematical theory. Rigorous mathematical model of probability was introduces by Andrei N. Kolmogorov in 1933.

Definition 1 (Probability space)**.** Let Ω be a nonempty set and let $\mathcal{F}$ be a system of subsets of Ω. We consider a probability measure P defined for all $F \in \mathcal{F}$ such that the mapping $\mathrm{P} : \mathcal{F} \to [0, 1]$ satisfies

(1) $\mathrm{P}(\Omega) = 1$
(2) for any pairwice disjoint $A_1, A_2, \ldots, \in \mathcal{F}$ it holds

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i).$$

(countable or σ-aditivity)

.

We have skipped the conditions for $\mathcal{F}$ so far. We may see in Definition 1:(2)that the union of pairwise disjoint sets form $\mathcal{F}$ should be again in $\mathcal{F}$. But for rigorous theory we need bit more. We need that $\mathcal{F}$ is a σ-algebra.

Definition (σ-algebra)**.** Nechť Ω je neprázdná množina. Třídu podmnožin Ω označená $\mathcal{F}$ nazveme *σ-algebrou* pokud

(1) $\emptyset \in \mathcal{F}$, $\Omega \in \mathcal{F}$.
(2) $A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$.
(3) $A_1, A_2, \dots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

So beside the countable union the closure to complements and $\Omega \in \mathcal{F}$ are needed conditions.

Definition 2 (Classical probability space)**.** Let Ω be nonempty finite set, $\mathcal{F} = 2^{\Omega}$. Define $P(A) = \frac{|A|}{|\Omega|}$ $\forall A \subset \Omega$. Then $(\Omega, \mathcal{F}, \mathrm{P})$ is called *classical probability space.*

Definition 3 (Discrete probability space)**.** Let Ω be nonempty finite or at most countable set, $\mathcal{F} = 2^{\Omega}$. Let $p : \omega \to \mathbb{R}$ be a function such that $p(\omega) \in [0, 1]$ $\forall \omega \in \Omega$, and $\sum_{\omega\in\Omega} p(\omega) = 1$. Define $\mathrm{P}(A) = \sum_{\omega\in A} p(\omega)$. Then $(\Omega, \mathcal{F}, \mathrm{P})$ is called *discrete probability space.*

Definition 4 (Real continuous probability space)**.** Let Ω be bounded or unbounded real interval, $\mathcal{F}$ contains all open and closed subintervals of Ω and their ciuntable unions and complements. Let $f : \omega \to \mathbb{R}$ be a function such that $f(\omega) \geq 0$, and $\int_{\Omega} f(\omega)\mathrm{d}\omega = 1$. For any $A \in \mathcal{F}$ define $\mathrm{P}(A) = \int_A f(\omega)\mathrm{d}\omega$. Then $(\Omega, \mathcal{F}, \mathrm{P})$ is called *real continuous probability space.*

The real interval Ω in the above definition may be an interval in k-dimensional reals space. In such case we need to replace the integral by the multiple integral in the definition.

1.2. **Theorems for computing the probability.**

- $\omega \in \Omega$ is called *elementary event.* $A \in \mathcal{F}$ is called *random event.*
- If $\mathrm{P}(A) = 1$, then A is *almost sure* event.
- If $\mathrm{P}(A) = 0$, then A is *null* event.
- If A is a random event then $A^C = \Omega \setminus A$ is *complementary event* of A.

Theorem 1 (Elemetary calculation)**.** *Let P be a probability measure defined on $\mathcal{F}$.*

(1) $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)\ \forall A \in \mathcal{F}$

(2) *If $A, B \in \mathcal{F} : A \subset B$ then* $\mathrm{P}(A) \leq \mathrm{P}(B)$ *a* $\mathrm{P}(B \setminus A) = \mathrm{P}(B) - \mathrm{P}(A)$

Důkaz. Immediately form the definition

(1) $\mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c)$. Since $A \cup A^c = \Omega$ then the properties of the probability P imply $\mathrm{P}(\Omega) = 1$, $\mathrm{P}(A) + \mathrm{P}(A^c) = 1$.

(2) Clearlz $A \subset B \Rightarrow B = A \cup (B \cap A^c)$ the union of two disjoint set. Hence $\mathrm{P}(B) = \mathrm{P}(A) + \mathrm{P}(B \setminus A) \geq \mathrm{P}(A)$.

□

Allprobability measures have one important property: continuity. As the continuity is used only for the proof of right continuity of distribution functions we give this definition only as a remark.

Consider a system of sets $\{A_i\}_{i=1}^{\infty} \subset \mathcal{F}$ such that $A_i \subset A_{i+1}$. Then we denote by $A_i \nearrow A$ the intersection $A = \bigcup_{i=1}^{\infty} A_i$ a and we say that the system $\{A_i\}_{i=1}^{\infty}$ converges monotonically to A.

The system $A_i \supset A_{i+1}$ converges monotonically to $A = \bigcap_{i=1}^{\infty} A_i$, what is denoted by $A_i \searrow A$. Note that in both cases $A \in \mathcal{F}$ since $\mathcal{F}$ is a σ-algebra.

Theorem (Continuity of probability measure)**.** *Let $\{A_i\}_{i=1}^{\infty} \subset \mathcal{F}$ be such that $A_i \searrow \emptyset$. Then*

$$\lim_{i \to \infty} \mathrm{P}(A_i) = 0.$$

Důkaz. Clearly $\mathrm{P}(\bigcap_{i=1}^{\infty} A_i) = \mathrm{P}(\emptyset) = 0$ which is equivalent to $1 - \mathrm{P}((\bigcap_{\imath=1}^{\infty} A_i)^c) = 1 - \mathrm{P}(\bigcup_{i=1}^{\infty} A_i^c) = 1$. Further $A_i^c \subseteq A_{i+1}^c$ foolows form $A_i \supseteq A_{i+1}$ and $A_i^c \nearrow \Omega$. Let us define a sequence $B_1 = A_1^c, B_{i+1} = A_{i+1}^c \setminus A_i^c$. Clearly B_i are disjoint such that $\bigcup_{i=1}^{\infty} B_i = \Omega$ and $\mathrm{P}(\bigcup_{i=1}^{\infty} B_i) = \mathrm{P}(\omega) = 1$. Therefore $\sum_{i=1}^{\infty} \mathrm{P}(B_i) = 1$. The last sum may be decomposed into two parts

$$1 = \underbrace{\sum_{i=1}^{n} \mathrm{P}(B_i)}_{\to 1} + \underbrace{\sum_{i=n+1}^{\infty} \mathrm{P}(B_i)}_{\to 0}$$

since P is bounded and σ-aditive. The definition of B_i and the properties of A_i give $\mathrm{P}(\bigcup_{i=1}^{n} B_i) = \mathrm{P}(\bigcup_{i=1}^{n} A_i^c) = 1 - \mathrm{P}(\bigcap_{i=1}^{n} A_i) = 1 - \mathrm{P}(A_n)$. Since the left hand side converges to 1 it follows that $\mathrm{P}(A_n) \to 0$. □

Theorem 2 (Inclusion adn exclusion principle)**.** *Let $A_1, \ldots, A_n$ be random events. Then*

$$\mathrm{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathrm{P}(A_i) - \sum_{1 \leq i \leq j \leq n} \mathrm{P}(A_i \cap A_j) + \sum_{1 \leq i \leq j \leq k\ \leq n} \mathrm{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} \mathrm{P}(\bigcap_{i=1}^{n} A_i)$$

Důkaz. Mathematical induction.

First step: for $n = 2$ clearly $A = (A \setminus B) \cup (A \cap B)$, $B = (B \setminus A) \cup (A \cap B)$, hence

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B).$$

Induction step $n - 1 \to n$:

$$\begin{aligned}
\mathrm{P}(\bigcup_{i=1}^{n} A_i) =& \mathrm{P}((\bigcup_{i=1}^{n-1} A_i) \cup A_n) = \mathrm{P}(\bigcup_{i=1}^{n-1} A_i) + \mathrm{P}(A_n) - \mathrm{P}(\bigcup_{i=1}^{n-1} (A_i \cap A_n)) \\
=& \sum_{i=1}^{n-1} \mathrm{P}(A_i) - \sum_{1 \leq i \leq j \leq n-1} \mathrm{P}(A_i \cap A_j) + \ldots + (-1)^{n-2} \mathrm{P}(\bigcap_{i=1}^{n-1} A_i) + \mathrm{P}(A_n) \\
& - \sum_{1 \leq i \leq j \leq n-1} \mathrm{P}(A_i \cap A_n) + \ldots + (-1)^{n-2} \mathrm{P}(\bigcap_{i=1}^{n} A_i).
\end{aligned}$$

Rearranging the sum and adding the appropriate terms together we get

$$\sum_{i=1}^{n} \mathrm{P}(A_i) - \sum_{1\leq i\leq j\leq n} \mathrm{P}(A_i \cap A_j) + \ldots + (-1)^{n-1}\mathrm{P}(\bigcap_{i=1}^{n} A_i)$$

□

The next theorems deal with an inportant concept of conditional probability.

Consider random event A occuring with the probability $\mathrm{P}(A)$. So before the experimant is made we know that the random event A will be the result of the experiment with probability $\mathrm{P}(A)$. Let us consider that there is an aditional information about the event B, namely that event B occurs in the experiment. May this information be used to improve our knowledge about the probability of the event A? Yes it may. An trivial example is a dice. Before we throw a dice the probability of 6 is 1/6. But if we know that the result is even the probability of 6 is 1/3 and if we know that the result is odd the probability of 6 is clearly 0. So the knowledge of parity of the result changes the probability of 6 substantially.

Definition 5 (Conditional probability)**.** Let $A, B \in \mathcal{F}, \mathrm{P}(B) > 0$ be a two random events. The conditional probability of A given B is $\mathrm{P}(A|B) = \frac{\mathrm{P}(A\cap B)}{\mathrm{P}(B)}$.

Problem. Show that the conditional probability satisfies the conditions for probability measure.

In general $\mathrm{P}(A|B \cup C) \neq \mathrm{P}(A|B) + \mathrm{P}(A|C)$. Find an example.

Note that $\mathrm{P}(A|\Omega) = \mathrm{P}(A)$. Clearly the information that Ω occurs brings no new information.

Theorem 3 (On product of probabilities)**.** *Let* $A_1, \ldots, A_n$ *be random events such that* $\mathrm{P}(\bigcap_{i=1}^{n} A_i) > 0$*. Then*

$$\mathrm{P}(\bigcap_{i=1}^{n} A_i) = \mathrm{P}(A_1|\bigcap_{i=2}^{n} A_i) \cdot \mathrm{P}(A_2|\bigcap_{i=3}^{n} A_i) \cdots \mathrm{P}(A_{n-1}|A_n) \cdot \mathrm{P}(A_n).$$

Důkaz. Mathematical induction. □

Definition 6 (Disjount partition)**.** Any finite or countable system of random events $\{B_i\}_{i\in I} \subset \mathcal{F}$ is called *disjoint partition* Ω if:

(1) $B_i \cap B_j = \emptyset \; \forall i \neq j$
(2) $\bigcup_i B_i = \Omega$ (or sufficiently $\mathrm{P}(\bigcup_i B_i) = 1$).
(3) $\mathrm{P}(B_i) > 0 \; \forall i$

Theorem 4 (On total probability)**.** *Let* A *be a random event and* $\{B_i\}$ *a disjoint partition. Then*

$$\mathrm{P}(A) = \sum_i \mathrm{P}(A|B_i) \cdot \mathrm{P}(B_i)$$

Důkaz. Clearly $A \cap B_i, i = 1, 2, \ldots$ are pairwise disjoint set. Hence $\bigcup_i A \cap B_i = A \cap \bigcup_i B_i = A \cap \Omega = A$ and it follows that

$$\mathrm{P}(A) = \mathrm{P}(\bigcup_i A \;\cap B_i) = \sum_i \mathrm{P}(A \cap B_i) = \sum_i \mathrm{P}(A|B_i) \cdot \mathrm{P}(B_i)$$

□

Theorem 5 (Bayes)**.** *Let* A *be a random event and* $\{B_i\}_i$ *a disjoint partition of* Ω*, and let moreover* $\mathrm{P}(A) > 0$*. Then*

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(A|B_i) \cdot \mathrm{P}(B_i)}{\sum_j \mathrm{P}(A|B_j) \cdot \mathrm{P}(B_j)}.$$

Důkaz. Immediately by the definition of conditional probability and the total probability theorem. □

Theorem 3 gives the exact probability of a simultaneous occurence of n random events. The computation of conditional probabilities in Theorem 3 may be sometimes too difficult. Therefore different estimates of the probabilities of complex events are used. One of them is the Bonferroni inequality.

Theorem 6 (Bonferroni inequality)**.** *Let $A_1, \dots, A_n$ be random events. Then*

$$P(\bigcap_{i=1}^{n} A_i) \geq 1 - \sum_{i=1}^{n}(1 - P(A_i))$$

Důkaz. $P(\bigcap_{i=1}^n A_i) = 1 - P\Big((\bigcap_{i=1}^n A_i)^C\Big) = 1 - P(\bigcup_{i=1}^n A_i^C) \geq 1 - \sum_{i=1}^n P(A_i^C) = 1 - \sum_{i=1}^n (1 - P(A_i))$. □

1.3. **Independence.** Independence (stochastical independence) of random events means that the occurence of some of the random events doesn't modify the probability of the other.

Definition 7 (Independence)**.** Two random events $A, B \in \mathcal{F}$ are *independent* if $\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B)$.

The definition of independence easily follows from the definition of conditional probability. If the events A, B are independent then

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A)\mathrm{P}(B)}{\mathrm{P}(B)} = \mathrm{P}(A)$$

and vice-versa this identity defines the independence. Independence may be defined also for null sets (which are so far excluded from the conditioning).

The independence must be defined for arbitrary number of random events.

Definition 8 (Mutual independence)**.** Let $A_1, \dots A_n$ be random events. Then A_i are *mutually independent* if it holds $\forall i_1, \dots, i_k \subset \{1, \dots, n\}$

$$\mathrm{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}).$$

Note that the equality must hold *for all subsets of indexes.*

Problem. Find a triple of random events A, B, and C such that all pairs form independent random events while $\mathrm{P}(A \cap B \cap C) \neq \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$.

It is sufficient to consider classical probability space $\Omega = \{\omega_1, \dots, \omega_n\}$ and find appropriate A, B, and C. What is the minimal sufficient number n?

Sometimes it is necessary to extend the notion of independence to *arbitrary* set of random events. However we cannot use infinite (even uncountable) intersections of sets in the definition of independence. The independence is correctly defined via all finite subsets of the index set as we shall see later in the definition of independence of random variables.

Note that the assumption of independence is quite strong and computation of probabilities of independent events is much simpler—try to apply the theorems above to independent random events.

2. Random variables and vectors

2.1. **Random variable and its distribution.** Random variable is a tool how to work with apriori unknown values which are results of experiments aith random outcome.

Definition 9. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. A mapping $X : \Omega \to \mathbb{R}$ such that $X^{-1}(-\infty, a] = \{\omega, X(\omega) \leq a\} \in \mathcal{F}\ \forall a \in \mathbb{R}$ is called *random variable.*

The definition of random variable is very general. Note that the condition on the pre-images and the properties of probability measure P allow to compute for *any* real interval the probability that the result (realisation) of the random variable will be the interval. This fact is reflected in the next definition.

This property of mapping is called *measurability* of the mapping X. Measurability allows to prove many different properties of the mapping although it is much weaker then, e.g., continuity. Nevertheless, when working woth only finite or countable number of possible outcomes of random variable the measurability is always satisfied. Hence, the measurability is not needed for the first reading.

Definition 10 (Distribution of random variable)**.** Let $X : \Omega \to \mathbb{R}$ be a random variable. Probability measure P_X defined on all open and cloed subsets of $\mathbb{R}$ such that for any $a \in \mathbb{R}$ it holds $P_X(-\infty, a] = \mathrm{P}[X \leq a]$ is called *the distribution of random variable* X.

Note that the probability P_X is a translation of the original probability P from the space Ω to the real line.

A random variable is a model of realisation of randomness in real numbers. It is important to note that **there may be several random variables with different or the same distribution defined on the same probability space**. One may use *canonical* construction of probabiolity space and random variables if only one or few random variables are needed, see below. However, in general it is more convenient to work with an abstract probability space when one needs to consider many different random variables.

Example. Canonical construction: consider

$$\Omega = \{1, 2, 3, 4, 5, 6\}^2, \mathcal{F} = 2^\Omega, \mathrm{P}(\{\omega\}) = \frac{1}{36}$$

and define

$$X_1(\omega_1, \omega_2) = \omega_1, X_2(\omega_1, \omega_2) = \omega_2, Y(\omega_1, \omega_2) = \omega_1 + \omega_2.$$

Then the distributions of X_1 and X_2 are the same while the random variables are not identical.

Find the distributions of X_i and Y. Since X_i have 6 different possible values, the distribution is defined as a probability measure on $2^6 = 64$ different sets. The distribution of Y is defined on $2^{11} = 2048$ different sets! Simpler description is needed for the distribution.

What are the sets $A \subset \mathbb{R}$ for which their probability P_X may be defined?

1. $(\infty, a]$
2. $(a, b], a < b = (-\infty, b] \setminus (-\infty, a]$
3. $(a, b) = \bigcup_{n=1}^{\infty} (a, b - \frac{1}{n}]$
4. All open sets
5. $A \in \mathcal{B} \Rightarrow X^{-1} \in \mathcal{F}$ for all $A \in \mathcal{B}$, X is *Borel measurable*

Definition (Random events generated by X)**.** Consider a random variable X and denote by $\mathcal{F}_X$ a system of sets such that $\mathcal{F}_X = \{B : B = X^{-1}(A) \text{ for some } A \in \mathcal{B}\}$.

Theorem. *The system $\mathcal{F}_x$ is the σ-algebra of random events generated by X.*

$\mathcal{F}_X$ is a σ-algebra, and $\mathcal{F}_X \subset \mathcal{F}$. Then

$$P_X(A) = \mathrm{P}[X \in A] = \mathrm{P}\big(\underbrace{X^{-1}(A)}_{\in \mathcal{F}_X \subset \mathcal{F}}\big).$$

The measure P_X is defined on $(\mathbb{R}, \mathcal{B})$ while the measure P is defined on $(\Omega, \mathcal{F})$.

It follows from Definition 10 that the (probability) distribution P_X of a random variable X is uniquelly defined by the probabilities of the sets $(-\infty, a]$ for all $a \in \mathbb{R}$. Hence we shall define distribution function of probabilities.

Definition 11. Let X be a random variable and P_X be its distribution. The function $F_X : \mathbb{R} \to [0, 1]$ defined by $F_X(x) = P_X[-\infty, x] = \mathrm{P}[X \leq x]$ is called *cummulative distribution function* (CDF) of the random variable X.

Cummulative distribution function F characterises the distribution P in the sense that $F_X = F_Y$ implies $P_X = P_Y$.

Theorem 7 (Properties of cdf)**.** *Consider a random variable X and its cdf F_X. Then*

1. $\lim_{x\to-\infty} F_X(x) = 0$
2. $\lim_{x\to\infty} F_X(x) = 1$
3. *F_X is nondecreasing and right continuous.*

Důkaz. Follows for the properties of probability measure.

The continuity of probability measure in empty set is needed for correct proof of the right continuity. So we shall just say that probability measure is continuous in empty set, see the remark above. □

There is also a reverse result. Any funciton which has the properties of cdf is a cdf of some random variable.

Theorem 8. *Assume that F satisfies the properties in Theorem 7. Then there is a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and a random variable X such that F is the cdf of X.*

To prove the theorem it is sufficient to construct some probability space and random variable X and show that F is the cdf of X.

Důkaz. Denote $\Omega = \mathbb{R}$, $\mathcal{F}$ the system containing all open sets, and their countable unions and intersections and complements. Define P a probability measure on $\mathcal{F}$ such that $P(-\infty, x] = F(x)$ (such measure is unique).

Define further $X : \Omega \to \mathbb{R}$ such that $X(\omega) = \omega$. Then $F_X(a) = P[X \leq a] = P(-\infty, a] = F(a)$. □

2.2. **Random vector and its distribution.** In what follows the notation

$$\boldsymbol{a} \leq \boldsymbol{b} \Leftrightarrow a_i \leq b_i \ \forall i = 1, 2, \ldots, d$$
$$\boldsymbol{a} < \boldsymbol{b} \Leftrightarrow a_i \leq b_i \ \forall i = 1, 2, \ldots, d, \text{ a existuje } j : a_j < b_j.$$

is used for any two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$.

Definition 12 (Random vector)**.** A mapping $\boldsymbol{X} : \Omega \to \mathbb{R}^d$ such that $\{\omega : \boldsymbol{X}(\omega) \leq \boldsymbol{a}\} \in \mathcal{F} \ \forall \boldsymbol{a} \in \mathbb{R}^d$ is called *random vector.* $(\Omega, \mathcal{F}, \mathrm{P})$ is some probability space and $d \in \mathbb{N}, d \geq 2$ is the dimension of the random vector.

Mathematically: Random vector must be measurable like the random variable is.

Definition 13 (Distribution and cummulative distribution function of random vector)**.** Let $\boldsymbol{X}$ be a d-dimensional random vector. Probability measurte $P_{\boldsymbol{X}}$ defined on (all open and closed subsets of) $\mathbb{R}^d$ such that

$$P_{\boldsymbol{X}}\big(\prod_{i=1}^{d}(-\infty, a_i]\big) = \mathrm{P}[\boldsymbol{X} \leq \boldsymbol{a}] = \mathrm{P}\big(\bigcap_{i=1}^{d}[X_i \leq a_i]\big)$$

is called the distribution of random vector $\boldsymbol{X}$.

The function $F_{\boldsymbol{X}} : \mathbb{R}^d \to [0,1]$ defined by $F_{\boldsymbol{X}}(\boldsymbol{a}) = P[\boldsymbol{X} \leq \boldsymbol{a}]$ is called *the cummulative distribution function* of random vector $\boldsymbol{X}$.

The probability measure $P_{\boldsymbol{X}}$ must be defined obviously also on all countable unions and intersections of open and closed sets, etc. Such extension is, however, unique, so the cummulative distribution function is sufficient to fully characterise the distribution of random vector. The cdf of random vectors may be still quite complex.

Notation: for any $\boldsymbol{a} < \boldsymbol{b}$ denote by $\Delta_k(\boldsymbol{a}, \boldsymbol{b})$ the set of all $\boldsymbol{c}$ such that there is exactly k indexes $i_1, i_2, \ldots, i_k$ satisfying $c_{i_j} = a_{i_j}$, and $c_l = b_l$ for the other indexes.

For example, let $\boldsymbol{a} = (a_1, a_2, a_3), \boldsymbol{b} = (b_1, b_2, b_3)$ then $\Delta_1(\boldsymbol{a}, \boldsymbol{b}) = \{(a_1, b_2, b_3), (b_1, a_2, b_3), (b_1, b_2, a_3)\}$.

Theorem 9 (Properties of cdf)**.** *Let $F_{\boldsymbol{X}}$ be the cdf of a random vector $\boldsymbol{X}$. Then:*

(1) $\lim_{a_i \to -\infty} F_{\boldsymbol{X}}(\boldsymbol{a}) = 0$ *for any i.*
(2) $\lim_{a_i \to \infty \ \forall i} F_{\boldsymbol{X}}(\boldsymbol{a}) = 1$.
(3) *The cdf $F_{\boldsymbol{X}}$ is rigth continuous and non-decreasing in all arguments.*
(4) $\forall \boldsymbol{a} < \boldsymbol{b}$ *it holds* $\sum_{k=0}^{d}(-1)^k \sum_{\boldsymbol{c} \in \Delta k(\boldsymbol{a}, \boldsymbol{b})} F_{\boldsymbol{X}}(\boldsymbol{c}) \geq 0$.

What does the last property mean in two dimensions? Consider $\boldsymbol{a} = (a_1, a_2), \boldsymbol{b} = (b_1, b_2)$. Then if $\boldsymbol{a} < \boldsymbol{b}$ it holds that $F_{\boldsymbol{X}}(b_1, b_2) - F_{\boldsymbol{X}}(a_1, b_2) - F_{\boldsymbol{X}}(b_1, a_2) + F_{\boldsymbol{X}}(a_1, a_2) \geq 0$. Clearly this corresponds to probability of the rectangle $(a_1, b_1] \times (a_2, b_2]$ and probability must be clearly non-negative.

For random vectors the theorem may be also reversed in the sense that any function satisfying (1)–(4) of Theorem 9 is the cdf of some random vector. Note that property (4) doesn't follow from (3) and is crucial. A function satisfying just (1)–(3) needs not to correspond to any probability measure since it may lead to negative probability.

Definition 14 (Marginal distribution)**.** Let $\boldsymbol{X}$ be a random vector, and $P_{\boldsymbol{X}}$ its distribution. The distribution $P_{X_i}(-\infty, a] = \lim_{a_j \to \infty, j \neq i} P_{\boldsymbol{X}}(X_{j=1}^{d}(-\infty, a_j])$ is called the *marginal distribution* of X_i, and $F_{X_i} = \lim_{a_j \to \infty, j \neq i} F_{\boldsymbol{X}}(\boldsymbol{a})$ is called the marginal cdf of X_i.

Terminology: A random vector (variable) $\boldsymbol{X}$, is *discrete*, if it attains up to countable many possible value only. There exists at most countable set $\mathbb{S}$ and non-negative values $p_{\boldsymbol{s}}, \boldsymbol{s} \in \mathbb{S}$ such that $\mathrm{P}[\boldsymbol{X} = \boldsymbol{s}] = p_s$, and $\mathrm{P}[\in A] = \sum_{\boldsymbol{s} \in A} p_{\boldsymbol{s}}$. In other words $\boldsymbol{X}(\omega) \in \mathbb{S}\ \forall \omega$.

A random vector (variable) $\boldsymbol{X}$ is *absolutely continuous* if for any $\boldsymbol{a}$ it holds $\mathrm{P}[\boldsymbol{X} = \boldsymbol{a}] = 0$, and if there exists a non-negative function f such that $\mathrm{P}[\boldsymbol{X} \in A] = \int_A f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$.

Agreemnet: For the sake of simplicity we shall assume that a discrete random vector attains values in a subset of $\mathbb{N}_0^d$ if not stated otherwise.

The distribution of a discrete random vector $\boldsymbol{X}$ is fully characterised by the set of values $\{p_{\boldsymbol{s}}\}_{\boldsymbol{s} \in S}$ while the distribution of a (absolutely) continuous random vector is fully characterised y the function f. Both the values $p_{\boldsymbol{s}}$ (for dicrete r.v.) and the function f (for continuous r.v.) are called *probability density function* (pdf) of the random vector $\boldsymbol{X}$.

Problem. Clearly

$$\sum_{\boldsymbol{x} \in \mathbb{S}} p_{\boldsymbol{x}} = 1, \quad \int_{\mathbb{R}^d} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1.$$

Note that we are working with multiple integrals. Multiple integrals are similar to multiple sums in the sense that the most inner integral must be evaluated first and the most outer integral is the last to be calculated.

Also note that the density f may be changed in finitely many púoints without changing the distribution.

Example. Basic discrete distributions of random variable:

(1) Alternative (Bernoulliho). X is $\{0, 1\}$ valued, $\mathrm{P}[X = 1] = P_x(\{1\}) = p \in (0, 1)$, and $\mathrm{P}[X = 0] = 1 - p$. Parameter p is interpreted as the probablity of success. The distribution is usually denoted $\mathrm{Alt}(p)$.

(2) Binomiial. X is the number of successes in n independent trials (result of one trial doesn't influence the other trials) with the same probability p of success. $\mathrm{P}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}$, the distribution is denoted $\mathrm{Bi}(n, p)$.

(3) Geometric. X is the number of failures up to the first success in independent trials with the same probability p of success. $\mathrm{P}[X = k] = p(1 - p)^k$, this distribution is denoted $\mathrm{Geom}(p)$.

(4) Poisson. X is the number of events appearing in a unit time interval. $\mathrm{P}[X = k] = \exp(-\lambda) \lambda^k / k!$, where $\lambda > 0$, $k = 0, 1, \ldots$. This distribution is denoted $\mathrm{Po}(\lambda)$

Example. Basic continuous distributions of random variable:

(1) Uniform distribution on the interval (a, b), $-\infty < a < b < \infty$. The pdf is

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & x \notin (a, b). \end{cases}$$

This distribution is denoted $\mathrm{U}(a, b)$ and it models random variable with values in the given interval without preferring any subinterval (all subitervals of the same length have the same probability).

(2) Exponential distribution is defined by the pdf

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & x > 0 \\ 0 & x < 0, \end{cases}$$

where $\lambda > 0$ is the parameter of the distribution. This is the basic model for "time to event" random variable. It is denoted $\mathrm{Exp}(\lambda)$.

(3) Gaussian (normal) distribution is characterised by the pdf

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2} \right), x \in \mathbb{R}.$$

The parameters are $\mu \in \mathbb{R}$ (location), and $\sigma^2 > 0$ (scale). It is denoted $\mathrm{N}(\mu, \sigma^2)$.

The Gaussian distribution is called "standard" if $\mu = 0$, and $\sigma^2 = 1$. The name emphasizes the importance of the distribution in probability and statistics.

Example. The basic one discrete and continuous distribution of random vectors are the multinomial and the multivariate normal distributions.

(1) Multinomial distribution is a generalisation of the binomial distribution. Let n denotes the number of trials in which one of the k possible outcomes occurs where the outcome i occurs with the probability p_i, $\sum_{i=1}^k = 1$ and the trials are independent. The random vector $\boldsymbol{X} = (X_1, X_2, \dots, X_k)$ contains the numbers of the respective results in n trials and its distribution is given by the pdf

$$\mathrm{P}[\boldsymbol{X} = (n_1, n_2, \dots, n_k] = \begin{cases} \frac{n!}{n_1!n_2!\dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} & n_i \in \mathbb{N}_0, \sum_{i=1}^k n_i = n \\ 0 & \text{else.} \end{cases}$$

This distribution is denoted by $\mathrm{Mult}(n, p_1, \dots, p_k)$.

(2) The multivariate (here the d-variate) Gaussian or normaldistribution is given by the pdf

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right), \quad \boldsymbol{x} \in \mathbb{R}^d.$$

The parameters of the distribution are $\mu \in \mathbb{R}^d$, and Σ a symmetric positive definite $d \times d$ matrix. The distribution is denoted $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$.

In what follows we shall restrict to bivariate random vectors since everything may be easily extended to higher dimensions. The bivariate normal distribution is an important example.

Example (Bivariate normal distribution). Let us show three basic version of this distribution.

(1) *Standard normalised* bivariate normal (Gaussian) distribution is given by $\boldsymbol{\mu} = \mathbf{0}$, and $\Sigma = I_2$ the identity matrix. The pdf is very simpple, namely

$$f(\boldsymbol{x}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1^2 + x_2^2) \right).$$

(2) If $\boldsymbol{\mu} = \mathbf{0}$, and the matrix Σ is

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \text{ kde } |\rho| < 1$$

then the distribution is called also normalised (but not standard) and the pdf is

$$f(\boldsymbol{x}) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)} \right).$$

(3) The general case is given by the pdf

$$f(\boldsymbol{x}) = \frac{1}{2\pi\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x_1 - \mu_1, x_2 - \mu_2) \Sigma^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right).$$

2.3. **Independence, random sample and empirical distribution.** The following theorem is only a simple observation.

Theorem 10 (On marginal distribution)**.** *Consider a random vector $\boldsymbol{X}$ and $P_{\boldsymbol{X}}$ be its distribution. Then all marginal distributions of the coordinates $X_1, \dots, X_d$ are uniquelly determined by the distribution $P_{\boldsymbol{X}}$.*

Důkaz. Immediate. □

Problem. The joint distribution **is not** given by the marginals.

Throw two dices. The results are two random variables A, B. Define further $C = A - 1$ if B sudé a $C = A + 1$ pro B liché, where 0 is replaced by 6 and 7 is replaced by 1. The joint distributions of (A, B), and of (A, C) are:

$A \backslash B$	1	2	3	4	5	6	
1	1/36	1/36	1/36	1/36	1/36	1/36	$\frac{1}{6}$
2	1/36	1/36	1/36	1/36	1/36	1/36	$\frac{1}{6}$
3	1/36	1/36	1/36	1/36	1/36	1/36	$\frac{1}{6}$
4	1/36	1/36	1/36	1/36	1/36	1/36	$\frac{1}{6}$
5	1/36	1/36	1/36	1/36	1/36	1/36	$\frac{1}{6}$
6	1/36	1/36	1/36	1/36	1/36	1/36	$\frac{1}{6}$
	1/6	1/6	1/6	1/6	1/6	1/6	

$A \backslash C$	1	2	3	4	5	6	
1	0	1/12	0	0	0	1/12	$\frac{1}{6}$
2	1/12	0	1/12	0	0	0	$\frac{1}{6}$
3	0	1/12	0	1/12	0	0	$\frac{1}{6}$
4	0	0	1/12	0	1/12	0	$\frac{1}{6}$
5	0	0	0	1/12	0	1/12	$\frac{1}{6}$
6	1/12	0	0	0	1/12	0	$\frac{1}{6}$
	1/6	1/6	1/6	1/6	1/6	1/6	

Hence it si clear that the marginal distributions of (A, B), and (A, C) are the same while the joint distributions are completely different.

There is a beautiful mathematical theorem that any multivariate joint cdf is combined form its marginal cdf's by a function called *copula*. Such copula is unique for continuous distributions and unique up to the null sets for discrete distributions. Namely, if $F_{\boldsymbol{X}}$ is a joint cdf, and $F_{X_i}, i = 1, \ldots, d$ its marginal cdf's then there is (in some sense unique) function $C : [0, 1]^d \to [0, 1]$ such that

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = C\big(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)\big).$$

The function C is a cdf itself, and all marginal cdf's of C are uniform on $(0, 1)$. There is a lot of interest and research related to copulas.

The joint distribution of a random vector $\boldsymbol{X}$ completely determines the *stochastic* relation between the components of the vector. There is a special case of the relation for which the joint distribution is uniuqelly given by the marginals.

Definition 15 (Independence of random variables)**.** Let $X_1, X_2, \ldots, X_k$ be random variables defined on $(\Omega, \mathcal{F}, \mathrm{P})$. These variables are (stochasticaly) *independent* if

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{k} F_{X_i}(x_i)\ \forall \boldsymbol{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k$$

holds for $\boldsymbol{X} = (X_1, \ldots, X_k)$.

Theorem 11 (Distribution of independent random variables)**.** *Random variables $X_1, \ldots, X_k$ are independent iff*

$$\mathrm{P}\left(\bigcap_{i=1}^{k} [X_i \in A_i]\right) = \prod_{i=1}^{k} P[X_i \in A_i] \quad \forall A_1, \ldots, A_k.$$

Důkaz. Immediately from the characterisation of the distribution by the cdf. □

We should restrict to $A_i \in \mathcal{B}$ only since we must work with measurable sets.

Theorem 12 (Equivalent conditions for independence)**.** *Discrete random variables $X_1, \ldots, X_k$ are independent iff $p_{\boldsymbol{X}}(a_1, \ldots, a_k) = \prod_{i=1}^{k} p_{X_i}(a_i)$.*

Continuous random variables $X_1, \ldots, X_k$ are independent iff $f_{\boldsymbol{X}}(x_1, \ldots, x_k) = \prod_{i=1}^{k} f_{X_i}(x_i)$.

Důkaz. Foíllows form the definition of probability density function and its relation to cdf. □

Problem: Model of distribution of random variable or vector is a theoretical description and simplification. The model may be very precise (like binomial distribution if the independence and identical success probability p are—somehow—ensured for al trials) some models are just „reasonably good“.

We usually try to fit, propose or test model based on some empirical observations. Of course, we must get some empirical experience before. This may be achieved if the random event (trial, experiment) is observed under the "same" conditions repeatedly.

The basic setting for observations is *random sample*. The essential feature of a random sample-Jeho podstatou is that the same „random phenomenon“ is observed in independent trials which allow generalisation (mathematical induction).

Definition 16 (Random sample)**.** The sequence $X_1, X_2, \ldots, X_n$ of independent random variables or vectors such that the distribution of each X_i is the same distribution P is called random sample from the distribution P of sample size n. We shartly say that $X_1, \dots, X_n$ are iid (independent and identically distributed).

Compare the independence of random events and independence of random variables. For random events we need to check that the probability of intersection is the product of probabilities for all subsets of indexes. This requirement (all subsets of indexes) is replaced by the fact that we need equality for the joint cdf and product of marginal distribution functions in *all* points.

We may need to define independence of infinitely many random variables. We say that random variables (indexed by arbitrary set) are independent iff for any *finite subset of indexes* the corresponding random variables are independent. We shall need only one special case, in particular the independence of $X_1, X_2, \dots$. This is equivalent to independence of $X_1, \dots, X_n$ for all finite n.

The random sample is the basic stone for model selection, estimates of distributions, characteristics, parameters, etc. Especially the estimate of the cummulative distribution function is simlpe and straightforward.

Definition 17 (Empirical distribution function)**.** Let $X_1, \dots, X_n$ be a random sample from distribution P and with the cdf F. Then we define *empirical distribution function* (edf) as $\widehat{F_n}(x) = \frac{1}{n}\sum_{i=1}^n \chi(X_i \le x)$, where χ is the indicator of a set.

The empirical distribution function is quite good estimator of the real (and typically unknown) cummulative distribution function. We need more characteristics of random variables before we may specify what *good estimate* means.

Remark. Note that the edf is a sum of random variables, hence it is random variable itself (you should know why). We have obtained a formula which contains apriori unknown values (random variables) and it gives different results in repeated experiments. Naturally, the random results have some distribution which may be, hopefully, described.

Theorem 13 (Distribution of edf I)**.** *Let $X_1, \dots, X_n$ be a random sample from distribution with cdf F and fix some x. Then the distribution of $\widehat{F_n}(x)$ is given by*

$$\mathrm{P}\left[\widehat{F_n}(x) = \frac{k}{n}\right] = \binom{n}{k}\big(F(x)\big)^k\big(1 - F(x)\big)^{n-k}.$$

Důkaz. Obviously $n\widehat{F_n}(x) = k$ iff exactly k out of n values in the random sample are at most x. All random variables X_i in the sample satisfy

$$\mathrm{P}[\chi(X_i \le x) = 1] = \mathrm{P}[X_i \le x] = F(x) = 1 - \mathrm{P}[\chi(X_i \le x) = 0],$$

hence the sum $\sum_{i=1}^n \chi(X_i \le x)$ follows the binomial distribution. The rest of the proof is obvious. □

3. Mean value and other moments

We shall define numerical characteristics of the distribution and corresponding random variables. The behaviour of random variable is fully described by its distribution or cdf or pdf. But the cdf is a function and it is not easy to understand what does it say exactly and how may be two or more random variables compared based on their distribution functions. The numerical characteristics allow to say some features of the random variables in few numbers. Other applications of these characteristics will be shown later.

Definition 18 (Mean value, mathematical definition)**.** Let X be a random variable defined on probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The value

$$\mathrm{E}\, X := \int_\Omega X(\omega) \mathrm{d}\mathrm{P}(\omega)$$

is called *mean value* of X if the integral exists.

The mean value is also called (mathematical) expectation of X.

The mean value is calculated using the density function in practice.

Theorem 14 (Mean value of discrete random variable)**.** *Let X be a discrete random variable with values in $\mathbb{S}$ and with density function $p_X s$. Then*

$$\mathrm{E}\, X = \sum_{s\in\mathbb{S}} s\mathrm{P}[X = s] = \sum_{s\in\mathbb{S}} = \sum_{s\in\mathbb{S}} sp_X(s),$$

if the right hand side exists.

Důkaz. Use the definition of $\mathrm{E}\, X$. Sets $A_s := \{\omega : X(\omega) = s\}$ form disjoint partition of Ω. Therefore

$$\begin{aligned}\mathrm{E}\, X &= \int_{\bigcup A_s} X(\omega)\mathrm{dP}(\omega) = \sum_{s\in\mathbb{S}} \int_{A_s} X(\omega)\mathrm{dP}(\omega)\\ &= \sum_{s\in\mathbb{S}} s \int_{A_s} \mathrm{dP}(\omega) = \sum_{s\in\mathbb{S}} s\mathrm{P}(A_s)\\ &= \sum_{s\in\mathbb{S}} s\mathrm{P}[X = s].\end{aligned}$$

□

Theorem 15 (Mean value of continuous random variable)**.** *Let X be a continuous random variable with density dunction $f_X x$. Then*

$$\mathrm{E}\, X = \int_{-\infty}^{\infty} xf_X(x)\mathrm{d}x$$

if the integral exists.

Remark. The mean **may exists** although being infinite. The random variable X has infinite mean if the integral in Definition 18 is infinite ($\pm\infty$). The integral is meaningless if if it is of type „$\infty - \infty$“. In particular $\int_{\mathbb{R}} 1\mathrm{d}x$ does exists (although being infinite) while $\int_{\mathbb{R}} x^{-1}\mathrm{d}x$ doesn't exist.

Problem. Find a function $p_X(s), s \in \mathbb{Z}$ such that:

(1) $p_X(s) \geq 0$
(2) $\sum_{s\in\mathbb{Z}} p_X(s) = 1$
(3) $\sum_{s>0} sp_X(s) = \infty$, $\sum_{s<0} sp_X(s) = -\infty$

There is no mean for such probaboility density function since the sum is not well defined ($\infty - \infty$). Finding of probability density f_X of a continuous random variable such that the mean doesn't exist is also easy.

Remark (Terminology and trivial facts).
- If $\mathrm{E}\, X$ exists and is finite then we say that X has finite mean.
- If $P[X \geq b] = 1$ for some finite constant b then $\mathrm{E}\, X$ does exist and $\mathrm{E}\, X > b$. In particular non-negative random variable has non-negative mean. (And vice-versa for the reverse inequality.)
- If $\exists a, b < \infty$ and $P[a \leq X \leq b] = 1$ then $\mathrm{E}\, X$ exists and is finite. Moreover $a \leq \mathrm{E}\, X \leq b$.
- Define $\mathrm{E}\, |X| = \int_\Omega |X(\omega)|\, dP(\omega)$ (exists always). If $\mathrm{E}\, |X| < \infty$ then $X \in L_1(\mathrm{P})$ and there is a finite mean of X, $|\mathrm{E}\, X| < \infty$.

We may define also moment of a function of random variable and higher order moments.

Definition 19 (General moments of random variable)**.** Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$. Then $Eg(X) = \int_\Omega g(X(\omega))\mathrm{dP}(\omega)$ if the integral exists.

Theorem 16 (Calculation of the mean)**.** *Let X be a discrete $\mathbb{S}$-valued random variable and $p_X s$ its probability density, $g : \mathbb{R} \to \mathbb{R}$. Then*

$$\mathrm{E}\, g(X) = \sum_{s\in\mathbb{S}} g(s)\mathrm{P}[X = s] = \sum_{s\in\mathbb{S}} = \sum_{s\in\mathbb{S}} g(s)p_X(s),$$

if the sum exists.

Let X be a continuous random variable and $f_X x$ its probability density function. Then

$$\mathrm{E}\, g(X) = \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x,$$

if the integral exists.

Theorem 17 (Linearity of the mean value)**.** *Let X be a random variable with finite mean. Then*

$$\mathrm{E}(a + bX) = a + b\, \mathrm{E}\, X, \forall a, b \in \mathbb{R}.$$

Důkaz. Follows directly form the definition. □

Theorem 18 (Jensen inequality)**.** *Let X be a random variable with finite mean $\mathrm{E}\, X$. Let φ be a convex function. Then $\mathrm{E}\, \varphi(x) \geq \varphi(\mathrm{E}\, X)$.*

Důkaz. Since the function φ is convex for any a there exists a constant λ such that $\varphi(x) \geq \varphi(a) + \lambda(x - a)$. Choosing $a = \mathrm{E}\, X$ we get

$$\varphi(X) \geq \varphi(\mathrm{E}\, X) + \lambda(X - \mathrm{E}\, X).$$

Since the mean of non-negative random variable is non-negative and since $\mathrm{E}(X - \mathrm{E}\, X) = 0$ the proof follows easily. □

Note that (real) function of a random variable is a random variable itself iff the conditions of Definition 9 hold. However, all common functions are such. In measure theroy such funcitons are called measurable.

Definition 20 (Important moments and moment generating function)**.** Let X be a random variable and consider $r \in \mathbb{N}$. Then

(1) $\mathrm{E}\, X^r$ is the r-th moment of X.
(2) $\mathrm{E}\, |X|^r$ is the r-th absolut moment of X.
(3) For $r \in \mathbb{N}$ $\mathrm{E}(X - EX)^r$ is the r-th central moment of X.
(4) For $r = 2$ the $\operatorname{var} x = \mathrm{E}(X - \mathrm{E}\, X)^2$ is variance of X.
(5) $\mu_3 = \frac{\mathrm{E}(X - \mathrm{E}\, X)^3}{(\mathrm{E}(X - \mathrm{E}\, X)^2)^{\frac{3}{2}}}$ is the skewness of the distribution of X (measure of the asymmetry).
(6) $\Psi_X(t) = \mathrm{E}\, e^{tX}$ is the moment generating function defined for those t for which the expectation does exist. For any random variable X we have $\Psi_X(0) = 1$.

All moments are defined if the integrals and sums do exists.

Remark (Basic properties of moments). Moments numerically characterise some features of random variables and their distributions. There are also some interesting relations between moments.

- The mean $\mathrm{E}\, X$ characterises the location of random variable.
- The variance characterises the dispersion. Namely it is the mean squared deviation of random variable.
- There are other characteristics of dispersion like the absolute central moment $\mathrm{E}\, |X - \mathrm{E}\, X|$.
- Let $0 < s < r$. If $\mathrm{E}\, |X|^r < \infty$ then $\mathrm{E}\, |X|^s < \infty$ and $|\,\mathrm{E}\, X^r| < \infty$ if the integral exists (e.g. $r \in \mathbb{N}$). Moreover $\left(\mathrm{E}\, |X|^s\right)^{1/s} \leq \left(\mathrm{E}\, |X|^r\right)^{1/r}$, in particular $\mathrm{E}\, |X| \leq \left(\mathrm{E}\, |X|^2\right)^{1/2} \leq \left(\mathrm{E}\, |X|^3\right)^{1/3}$.
- If there is $\delta > 0$ such that the moment generating function $\psi_X(t)$ exists finite $\forall |t| < \delta$ then $\forall r \in \mathbb{N}\, \mathrm{E}\, X^r = \psi_X^{(r)}(0)$ (r-th derivative of $\psi_X(t)$ at 0).

Theorem 19 (Computation and properties of variance)**.** *Let X be a random variable with finite variance. Then*

$$\operatorname{var} X = \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2 = \mathrm{E}(X(X-1)) - \mathrm{E}\, X(\mathrm{E}\, X - 1).$$

Let a and b be any constants. Then

$$\operatorname{var}(a + bX) = b^2 \operatorname{var} X.$$

Důkaz. Direct calculation. □

Remark. The variance is the mean of non-negative function hence it is non-negative itself. In particular $\mathrm{E}\, X^2 \geq (\mathrm{E}\, X)^2$. as follows also from the Jensen inequality.

Let us define moments of random vector. Namely we need just one moment characterising the joint behaviour of two components or random vector the other moments we know already.

Definition 21 (Mean value of random vector)**.** Let $\boldsymbol{X}$ be a random vector. Then we define:

(1) $\mathrm{E}\,\boldsymbol{X} = (\mathrm{E}\,X_1, \dots, \mathrm{E}\,X_d)^T$
(2) Let $g : \mathbb{R}^d \to \mathbb{R}$ (such that $g(\boldsymbol{X})$ is a random variable). Then define $\mathrm{E}\,g(\boldsymbol{X}) = \int_\Omega g(\boldsymbol{X}(\omega))\,dP(\omega)$ if all integrals exist.

Theorem 20. *Let* $\boldsymbol{X}$ *be a* discrete $\mathbb{N}_0^d$ *valued random vector. Then* $\mathrm{E}\,g(\boldsymbol{X}) = \sum_{\boldsymbol{z}\in\mathbb{N}_0^d} g(\boldsymbol{z})P[\boldsymbol{X} = \boldsymbol{z}]$ *if the sum exists.*

Let $\boldsymbol{X}$ *be a* continuous *(*$\mathbb{R}^d$*/valued) random vector and* $f_{\boldsymbol{X}}$ *its probability density function. Then* $\mathrm{E}\,g(\boldsymbol{X}) = \int_{\mathbb{R}^d} g(\boldsymbol{z})f_{\boldsymbol{X}}(\boldsymbol{z})\mathrm{d}\boldsymbol{z}$ *if the integral exists.*

In both cases we have multiple sums and integrals.