Chapter 1
Introduction to Probability Theory

1.1 Summary of Basic Notions of Probability Theory

In this chapter we summarize the most important notions and facts of probability theory that are necessary for an elaboration of our topic. In the present summary, we will apply the more specific mathematical concepts and facts – mainly measure theory and analysis – only to the necessary extent while, however, maintaining mathematical precision.

Random Event We consider experiments whose outcomes are uncertain, where the totality of the circumstances that are or can be considered does not determine the outcome of the experiment. A set consisting of all possible outcomes is called a **sample space**. We define **random events** (**events** for short) as certain sets of outcomes (subsets of the sample space). It is assumed that the set of events is closed under countable set operations, and we assign probability to events only; they characterize the quantitative measure of the degree of uncertainty. Henceforth countable means finite or countably infinite.

Denote the sample space by $\Omega = \{\omega\}$. If Ω is countable, then the space Ω is called **discrete**. In a mathematical approach, events can be defined as subsets $A \subset \Omega$ of the possible outcomes Ω having the properties (σ-algebra properties) defined subsequently.

A given event A occurs in the course of an experiment if the outcome of the experiment belongs to the given event, that is, if an outcome $\omega \in A$ exists. An event is called simple if it contains only one outcome ω. It is always assumed that the whole set Ω and the empty set $\varnothing$ are events that are called a **certain** event and an **impossible event**, respectively.

Operation with Events; Notion of σ-Algebra Let A and B be two events. The **union** $A \cup B$ of A and B is defined as an event consisting of all elements $\omega \in \Omega$ belonging to either event A or B, i.e., $A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$.

The **intersection** (**product**) $A \cap B$ (AB) of events A and B is defined as an event consisting of all elements $\omega \in \Omega$ belonging to both A and B, i.e.,

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, DOI 10.1007/978-1-4614-5317-8_1, © Springer Science+Business Media, LLC 2013

$$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}.$$

The **difference** $A \backslash B$, which is not a symmetric operation, is defined as the set of all elements $\omega \in \Omega$ belonging to event A but not to event B, i.e.,

$$A \backslash B = \{\omega : \omega \in A \text{ and } \omega \notin B\}.$$

A **complementary event** $\overline{A}$ of A is defined as a set of all elements $\omega \in \Omega$ that does not belong to A, i.e.,

$$\overline{A} = \Omega \backslash A.$$

If $A \cap B = \oslash$, then sets A and B are said to be **disjoint** or **mutually exclusive**.

Note that the operations $\cup$ and $\cap$ satisfy the associative, commutative, and distributive properties

$$(A \cup B) \cup C = A \cup (B \cup C), \quad \text{and} \quad (A \cap B) \cap C = A \cap (B \cap C),$$
$$A \cup B = B \cup A, \quad \text{and} \quad A \cap B = B \cap A,$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad \text{and} \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

DeMorgan identities are valid also for the operations union, intersection, and complementarity of events as follows:

$$\overline{A \cup B} = \overline{A} \cap \overline{B}, \qquad \overline{A \cap B} = \overline{A} \cup \overline{B}.$$

With the use of the preceding definitions introduced, we can define the notion of σ-algebra of events.

Definition 1.1. Let Ω be a nonempty (abstract) set, and let $\mathcal{A}$ be a certain family of subsets of the set Ω satisfying the following conditions:

(1) $\Omega \in \mathcal{A}$.
(2) If $A \in \mathcal{A}$, then $\overline{A} \in \mathcal{A}$.
(3) If $A_1, A_2, \ldots \in \mathcal{A}$ is a countable sequence of elements, then

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}.$$

The family $\mathcal{A}$ of subsets of the set Ω satisfying conditions (1)–(3) is called a **σ-algebra**. The elements of $\mathcal{A}$ are called **random events**, or simply **events**.

Comment 1.2. *The pair* $(\Omega, \mathcal{A})$ *is usually called a* ***measurable space****, which forms the general mathematical basis of the notion of probability.*

Probability Space, Kolmogorov Axioms of Probability Theory Let Ω be a nonempty sample set, and let $\mathcal{A}$ be a given σ-algebra of subsets of Ω, i.e., the pair

$(\Omega, \mathcal{A})$ is a measurable space. A nonnegative number $\mathbf{P}(A)$ is assigned to all events A of σ-algebra satisfying the axioms as follows.

A1. $0 \le \mathbf{P}(A) \le 1$, $A \in \mathcal{A}$.
A2. $\mathbf{P}(\Omega) = 1$.
A3. If the events $A_i \in \mathcal{A}$, $i = 1, 2, \ldots$, are disjoint (i.e., $A_i A_j = \oslash$, $i \ne j$), then

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

The number $\mathbf{P}(A)$ is called the **probability** of event A, axioms A1, A2, and A3 are called the Kolmogorov axioms, and the triplet $(\Omega, \mathcal{A}, \mathbf{P})$ is called the probability space. As usual, axiom A3 is called the σ-additivity property of the probability. The probability space characterizes completely a random experiment.

Comment 1.3. *In the measure theory context of probability theory, the function* $\boldsymbol{P}$ *defined on* $\mathcal{A}$ *is called a probability measure. Conditions A1–A3 ensure that* $\boldsymbol{P}$ *is nonnegative and that* σ *is an additive and normed [*$\boldsymbol{P}(\Omega) = 1$*] set function on* $\mathcal{A}$*, i.e., a normed measure on* $\mathcal{A}$*. Our discussion basically does not require the direct use of measure theory, but some assertions cited in this work essentially depend on this theory.*

Main Properties of Probability Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space. The following properties of probability are valid for all probability spaces.

Elementary properties:

(a) The probability of an impossible event is zero, i.e.,
$\mathbf{P}(\oslash) = 0$.
(b) $\mathbf{P}\left(\overline{A}\right) = 1 - \mathbf{P}(A)$ for all $A \in \mathcal{A}$.
(c) If the relationship $A \subseteq B$ is satisfied for given events $A, B \in \mathcal{A}$, then
$\mathbf{P}(A) \le \mathbf{P}(B)$,
$\mathbf{P}(B - A) = \mathbf{P}(B) - \mathbf{P}(A)$.

Definition 1.4. A collection $\{A_i,\ i \in I\}$ of a countable set of events is called a **complete system** of events if $A_i,\ i \in I$ are disjoint (i.e., $A_i \cap A_j = \oslash$ if $i \ne j$, $i, j \in I$) and $\bigcup\limits_{i \in I} A_i = \Omega$.

Comment 1.5. *If the collection of events* $\{A_i,\ i \in I\}$ *forms a complete system of events, then*

$$\boldsymbol{P}\left(\bigcup_{i \in I} A_i\right) = 1.$$

Probability of Sum of Events, Poincaré Formula For any events A and B it is true that

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(AB).$$

Using this relation, a more general formula, called the **Poincaré formula**, can be proved. Let n be a positive integer number; then, for any events $A_1, A_2, \ldots, A_i \in \mathcal{A}$,

$$\mathbf{P}(A_1 + \ldots + A_n) = \sum_{k=1}^{n} (-1)^{k-1} S_k^{(n)},$$

where $S_k^{(n)} = \sum_{1 \le i_1 \le \ldots \le i_k \le n} \mathbf{P}\left(A_{i_1} \ldots A_{i_k}\right)$.

Subadditive Property of Probability For any countable set of events $\{A_i,\ i \in I\}$ the inequality

$$\mathbf{P}\left(\bigcup_{i \in I} A_i\right) \le \sum_{i \in I} \mathbf{P}(A_i)$$

is true.

Continuity Properties of Probability Continuity properties of probability are valid for monotonically sequences of events, each of which is equivalent to axiom A3 of probability. A sequence of events $A_1, A_2, \ldots$ is called monotonically increasing (resp. decreasing) if $A_1 \subset A_2 \subset \ldots$ (resp. $A_1 \supset A_2 \supset \ldots$).

Theorem 1.6. *If the sequence of events $A_1, A_2, \ldots$ is monotonically decreasing, then*

$$\boldsymbol{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n \to \infty} \boldsymbol{P}(A_n).$$

If the sequence of events $A_1, A_2, \ldots$ is monotonically increasing, then

$$\boldsymbol{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n \to \infty} \boldsymbol{P}(A_n).$$

Conditional Probability and Its Properties, Independence of Events In practice, the following obvious question arises: if we know that event B occurs (i.e., the outcome is in $B \in \mathcal{A}$), what is the probability that the outcome is in $A \in \mathcal{A}$? In other words, how does the occurrence of an event B influence the occurrence of another event A? This effect is characterized by the notion of conditional probability $\mathbf{P}(A|B)$ as follows.

Definition 1.7. Let A and B be two events, and assume that $\mathbf{P}(B) > 0$. The quantity

$$\mathbf{P}(A|B) = \mathbf{P}(AB)/\mathbf{P}(B)$$

is called the **conditional probability of A given B**.

It is easy to verify that the conditional probability possesses the following properties:

1. $0 \le \mathbf{P}(A|B) \le 1$.
2. $\mathbf{P}(B|B) = 1$.
3. If the events $A_1, A_2, \ldots$ are disjoint, then

$$\mathbf{P}\left(\sum_{i=1}^{\infty} A_i | B\right) = \sum_{i=1}^{\infty} \mathbf{P}(A_i|B).$$

4. The definition of conditional probability $\mathbf{P}(A|B) = \mathbf{P}(AB)/\mathbf{P}(B)$ is equivalent to the so-called theorem of multiplication

$$\mathbf{P}(AB) = \mathbf{P}(A|B)\mathbf{P}(B) \text{ and } \mathbf{P}(AB) = \mathbf{P}(B|A)\mathbf{P}(A).$$

Note that these equations are valid in the cases $\mathbf{P}(B) = 0$ and $\mathbf{P}(A) = 0$ as well.

One of the most important concepts of probability theory, the independence of events, is defined as follows.

Definition 1.8. We say that events A and B are **independent** if the equation

$$\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B)$$

is satisfied.

Comment 1.9. *If A and B are independent events and $\boldsymbol{P}(B) > 0$, then the conditional probability $\boldsymbol{P}(A|B)$ does not depend on event B since*

$$\boldsymbol{P}(A|B) = \frac{\boldsymbol{P}(AB)}{\boldsymbol{P}(B)} = \frac{\boldsymbol{P}(A)\boldsymbol{P}(B)}{\boldsymbol{P}(B)} = \boldsymbol{P}(A).$$

This relation means that knowing that an event B occurs does not change the probability of another event A.

The notion of independence of an arbitrary collection A_i, $i \in I$ of events is defined as follows.

Definition 1.10. A given collection of events A_i, $i \in I$ is said to be **mutually independent (independent** for short) if, having chosen from among them any finite number of events, the probability of the product of the chosen events equals the product of the probabilities of the given events. In other words, if $\{i_1, \ldots, i_k\}$ is any subcollection of I, then one has

$$\mathbf{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right) = \mathbf{P}(A_{i_1}) \ldots \mathbf{P}\left(A_{i_k}\right).$$

This notion of independence is stricter when pairs are concerned since it is easy to create an example where pairwise independence occurs but mutual independence does not.

Example 1.11. We roll two dice and denote the pair of results by

$$(\omega_1, \omega_2) \in \Omega = \{(i, j), 1 \leq i, j \leq 6\}.$$

The number of elements of the set Ω is $|\Omega| = 36$, and we assume that the dice are standard, that is, $P\{(\omega_1, \omega_2)\} = 1/36$ for every $(\omega_1, \omega_2) \in \Omega$. Events A_1, A_2, and A_3 are defined as follows:

$$\begin{aligned} A_1 &= \{\text{the result of the first die is even}\}, \\ A_2 &= \{\text{the result of the second die is odd}\}, \\ A_3 &= \{\text{both the first and second dice are odd or both of them are even}\}. \end{aligned}$$

We check that events A_1, A_2, and A_3 are pairwise independent, but they are not (mutually) independent. It is clear that

$$\begin{aligned} A_1 &= \{(2,1), \ldots, (2,6), (4,1), \ldots, (4,6), (6,1), \ldots, (6,6)\}, \\ A_2 &= \{(1,1), \ldots, (6,1), (1,3), \ldots, (6,3), (1,5), \ldots, (6,5)\}, \\ A_3 &= \{(1,1), (1,3), (1,5), (2,2), (2,4), (2,6), (3,1), (3,3), \\ &\quad (3,5), \ldots, (6,2), (6,4)(6,6)\}, \end{aligned}$$

thus

$$|A_1| = 3 \cdot 6 = 18, \quad |A_2| = 6 \cdot 3 = 18, \quad |A_3| = 6 \cdot 3 = 18.$$

We have, then, $\mathbf{P}(A_i) = \frac{1}{2}$, $i = 1, 2, 3$, and the relations

$$\mathbf{P}\left(A_i A_j\right) = \frac{1}{4} = \mathbf{P}(A_i)\mathbf{P}\left(A_j\right), \ 1 \leq i, j \leq 3, \ i \neq j,$$

which means events A_1, A_2, and A_3 are pairwise independent. On the other hand,

$$\mathbf{P}(A_1 A_2 A_3) = 0 \neq \frac{1}{8} = \mathbf{P}(A_1)\mathbf{P}(A_2)\mathbf{P}(A_3);$$

consequently, the mutual independence of events A_1, A_2, and A_3 does not follow from their pairwise independence.

Formula of Total Probability, Bayes' Rule Using the theorem of multiplication for conditional probability we can easily derive the following two theorems. Despite the fact that the two theorems are not complicated, they represent quite effective tools in the course of the various considerations.

Theorem 1.12 (*Formula of total probability*). *Let the sequence* $\{A_i,\ i \in I\}$ *be a complete system of events with* $\boldsymbol{P}(A_i > 0),\ i \in I$*; then for all events* B

$$\boldsymbol{P}(B) = \sum_{i \in I} \boldsymbol{P}(B|A_i)\boldsymbol{P}(A_i)$$

is true.

Theorem 1.13 (*Bayes' rule*). *Under the conditions of the preceding theorem, the following relation holds for all indices* $n \in I$*:*

$$\mathbf{P}(A_n|B) = \frac{\mathbf{P}(B|A_n)\mathbf{P}(A_n)}{\sum\limits_{i \in I} \mathbf{P}(B|A_i)\mathbf{P}(A_i)}.$$

Concept of Random Variables Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space that is to be fixed later on. In the course of random experiments, the experiments usually result in some kind of value. This means that the occurrence of a simple event ω results in a random $X(\omega)$ value. Different values might belong to different simple events; however, the function $X(\omega)$, depending on the simple event ω, will have a specific property. We must answer such basic questions as, for example, what is the probability that the result of the experiment will be smaller than a certain given value x? We have only determined probabilities of events (only for elements of the set $\mathcal{A}$) in connection with the definition of probability space; therefore, it has the immediate consequence that we may only consider the probability of the set if the set $\{\omega : X(\omega) \leq x\}$ is an event, which means that the set belongs to σ-algebra $\mathcal{A}$:

$$\{\omega : X(\omega) \leq x\} \in \mathcal{A}.$$

This fact led to one of the most important notions of probability theory.

Definition 1.14. The real-valued function $X : \Omega \to \mathbb{R}$ is called a **random variable** if the relationship

$$\{\omega : X(\omega) \leq x\} \in \mathcal{A}$$

is valid for all real numbers $x \in \mathbb{R}$. A function satisfying this condition is called $\mathcal{A}$ **measurable**.

A property of random variables should be mentioned here. Define by $\mathcal{B} = \mathcal{B}_1$ the σ-algebra of Borel sets of $\mathbb{R}$ as the minimal σ-algebra containing all intervals of $\mathbb{R}$; the elements of $\mathcal{B}$ are called the Borel sets of $\mathbb{R}$. If X is $\mathcal{A}$ measurable, then for all Borel sets D of $\mathbb{R}$ the set $\{\omega : X(\omega) \in D\}$ is also an element of $\mathcal{A}$, i.e., $\{\omega : X(\omega) \in D\}$ is an event. Thus the probability $\mathbf{P}_X[D] = \mathbf{P}(\{\omega : X(\omega) \in D\})$, and so $\mathbf{P}(\{\omega : X(\omega) \leq x\})$ are well defined. An important special case of random variables are the so-called **indicator variables** defined as follows. Let $A \in \mathcal{A}$ be an event, and let us introduce the random variable $\mathcal{I}_{\{A\}}$, $A \in \mathcal{A}$:

$$\mathcal{I}_{\{A\}} = \mathcal{I}_{\{A\}}(\omega) = \begin{cases} 1, \text{ if } \omega \in A, \\ 0, \text{ if } \omega \notin A. \end{cases}$$

Distribution Function Let $X = X(\omega)$ be a random variable; then the probability $\mathbf{P}(X \leq x)$, $x \in \mathbb{R}$, is well defined.

Definition 1.15. The function $F_X(x) = \mathbf{P}(X \leq x)$ for all real numbers $x \in \mathbb{R}$ is called a **cumulative distribution function**(CDF) of random variable X .

Note that the CDFs F_X and function $\mathbf{P}_X$ determine each other mutually and unambiguously. It is also clear that if the real line $\mathbb{R}$ is chosen as a new sample space, and $\mathcal{B}$ is a σ-algebra of Borel sets as the σ-algebra of events, then the triplet $(\mathbb{R}, \mathcal{B}, \mathbf{P}_X)$ determines a new probability space, where $\mathbf{P}_X$ is referred to as a probability measure induced by the random variable X.

The CDF F_X has the following properties.

(1) In all points of a real line $-\infty < x_0 < \infty$ the function $F_X(x)$ is continuous from the right, that is,

$$\lim_{x \to x_0 + 0} F_X(x) = F_X(x_0).$$

(2) The function $F_X(x)$, $-\infty < x < \infty$ is a monotonically increasing function of the variable x, that is, for all $-\infty < x < y < \infty$ the inequality $F_X(x) \leq F_X(y)$ holds.

(3) The limiting values of the function $F_X(x)$ exist under the conditions $x \to -\infty$ and $x \to \infty$ as follows:

$$\lim_{x \to -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F_X(x) = 1.$$

(4) The set of discontinuity points of the function $F_X(x)$, that is, the set of points $x \in \mathbb{R}$ for which $F_X(x) \neq F_X(x - 0)$, is countable.

Comment 1.16. *It should be noted in connection with the definition of the CDF that the literature is not consistent. The use of $F_X(x) = \boldsymbol{P}(X < x)$, $-\infty < x < \infty$ as a CDF is also widely applied. The only difference between the two definitions lies within property (1) (see preceding discussion), which means that in the latter case the CDF is continuous from the left and not from the right, but all the other properties remain the same. It is also clear that if the CDF is continuous in all $x \in \mathbb{R}$, then there is no difference between the two definitions.*

Comment 1.17. *From a practical point of view, it is sometimes useful to allow that property (3) (see preceding discussion) does not satisfy the CDF F_X of random variable X, which means that, instead, one or both of the following relations hold: In this case $\boldsymbol{P}(|X| < \infty) < 1$, and the CDF of random variable X has a **defective distribution function**.*

Let a and b be two arbitrary real numbers for which $-\infty < a < b < \infty$; then we can determine the probability of some frequently occurring events with the use of the CDF of X as follows:

$$\begin{aligned}
\mathbf{P}(X = a) &= F_X(a) - F_X(a - 0),\\
\mathbf{P}(a < X < b) &= F_X(b - 0) - F_X(a),\\
\mathbf{P}(a \le X < b) &= F_X(b - 0) - F_X(a - 0),\\
\mathbf{P}(a < X \le b) &= F_X(b) - F_X(a),\\
\mathbf{P}(a \le X \le b) &= F_X(b) - F_X(a - 0).
\end{aligned}$$

These equations also determine the connection between the CDF F_X and the distribution $\mathbf{P}_X$ for special Borel sets of a real line.

Discrete and Continuous Distribution, Density Function We distinguish two important types of distributions in practice, the so-called discrete and continuous distributions. There is also a third type of distribution, the so-called singular distribution, in which case the CDF is continuous everywhere and its derivative (with respect to the Lebesgue measure) equals 0 almost everywhere; however, we will not consider this type. This classification follows from the Jordan decomposition theorem of monotonically functions, that is, an arbitrary CDF F can always be decomposed into the sum of three functions – the monotonically increasing absolutely continuous function, the step function with finite or countably infinite sets of jumps (this part corresponds to a discrete distribution), and the singular function.

Definition 1.18. Random variable X is **discrete** or has a **discrete distribution** if there is a finite or countably infinite set of values $\{x_k,\ k \in I\}$ such that $\sum\limits_{k\in I} p_k = 1$, where $p_k = \mathbf{P}(X = x_k),\ k \in I$. The associated function

$$f_X(x) = \begin{cases} p_k, & \text{if } x = x_k,\ k \in I,\\ 0, & \text{if } x \ne x_k,\ k \in I, \end{cases} \qquad x \in \mathbb{R},$$

is termed a **probability density function** (PDF) or **probability mass function** (PMF).

It is easy to see that if random variable X is discrete with possible values $\{x_k, k = 0, 1, \ldots\}$ and with distribution $\{p_k, k = 0, 1, \ldots\}$, then the relationship between the CDF F_X and the PMF can be given as

$$F_X(x) = \sum_{x_k < x} p_k,\ -\infty < x < \infty.$$

Definition 1.19. A random variable X is **continuous** or has a **continuous distribution** if there exists a nonnegative integrable function $f_X(x)$, $-\infty < x < \infty$ such that for all real numbers a and b, $-\infty < a < b < \infty$,

$$F_X(b) - F_X(a) = \int_a^b f_X(x)\mathrm{d}x$$

holds. The function $f_X(x)$ is called the PDF of random variable X, or just the **density function** of X.

Comment 1.20. *It is clear that*

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u, \ -\infty < x < \infty,$$

and it is also true that the PDF is not uniquely defined since if we take instead of $f_X(u)$ *the function* $f_X(u)+g(u)$*, where the function* $g(u)$ *is nonnegative, integrable, and* $\int_{-\infty}^{x} g(u)\mathrm{d}u = 0$*, then the function* $f_X(u) + g(u)$ *is also a PDF of random variable* X*, which can naturally differ from the original* f_X*.*

An arbitrary PDF $f_X(x)$ is nonnegative and integrable,

$$\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1,$$

and almost everywhere in $\mathbb{R}$ (with respect to the Lebesgue measure) the equation $F'_X(x) = f_X(x)$ is true.

Distribution of a Function of a Random Variable Let $X = X(\omega)$ be a random variable. Let $h(x)$, $x \in \mathbb{R}$ be a real-valued function, and let us define it as $Y = h(X)$. The equation $Y = h(X)$ determines a random variable if for all $y \in \mathbb{R}$ the set $\{\omega : Y(\omega) = h(X(\omega)) \le y\}$ is an event that is an element of σ-algebra $\mathcal{A}$. If h is a continuous function or, more generally, is a Borel-measurable function (h is Borel measurable if for all x the relationship $\{u : h(u) \le x\} \in \mathcal{B}$ is true), then Y, which is determined by the equation $Y = h(X)$, is a random variable. The question is how the CDF and the density function (if the latter exists) of random variable Y can be determined. It is usually true that

$$F_X(y) = \mathbf{P}(Y \le y) = \mathbf{P}(h(X) \le y) = \mathbf{P}_X[\{x : h(x) \le y\}], \ -\infty < y < \infty.$$

If h is a strictly monotonically increasing function, then this formula can be given in a simpler form. Let us denote by h^{-1} the inverse function of h, which in this case must exist. Then

$$F_X(y) = \mathbf{P}(h(X) \le y) = \mathbf{P}\left(X \le h^{-1}(y)\right) = F_X(h^{-1}(y)),\ -\infty < y < \infty.$$

If h is a strictly monotonically decreasing function, then

$$F_X(y) = \mathbf{P}(h(X) \le y) = \mathbf{P}\left(X \ge h^{-1}(y)\right) = 1 - F_X(h^{-1}(y) - 0),\ -\infty < y < \infty.$$

With these relations, a formula can be given for the PDF of Y in special cases.

Theorem 1.21. *Let us suppose that random variable X has a PDF f_X and h is a strictly monotonically, differentiable real function. Then*

$$f_Y(y) = f_X(h^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} h^{-1}(y) \right|,\ -\infty < y < \infty.$$

Comment 1.22. *If h is a linear function, that is, $h(y) = ay + b$, $a \neq 0$, and X has a PDF f_X, then the random variable $Y = h(X)$ also has a PDF and the formula*

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-a}{b}\right),\ -\infty < y < \infty,$$

is true.

Joint Distribution and Density Function of Random Variables, Marginal Distributions In the majority of problems arising in practice, we have not one but several random variables, and we examine the probability of events where random variables simultaneously satisfy certain conditions.

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space, and let there be two random variables X and Y on that space. The joint statistical behavior of the two random variables can be determined by a **joint CDF**. We should note that the joint analysis of the random variables X and Y corresponds to the examination of two-dimensional random vector variables such as (X, Y) that have random variable coordinates.

Definition 1.23. The function

$$F_{XY}(x, y) = \mathbf{P}(X \le x, Y \le y),\ -\infty < x, y < \infty,$$

is called the **joint CDF** of random variables X and Y.

From a practical point of view, the two most important types of distributions are the discrete and the continuous ones, as in the one-dimensional case.

Definition 1.24. The joint distribution function of random variables X and Y is called **discrete**; in other words, the random vector (X, Y) has a **discrete distribution** if random variables X and Y are discrete. If we denote the values

of random variables X and Y by $\{x_i,\ i \in I\}$ and $\{y_j,\ j \in J\}$, respectively, then the function

$$f_{X,Y}(x,y) = \begin{cases} p_{i,j}, & \text{if } x = x_i,\ y = y_j,\ i \in I,\ j \in J, \\ 0, & \text{if } x \neq x_i,\ y \neq y_j,\ i \in I,\ j \in J, \end{cases} \quad x \in \mathbb{R},$$

is called a **joint PMF** or **joint PDF**.

It is clear that in the discrete case the joint distribution function is

$$F_{XY}(x,y) = \sum_{x_i \leq x,\ y_j \leq y} p_{ij}.$$

The case of a joint continuous distribution is analogous to the discrete one.

Definition 1.25. The joint distribution of random variables X and Y is called **continuous**; in other words, the random vector (X,Y) has a **continuous distribution** if there exists a nonnegative, real-valued integrable function on the plane $f_{XY}(x,y)$, $-\infty < x, y < \infty$, for which the relation

$$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u,v)\mathrm{d}u\mathrm{d}v$$

holds for all $-\infty < x, y < \infty$.

Definition 1.26. If F_{XY} denotes the joint CDF of random variables X and Y, then the CDFs

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x,y),$$
$$F_Y(y) = \lim_{x \to \infty} F_{XY}(x,y)$$

are called **marginal distribution functions**.

It is not difficult to see that marginal distribution functions do not determine the joint CDF. It is also clear that if a joint PDF $f_{XY}(x,y)$ of random variables X and Y exists, then marginal PDFs can be given in the form

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}y,\ -\infty < x < \infty,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}x\ -\infty < y < \infty.$$

If there are more than two random variables $X_1, \ldots, X_n$, $n \geq 3$, i.e., in the case of an n-dimensional random vector $(X_1, \ldots, X_n)$, then the definitions of joint

distribution function and density functions can be given analogously to the case of two random variables, so there is no essential difference. We will return to this question when we introduce the concept of stochastic processes.

Conditional Distributions Let A be an arbitrary event, with $P(A) > 0$, and X an arbitrary random variable. Using the notion of conditional probability, we can define the **conditional distribution** of random variable X given event A as the function

$$F_X(x|A) = \mathbf{P}(X \leq x|A),\ x \in \mathbb{R}.$$

The function $F_X(x|A)$ has all the properties of a distribution function mentioned previously.

The function $f_X(x|A_i)$ is called a **conditional density function** of random variable X given event A if a nonnegative integrable function $f_X(x|A)$ exists for which the equation

$$F_X(x|A) = \int_{-\infty}^{x} f_X(u|A)\mathrm{d}u,\ -\infty < x < \infty,$$

holds.

The result for the distribution function $F_X(x)$ can be easily proved in the same way as the theorem of full events. If the sequence of events $A_1, A_2, \ldots$ is a complete system of events with the property $\mathbf{P}(A_i) > 0,\ i = 1, 2, \ldots$, then

$$F_X(x) = \sum_{i=1}^{\infty} F_X(x|A_i)\mathbf{P}(A_i),\ -\infty < x < \infty.$$

A similar relation holds for the conditional PDFs $f_X(x|A_i),\ i \geq 1$, if they exist:

$$f_X(x) = \sum_{i=1}^{\infty} f_X(x|A_i)\mathbf{P}(A_i),\ -\infty < x < \infty.$$

A different approach is required to define the conditional distribution function $F_{X|Y}(x|y)$ of random variable X given $Y = y$, where Y is another random variable. The difficulty is that if a random variable Y has a continuous distribution function, then the probability of the event $\{Y = y\}$ equals zero, and therefore the conditional distribution function $F_{X|Y}(x|y)$ cannot be defined with the help of the notion of conditional probability. In this case the conditional distribution function $F_{X|Y}(x|y)$ is defined as follows:

$$F_{X|Y}(x|y) = \lim_{\Delta y \to +0} \mathbf{P}(X \leq x|y \leq Y < y + \Delta y)$$

if the limit exists.

Let us assume that the joint density function $f_{XY}(x, y)$ of random variables X and Y exists. In such a case random variable X has the conditional CDF $F_{X|Y}(x|y)$ and conditional PDF $f_{X|Y}(x|y)$ given $Y = y$. If a joint PDF exists and $f_X(y) > 0$, then it is not difficult to see that the following relation holds:

$$\begin{aligned} F_{X|Y}(x|y) &= \lim_{\Delta y \to +0} \mathbf{P}(X \le x | y \le Y < y + \Delta y) \\ &= \lim_{\Delta y \to +0} \frac{\mathbf{P}(X \le x, y \le Y < y + \Delta y)}{\mathbf{P}(y \le Y < y + \Delta y)} \\ &= \lim_{\Delta y \to +0} \frac{\frac{F_{XY}(x, y+\Delta y) - F_{XY}(x,y)}{\Delta y}}{\frac{F_Y(y+\Delta y) - F_Y(y)}{\Delta y}} = \frac{1}{f_Y(y)} \frac{\partial}{\partial y} F_{XY}(x, y). \end{aligned}$$

From this relation we get the conditional PDF $f_{X|Y}(x|y)$ as follows:

$$f_{X|Y}(x|y) = \frac{\partial}{\partial x} F_{X|Y}(x|y) = \frac{1}{f_Y(y)} \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y) = \frac{f_{XY}(x, y)}{f_Y(y)}. \tag{1.1}$$

Independence of Random Variables Let X and Y be two random variables. Let $F_{XY}(x, y)$ be the joint distribution function of X and Y, and let $F_X(x)$ and $F_Y(y)$ be the marginal distribution functions.

Definition 1.27. Random variables X and Y are called independent of each other, or just independent, if the identity

$$F_{XY}(x, y) = F_X(x) F_Y(y)$$

holds for any $x, y, -\infty < x, y < \infty$.

In other words, random variables X and Y are independent if and only if the joint distribution function of X and Y equals the product of their marginal distribution functions.

The definition of independence of two random variables can be easily generalized to the case where an arbitrary collection of random variables $\{X_i,\ i \in I\}$ is given, analogously to the notion of the independence of events.

Definition 1.28. A collection of random variables $\{X_i,\ i \in I\}$ is called **mutually independent** (or just **independent**), if for any choice of a finite number of elements $X_{i_1}, \ldots, X_{i_n}$ the relation

$$F_{X_{i_1}, \ldots, X_{i_n}}(x_1, \ldots, x_n) = F_{X_{i_1}}(x_1) \cdot \ldots \cdot F_{X_{i_n}}(x_n),\ x_1, \ldots, x_n \in \mathbb{R}$$

holds.

Note that from the **pairwise independence** of random variables $\{X_i,\ i \in I\}$, which means that the condition

$$F_{X_{i_1},X_{i_2}}(x_1, x_2) = F_{X_{i_1}}(x_1)F_{X_{i_2}}(x_2),\ x_1, x_2 \in \mathbb{R},\ i_1, i_2 \in I,$$

is satisfied, mutual independence does not follow.

Example 1.29. Consider Example 1.11 given earlier and preserve the notation. Denote by $X_i = \mathcal{I}_{\{A_i\}}$ the indicator variables of the events $A_i,\ i = 1, 2, 3$. Then we can verify that random variables X_1, X_2, and X_3 are pairwise independent, but they do not satisfy mutual independence. The pairwise independence of random variables X_i can be easily proved. Since the events A_1, A_2, A_3 are independent and

$$\{X_i = 1\} = A_i \quad \text{and} \quad \{X_i = 0\} = \overline{A}_i,$$

then, using the relation proved in Example 1.11, we obtain for $i \neq j$

$$\begin{aligned}
\mathbf{P}\left(X_i = 1, X_j = 1\right) &= \mathbf{P}\left(A_i A_j\right) = \mathbf{P}\left(A_i\right)\mathbf{P}\left(A_j\right) = \frac{1}{4},\\
\mathbf{P}\left(X_i = 1, X_j = 0\right) &= \mathbf{P}\left(A_i \overline{A}_j\right) = \mathbf{P}\left(A_i\right)\mathbf{P}\left(\overline{A}_j\right) = \frac{1}{4},\\
\mathbf{P}\left(X_i = 0, X_j = 0\right) &= \mathbf{P}\left(\overline{A}_i \overline{A}_j\right) = \mathbf{P}\left(\overline{A}_i\right)\mathbf{P}\left(\overline{A}_j\right) = \frac{1}{4},
\end{aligned}$$

while, for example,

$$\begin{aligned}
&\mathbf{P}\left(X_1 = 1, X_2 = 1, X_3 = 1\right) = \mathbf{P}\left(A_1 A_2 A_3\right) = 0 \neq \frac{1}{8}\\
&\quad = \mathbf{P}\left(A_1\right)\mathbf{P}\left(A_2\right)\mathbf{P}\left(A_3\right) = \mathbf{P}\left(X_1 = 1\right)\mathbf{P}\left(X_2 = 1\right)\mathbf{P}\left(X_3 = 1\right).
\end{aligned}$$

Consider how we can characterize the notion of independence for two random variables in the discrete and continuous cases (if more than two random variables are given, then we may proceed in a similar manner).

Firstly, let us assume that the sets of values of discrete random variables X and Y are $\{x_i, i \geq 0\}$ and $\{y_j, j \geq 0\}$, respectively. If we denote the joint and marginal distributions of X and Y by

$$\begin{aligned}
&\left\{p_{ij} = \mathbf{P}\left(X = x_i, Y = y_j\right),\ i, j \geq 0\right\}, \left\{q_i = \mathbf{P}\left(X = x_i\right),\ i \geq 0\right\},\\
&\quad \text{and } \left\{r_j = \mathbf{P}\left(Y = y_j\right), j \geq 0\right\},
\end{aligned}$$

then the following assertion holds. Random variables X and Y are independent if and only if

$$p_{ij} = q_i r_j,\ i, j \geq 0.$$

Now assume that random variables X and Y have joint density $f_{XY}(x, y)$ and marginal densities $f_X(x)$ and $f_Y(y)$. Thus, in this case, random variables X and Y are independent if and only if the joint PDF takes a product form, that is,

$$f_{XY}(x, y) = f_X(x) f_Y(y), \ -\infty < x, y < \infty.$$

Convolution of Distributions Let X and Y be independent random variables with distribution functions $F_X(x)$ and $F_Y(y)$, respectively, and let us consider the distribution of the random variable $Z = X + Y$.

Definition 1.30. The distribution (CDF, PDF) of the random variable $Z = X+Y$ is called the **convolution** of the distribution (CDF, PDF), and the equations expressing the relation among them are called convolution formulas.

Definition 1.31. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with the common CDF F_X. The CDF F_X^{*n} of the sum $Z_n = X_1 + \ldots + X_n$ $(n \geq 1)$ is uniquely determined by F_X and is called the n**-fold convolution** of the CDF of F_X.

Note that the CDF $F_Z(z)$ of the random variable $Z = X + Y$, which is called the **convolution** of CDFs $F_X(x)$ and $F_Y(y)$, can be given in the general form

$$F_Z(z) = \mathbf{P}(Z \leq z) = \mathbf{P}(X + Y \leq z) = \int_{-\infty}^{\infty} F_X(z - y) \, \mathrm{d}F_Y(y).$$

This formula gets a simpler form in cases where the discrete random variables X and Y take only integer numbers, or if the PDFs $f_X(x)$ and $f_Y(y)$ of X and Y exist.

Let X and Y be independent discrete random variables taking values in $\{0, \pm 1, \pm 2, \ldots\}$ with probabilities $\{q_i = \mathbf{P}(X = x_i)\}$ and $\{r_j = \mathbf{P}(Y = y_j)\}$, respectively. Then the random variable $Z = X + Y$ takes values in $\{0, \pm 1, \pm 2, \ldots\}$, and its distribution satisfies the identity

$$s_k = \sum_{n=-\infty}^{\infty} q_{k-n} r_n, \ k = 0, \pm 1, \pm 2, \ldots .$$

If the independent random variables X and Y have a continuous distribution with the PDFs $f_X(x)$ and $f_Y(y)$, respectively, then random variable Z is continuous and its PDF $f_Z(z)$ can be given in the integral form

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) \mathrm{d}y.$$

Mixture of Distributions Let $F_1(x), \ldots, F_n(x)$ be a given collection of CDFs, and let $a_1, \ldots, a_n$ be nonnegative numbers with the sum $a_1 + \ldots + a_n = 1$. The function

$$F(x) = a_1 F_1(x) + \ldots + a_n F_n(x),\ -\infty < x < \infty,$$

is called a **mixture** of CDFs $F_1(x), \ldots, F_n(x)$ with weights $a_1, ..., a_n$.

Comment 1.32. *Any CDF can be given as a mixture of discrete, continuous, and singular CDFs, where the weights can also take a value of* 0.

Clearly, the function $F(x)$ possesses all the properties of CDFs; therefore it is also a CDF. In practice, the modeling of mixture distributions plays a basic role in stochastic simulation methods. A simple way to model mixture distributions is as follows.

Let us assume that the random variables $X_1, \ldots, X_n$ with distribution functions $F_1(x), \ldots, F_n(x)$ can be modeled. Let Y be a random variable taking values in $\{1, \ldots, n\}$ and independent of $X_1, \ldots, X_n$. Assume that Y has a distribution $P(Y = i) = a_i,\ 1 \le i \le n\ (a_i \ge 0,\ a_1 + \ldots + a_n = 1)$. Let us define random variable Z as follows:

$$Z = \sum_{i=1}^{n} \mathcal{I}_{\{Y=i\}} X_i,$$

where $\mathcal{I}_{\{\}}$ denotes the indicator variable. Then the CDF of random variable Z equals $F(z)$.

Proof. Using the formula of total probability, we have the relation

$$\mathbf{P}(Z \le z) = \sum_{i=1}^{n} \mathbf{P}(Z \le z | Y = i)\mathbf{P}(Y = i) = \sum_{i=1}^{n} \mathbf{P}(X_i \le z)a_i = F(z).$$

□

Concept and Properties of Expectation A random variable can be completely characterized in a statistical sense by its CDF. To define a distribution function $F(x)$, one needs to determine its values for all $x \in \mathbb{R}$, but this is not possible in many cases. Fortunately, there is no need to do so because in many cases it suffices to give some values that characterize the CDF in a certain sense depending on concrete practical considerations. One of the most important concepts is expectation, which we define in general form, and we give the definition for discrete and continuous distributions as special cases.

Definition 1.33. Let X be a random variable, and let $F_X(x)$ be its CDF. The **expected value** (or **mean value**) of random variable X is defined as

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

if the expectation exists.

Note that the finite expected value $\mathbf{E}(X)$ exists if and only if $\int_{-\infty}^{\infty} |x| \mathrm{d}F_X(x) < \infty$. It is conventional to denote the expected value of the random variable X by μ_X.

Expected Value of Discrete and Continuous Random Variables Let X be a discrete valued random variable with countable values $\{x_i, i \in I\}$ and with probabilities $\{p_i = \mathbf{P}(X = x_i), i \in I\}$. The finite expected value $\mathbf{E}(X)$ of random variable X exists and equals

$$\mathbf{E}(X) = \sum_{i \in I} p_i x_i$$

if and only if the sum is absolutely convergent, that is, $\sum_{i \in I} p_i |x_i| < \infty$. In the case of continuous random variables, the expected value can also be given in a simple form. Let $f_X(x)$ be the PDF of a random variable X. If the condition $\int_{-\infty}^{\infty} |x| f_X(x) \mathrm{d}x < \infty$ holds (i.e., the integral is absolutely convergent), then the finite expected value of X exists and can be given as

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

From a practical point of view, it is generally enough to give two special, discrete, and continuous cases. Let X be a random variable that has a mixed CDF with discrete and continuous components $F_1(x)$ and $F_2(x)$, respectively, and with weights a_1 and a_2, that is,

$$F(x) = a_1 F_1(x) + a_2 F_2(x),\ a_1, a_2 \geq 0,\ a_1 + a_2 = 1.$$

Assume that the set of discontinuities of $F_1(x)$ is $\{x_i,\ i \in I\}$ and denote $p_i = F_1(x_i) - F_1(x_i -), i \in I$. In addition, we assume that the continuous CDF $F_2(x)$ has the PDF $f(x)$. Then the expected value of random variable X is determined as follows:

$$\mathbf{E}(X) = a_1 \sum_{i \in I} p_i x_i + a_2 \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

if the series and the integral on the right-hand side of the last formula are absolutely convergent. The expected values related to special and different CDFs will be given later in this chapter.

The operation of expectation can be interpreted as a functional

$$\mathbf{E} : X \to \mathbf{E}(X)$$

that assigns a real value to the given random variable. We enumerate the basic properties of this functional as follows.

1. If random variable X is finite, i.e., if there are constants x_1 and x_2 for which the inequality $x_1 \le X \le x_2$ holds, then

$$x_1 \le \mathbf{E}(X) \le x_2.$$

If random variable X is nonnegative and the expected value $\mathbf{E}(X)$ exists, then

$$\mathbf{E}(X) \ge 0.$$

2. Let us assume that the expected value $\mathbf{E}(X)$ exists; then the expected value of random variable cX exists for an arbitrary given constant c, and the identity

$$\mathbf{E}(cX) = c\mathbf{E}(X)$$

is true.
3. If random variable X satisfies the condition $\mathbf{P}(X = c) = 1$, then

$$\mathbf{E}(X) = c.$$

4. If the expected values of random variables X and Y exist, then the sum $X + Y$ has an expected value, and the equality

$$\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y)$$

holds. This relation can usually be interpreted in such a way that the operation of expectation on the space of random variables is an additive functional.
5. The preceding properties can be expressed in a more general form. If there are finite expected values of random variables $X_1, \ldots, X_n$ and $c_1, \ldots, c_n$ are constants, then the equality

$$\mathbf{E}(c_1X_1 + \ldots + c_nX_n) = c_1\mathbf{E}(X_1) + \ldots + c_n\mathbf{E}(X_n)$$

holds. This property means that the functional $\mathbf{E}()$ is a linear one.
6. Let X and Y be independent random variables with finite expected value. Then the expected value of the product of random variables $X \cdot Y$ exists and equals the product of expected values, i.e., the equality

$$\mathbf{E}(XY) = \mathbf{E}(X) \cdot \mathbf{E}(Y)$$

is true.

Expectation of Functions of Random Variables, Moments and Properties Let X be a discrete random variable with finite or countable values $\{x_i, i \in I\}$ and with distribution $\{p_i, i \in I\}$. Let $h(x),\ x \in \mathbb{R}$ be a real-valued function for which the expected value of the random variable $Y = h(X)$ exists; then the equality

$$\mathbf{E}(Y) = \mathbf{E}(h(X)) = \sum_{i \in I} p_i h(x_i)$$

holds.

If the continuous random variable X has a PDF $f_X(x)$ and the expected value of the random variable $Y = h(X)$ exists, then the expected value of Y can be given in the form

$$\mathbf{E}(Y) = \int_{-\infty}^{\infty} h(x) f_X(x) \mathrm{d}x.$$

In cases where the expected value of functions of random variables (functions of random vectors) are investigated, analogous results to the one-dimensional case can be obtained. We give the formulas in connection with the two-dimensional case only. Let X and Y be two random variables, and let us assume that the expected value of the random variable $Z = h(X, Y)$ exists. With the appropriate notation, used earlier, for the cases of discrete and continuous distributions, the expected value of random variable Z can be given in the forms

$$\mathbf{E}(Z) = \sum_{i \in I} \sum_{j \in J} h(x_i, y_j) \mathbf{P}\left(X = x_i, Y = y_j\right),$$

$$\mathbf{E}(Z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f_{XY}(x, y) \mathrm{d}x \mathrm{d}y.$$

Consider the important case where h is a power function, i.e., for a given positive integer number k, $h(x) = x^k$. Assume that the expected value of X^k exists. Then the quantity

$$\mu_k = \mathbf{E}\left(X^k\right),\ k = 1, 2, \ldots,$$

is called the kth moment of random variable X. It stands to reason that the **first moment** $\mu = \mu_1 = \mathbf{E}\left(X^1\right)$ is the expected value of X and the frequently used **second moment** is $\mu_2 = \mathbf{E}\left(X^2\right)$.

Theorem 1.34. *Let j and k be integer numbers for which $1 \leq j \leq k$. If the kth moment of random variable X exists, then the jth moment also exists.*

Proof. From the existence of the kth moment it follows that $\mathbf{E}\left(|X|^k\right) < \infty$. Since $k/j \geq 1$, the function $x^{k/j},\ x \geq 0$, is convex, and by the use of Jensen's inequality we get the relation

$$\left[\mathbf{E}\left(|X|^j\right)\right]^{k/j} \leq \mathbf{E}\left(\left(|X|^j\right)^{k/j}\right) = \mathbf{E}\left(|X|^k\right) < \infty.$$

□

The kth central moment $\mathbf{E}\left((X-\mathbf{E}(X))^k\right)$ is also used in practice; it is defined as the kth moment of the random variable centered at the first moment (expected value). The kth central moment $\mathbf{E}\left((X-\mathbf{E}(X))^k\right)$ can be expressed by the noncentral moments $\mu_i, 1 \leq i \leq k$ of random variable X as follows:

$$\mathbf{E}\left((X-\mathbf{E}(X))^k\right) = \mathbf{E}\left(\sum_{i=0}^{k}\binom{k}{i}X^i(-\mathbf{E}(X))^{k-i}\right)$$
$$= \sum_{i=0}^{k}\binom{k}{i}\mathbf{E}\left(X^i\right)(-\mathbf{E}(X))^{k-i}.$$

In the course of a random experiment, the observed values fluctuate around the expected value. One of the most significant characteristics of the quantity of fluctuations is the variance. Assume that the second moment of random variable X is finite. Then the quantities

$$\mathbf{Var}(X) = \mathbf{E}\left((X-\mathbf{E}(X))^2\right)$$

are called the **variance** of random variable X. The **standard deviation** of a random variable X is the square root of its variance:

$$\mathbf{D}(X) = \sqrt{\mathbf{E}((X-\mathbf{E}(X))^2)}.$$

It is clear that the variance of X can be given with the help of the first and second moments as follows:

$$\mathbf{D}^2(X) = \mathbf{Var}(X) = \mathbf{E}\left((X-\mathbf{E}(X))^2\right) = \mathbf{E}\left(X^2\right) - 2\mathbf{E}(X)\cdot\mathbf{E}(X) + (\mathbf{E}(X)^2)$$
$$= \mathbf{E}\left(X^2\right) - (\mathbf{E}(X))^2 = \mu_2 - \mu^2.$$

It is conventional to denote the variance of the random variable X by $\sigma_X^2 = \mathbf{D}^2(X)$.

It should be noted that the variance of a random variable exists if and only if its second moment is finite. In addition, from the last inequality it follows that an upper estimation can be given for the variance as

$$\mathbf{D}^2(X) \leq \mathbf{E}\left(X^2\right).$$

It can also be seen that for every constant c the relation

$$\mathbf{E}\left((X-c)^2\right) = \mathbf{E}\left([(X-\mathbf{E}(X)) + (\mathbf{E}(X)-c)]^2\right) = \mathbf{D}^2(X) + (\mathbf{E}(X)-c)^2$$

holds, which is analogous to the Steiner formula, well known in the field of mechanics.

As an important consequence of this identity, we have the following result: the second moment $\mathbf{E}\left((X-c)^2\right)$ takes the minimal value for the constant $c = \mathbf{E}(X)$.

We will now mention some frequently used properties of variance.

1. If the variance of random variable X exists, then for all constants a and b the identity
$$\mathbf{D}^2(aX+b) = a^2\mathbf{D}^2(X)$$
is true.
2. Let $X_1, \ldots, X_n$ be independent random variables with finite variance; then
$$\mathbf{D}^2(X_1 + \ldots + X_n) = \mathbf{D}^2(X_1) + \ldots + \mathbf{D}^2(X_n). \tag{1.2}$$

The independence of random variables that play a role in formula (1.2) is not required for the last identity, and it is also true if instead of assuming the independence of the random variables $X_1, \ldots, X_n$ we assume that they are uncorrelated. The notion of correlation is to be defined later. If $X_1, \ldots, X_n$ are independent and identically distributed random variables with finite variance σ, then

$$\mathbf{D}^2(X_1 + \ldots + X_n) = \mathbf{D}^2(X_1) + \ldots + \mathbf{D}^2(X_n) = n\sigma^2,$$

from which
$$\mathbf{D}(X_1 + \ldots + X_n) = \sigma\sqrt{n}$$
follows.

In the literature on queueing theory, the notion of **relative variance** $\mathbf{CV}(X)^2$ is applied, which is defined as

$$\mathbf{CV}(X)^2 = \frac{\mathbf{D}^2(X)}{\mathbf{E}(|X|)^2} \ .$$

Its square root $\mathbf{CV}(X) = \mathbf{D}(X)/\mathbf{E}(|X|)$ is called the **coefficient of variation**, which serves as a normalized measure of variance of a distribution. The following inequalities hold:

Exponential distribution:	$CV = 1,$
Hyperexponential distribution:	$CV > 1,$
Erlang distribution:	$CV < 1.$

Markov and Chebyshev Inequalities The role of the Markov and Chebyshev inequalities is significant, not only because they provide information concerning distributions with the help of expected value and variance but because they are also effective tools for proving certain results.

Theorem 1.35 (*Markov inequality*). *If the expected value of a nonnegative random variable X exists, then the following inequality is true for any constant $\varepsilon > 0$,:*

$$\boldsymbol{P}(X \geq \varepsilon) \leq \frac{\boldsymbol{E}(X)}{\varepsilon}.$$

Proof. For an arbitrary positive constant $\varepsilon > 0$ we have the relation

$$\mathbf{E}(X) \geq \mathbf{E}\left(X\mathcal{I}_{\{X \geq \varepsilon\}}\right) \geq \varepsilon \mathbf{E}\left(\mathcal{I}_{\{X \geq \varepsilon\}}\right) = \varepsilon\, \mathbf{P}(X \geq \varepsilon),$$

from which the Markov inequality immediately follows. □

Theorem 1.36 (*Chebyshev inequality*). *If the variance of random variable X is finite, then for any constant $\varepsilon > 0$ the inequality*

$$\boldsymbol{P}(|X - \boldsymbol{E}(X)| \geq \varepsilon) \leq \frac{\boldsymbol{D}^2(X)}{\varepsilon^2}$$

holds.

Proof. Using the Markov inequality for a constant $\varepsilon > 0$ and for the random variable $(X - \mathbf{E}(X))^2$ we find that

$$\mathbf{P}(|X - \mathbf{E}(X)| \geq \varepsilon) = \mathbf{P}\left((X - \mathbf{E}(X))^2 \geq \varepsilon^2\right) \leq \frac{\mathbf{E}(X - \mathbf{E}(X))^2}{\varepsilon^2} = \frac{\mathbf{D}^2(X)}{\varepsilon^2},$$

from which the assertion of the theorem follows. □

Comment 1.37. *Let X be a random variable. If $h(x)$ is a convex function and $\boldsymbol{E}(h(X))$ exists, then the Jensen inequality $\boldsymbol{E}(h(X)) \geq h(\boldsymbol{E}(X))$ is true. Using this inequality we can obtain some other relations, similar to the case of the Markov inequality.*

Example 1.38. As a simple application of the Chebyshev inequality, let us consider the average $(X_1 + \ldots + X_n)/n$, where the random variables $X_1, \ldots, X_n$ are independent identically distributed with finite second moment. Let us denote the joint expected value and variance by μ and σ^2, respectively. Using the property (1.2) of variance and the Chebyshev inequality and applying $(n\varepsilon)$ instead of ε, we get the inequality

$$\begin{aligned}\mathbf{P}(|X_1 + \ldots + X_n - n\mu| \geq n\varepsilon) &= \mathbf{P}\left((X_1 + \ldots + X_n - n\mu)^2 \geq n^2\varepsilon^2\right) \\ &\leq \frac{n\sigma^2}{(n\varepsilon)^2} = \frac{\sigma^2}{n\varepsilon^2};\end{aligned}$$

then

$$\mathbf{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

As a consequence of the last inequality, for every fixed positive constant ε the probability $\mathbf{P}\left(\left|\frac{X_1+\ldots+X_n}{n}-\mu\right|\geq\varepsilon\right)$ tends to 0 as n goes to infinity. This assertion is known as the **weak law of large numbers**.

Generating and Characteristic Functions So far, certain quantities characterizing the distribution of random variables have been provided. Now such transformations of distributions will be given where the distributions and the functions obtained by the transformations uniquely determine each other. The investigated transformations provide effective tools for determining, for instance, distributions and moments and for proving limit theorems.

Definition 1.39. Let X be a random variable taking values in $\{0,1,\ldots\}$, with probabilities $p_0, p_1, \ldots$. Then the power series

$$G_X(z)=\mathbf{E}\left(z^X\right)=\sum_{i=0}^{\infty}p_i z^i$$

is convergent for all $z\in[-1,1]$, and the function $G_X(z)$ is called the **probability generating function** (or just **generating function**) of the discrete random variable X.

In engineering practice, the power series defining the generating function is applied in a more general approach instead of in the interval $[-1,1]$, and the generating function is defined on the closed complex unit circle $z\in\mathbb{C}$, $|z|\leq 1$, which is usually called a z**-transform** of the distribution $\{p_i, i=0,1,\ldots\}$. This notion is also applied if, instead of a distribution, a transformation is made to an arbitrary sequence of real numbers.

It should be noted that $|G_X(z)|\leq 1$ if $z\in\mathbb{C}$ and the function $G_X(z)$ is differentiable on the open unit circle of the complex plane $z\in\mathbb{C}$, $|z|<1$ infinitely many times and the kth derivative of $G_X(z)$ equals the sum of the kth derivative of the members of the series.

It is clear that

$$p_k=G_X^{(k)}(0)/k!,\ k=0,1,\ldots\ .$$

This formula makes it possible to compute the distribution if the generating function is given. It is also true that if the first and second derivatives $G'_X(1-)$ and $G''_X(1-)$ exist on the left-hand side at $z=1$, then the first and second moments of random variable X can be computed as follows:

$$\mathbf{E}(X)=G'_X(1-)\quad\text{and}\quad\mathbf{E}\left(X^2\right)=\left(zG'_X(z)\right)'\Big|_{z=1}=G''_X(1-)+G'_X(1-).$$

From this we can obtain the variance of X as follows:

$$\mathbf{D}^2(X)=G''_X(1-)+G'_X(1-)-(G'_X(1-))^2.$$

It can also be verified that if the nth derivative of the generating function $G_X(z)$ exists on the left-hand side at $z = 1$, then

$$\begin{aligned}\mathbf{E}\left(X(X-1)\dots(X-m+1)\right) &= \sum_{k=m}^{\infty} k(k-1)\dots(k-m+1)\,p_k \\ &= G_X^{(m)}(1-),\ 1 \le m \le n.\end{aligned}$$

Computing the expected values on the left-hand side of these identities, we can obtain linear equations between the moments $\mu_k = \mathbf{E}(X^k)$, $1 \le k \le m$, and the derivatives $G_X^{(m)}(1-)$ for all $1 \le m \le n$. The moments μ_m, $m = 1, 2, \dots, n$ can be determined in succession with the help of the derivatives $G_X^{(k)}(1-)$, $1 \le k \le m$. The special cases of $k = 1, 2$ give the preceding formulas for the first and second moments.

Characteristic Function

Definition 1.40. The complex valued function

$$\varphi_X(s) = \mathbf{E}\left(\mathrm{e}^{isX}\right) = \mathbf{E}\left(\cos(sX)\right) + i\,\mathbf{E}\left(\sin(sX)\right),\ s \in \mathbb{R},$$

is called the **characteristic function** of random variable X, where $i = \sqrt{-1}$.

Note that a characteristic function can be rewritten in the form

$$\varphi_X(s) = \int_{-\infty}^{\infty} \mathrm{e}^{isx}\,\mathrm{d}F_X(x),$$

which is the well-known Fourier–Stieltjes transform of the CDF $F_X(x)$.
Using conventional notation, in discrete and continuous cases we have

$$\varphi_X(s) = \sum_{k=0}^{\infty} p_k \mathrm{e}^{isx_k}, \quad \text{and} \quad \varphi_X(s) = \int_{-\infty}^{\infty} \mathrm{e}^{isx} f_X(x)\mathrm{d}x.$$

The characteristic function and the CDFs determine each other uniquely. Now some important properties of characteristic functions will be enumerated.

1. The characteristic function is real valued if and only if the distribution is symmetric.
2. If the kth moment $\mathbf{E}\left(X^k\right)$ exists at point 0, then

$$\mathbf{E}\left(X^k\right) = \frac{\varphi_X^{(k)}(0)}{i^k}.$$

3. If the derivative $\varphi_X^{(2k)}(0)$ is finite for a positive integer k, then the moment $\mathbf{E}\left(X^{2k}\right)$ exists. Note that from the existence of the finite derivative $\varphi_X^{(2k+1)}(0)$ only the existence of the finite moment $\mathbf{E}\left(X^{2k}\right)$ follows.
4. Let $X_1, \ldots, X_n$ be independent random variables; then the characteristic function of the sum $X_1 + \ldots + X_n$ equals the product of the characteristic functions of the random variables X_i, that is,

$$\begin{aligned}\varphi_{X_1+\ldots+X_n}(s) &= \mathbf{E}\left(\mathrm{e}^{is(X_1+\ldots+X_n)}\right) = \mathbf{E}\left(\mathrm{e}^{isX_1}\ldots\mathrm{e}^{isX_n}\right)\\ &= \mathbf{E}\left(\mathrm{e}^{isX_1}\right)\cdot\ldots\cdot\mathbf{E}\left(\mathrm{e}^{isX_n}\right) = \varphi_{X_1}(s)\ldots\varphi_{X_1}(s).\end{aligned}$$

Note that property 4 plays an important role in the limit theorems of probability theory.

Laplace–Stieltjes and Laplace Transforms If, instead of the CDFs, the Laplace–Stieltjes and Laplace transforms were used, the problem could be solved much easier in many practical cases and the results could additionally often be given in more compact form. Let X be a nonnegative random variable with the CDF $F(x)$ ($F(0) = 0$). Then the real or, in general, complex varying function

$$F^{\sim}(s) = \mathbf{E}\left(\mathrm{e}^{-sX}\right) = \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}F(x),\ \mathrm{Re}s \geq 0,\ F^{\sim}(0) = 1$$

is called the **Laplace–Stieltjes transform** of the CDF F. Since $\left|\mathrm{e}^{-sX}\right| \leq 1$ if $\mathrm{Re}s \geq 0$, then the function $F^{\sim}(s)$ is well defined. If f is a PDF, then the function

$$f^*(s) = \int_0^\infty \mathrm{e}^{-sx} f(x)\mathrm{d}x,\ \mathrm{Re}s \geq 0,$$

is called the **Laplace transform** of the function f. These notations will be used even if the functions F and f do not possess the necessary properties of distribution and PDFs but $F^{\sim}(s)$ and $f^*(s)$ are well defined. If f is a PDF related to the CDF F, then the equality

$$F^{\sim}(s) = f^*(s) = sF^*(s) \tag{1.3}$$

holds.

Proof. It is clear that

$$F^{\sim}(s) = \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}F(x) = \int_0^\infty \mathrm{e}^{-sx} f(x)\mathrm{d}x = f^*(s),$$

and integrating by parts we have

$$F^{\sim}(s) = \int_0^{\infty} \mathrm{e}^{-sx} \mathrm{d}F(x) = \int_0^{\infty} s\mathrm{e}^{-sx} F(x)\mathrm{d}x = sF^*(s).$$

□

Since the preceding equation is true between the two introduced transforms, it is enough to consider the Laplace–Stieltjes transform only and to enumerate its main properties.

(a) $F^{\sim}(s)$, $\mathrm{Re}s \geq 0$ is a continuous function and $0 \leq |F^{\sim}(s)| \leq 1$, $\mathrm{Re}s \geq 0$.
(b) $F^{\sim}_{aX+b}(s) = \mathrm{e}^{-bs} F^{\sim}(as)$.
(c) For all positive integers k

$$(-1)^k F^{\sim(k)}(s) = \int_0^{\infty} x^k \mathrm{e}^{-sx} \mathrm{d}F(x), \quad \mathrm{Re}s > 0.$$

If the kth moment $\mu_k = \mathbf{E}\left(X^k\right)$ exists, then $\mu_k = (-1)^k F^{\sim(k)}(0)$.
(d) If the nonnegative random variables X and Y are independent, then

$$F^{\sim}_{X+Y}(s) = F^{\sim}_X(s) F^{\sim}_Y(s).$$

(e) For all continuity points of the CDF F the inversion formula

$$F(x) = \lim_{a\to\infty} \sum_{n\leq ax} (-1)^n (F^{\sim}(a))^{(n)} \frac{a^n}{n!}$$

is true.

Covariance and Correlation Let X and Y be two random variables with finite variances σ_X^2 and σ_Y^2, respectively. The **covariance** between the pair of random variables (X, Y) is defined as

$$\mathrm{cov}(X, Y) = \mathbf{E}\left((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))\right).$$

The covariance can be rewritten in the simple computational form

$$\mathrm{cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y).$$

If the variances σ_X^2 and σ_Y^2 satisfy the conditions $\mathbf{D}(X) > 0$, $\mathbf{D}(Y) > 0$, then the quantity

$$\mathrm{corr}(X, Y) = \mathrm{cov}\left(\frac{X - \mathbf{E}(X)}{\mathbf{D}(X)}, \frac{Y - \mathbf{E}(Y)}{\mathbf{D}(Y)}\right) = \frac{\mathrm{cov}(X, Y)}{\mathbf{D}(X)\mathbf{D}(Y)}$$

is called the **correlation** between the pair of random variables (X, Y).

Correlation can be used as a measure of the dependence between random variables. It is always true that

$$-1 \leq \operatorname{corr}(X, Y) \leq 1,$$

provided that the variances of random variables X and Y are finite and nonzero.

Proof. Since by the Cauchy–Schwartz inequality for all random variables U and V with finite second moments

$$(\mathbf{E}(UV))^2 \leq \mathbf{E}\left(U^2\right)\mathbf{E}\left(V^2\right),$$

therefore

$$(\operatorname{cov}(X, Y))^2 \leq \mathbf{E}\left((X - \mathbf{E}(X))^2\right)\mathbf{E}\left((Y - \mathbf{E}(Y))^2\right) = \mathbf{D}^2(X)\mathbf{D}^2(Y),$$

from which the inequality $|\operatorname{corr}(X, Y)| \leq 1$ immediately follows. □

It can also be proved that the equality $|\operatorname{corr}(X, Y)| = 1$ holds if and only if a linear relation exists between random variables X and Y with probability 1, that is, there are two constants a and b for which $\mathbf{P}(Y = aX + b) = 1$.

Both covariance and correlation play essential roles in multivariate statistical analysis. Let $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ be a column vector whose n elements $X_1, \ldots, X_n$ are random variables. Here it should be noted that in probability theory and statistics usually column vectors are applied, but in queueing theory row vectors are used if Markov processes are considered. We define

$$\mathbf{E}(\mathbf{X}) = (\mathbf{E}(X_1), \ldots, \mathbf{E}(X_n))^{\mathrm{T}},$$

provided that the expected values of components exist. The upper index T denotes the transpose of vectors or matrices. Similarly, if a matrix $W = \left(W_{ij}\right) \in \mathbb{R}^{k \times m}$ is given whose elements W_{ij} are random variables of finite expected values, then we define

$$\mathbf{E}(W) = \left(\mathbf{E}\left(W_{ij}\right)\right),\ 1 \leq i \leq k,\ 1 \leq j \leq m).$$

If the variances of components of a random vector $X = (X_1, \ldots, X_k)^{\mathrm{T}}$ are finite, then the matrix

$$R = \mathbf{E}\left((X - \mathbf{E}(X))(X - \mathbf{E}(X))^{\mathrm{T}}\right) \tag{1.4}$$

is called a **covariance matrix** of X. It can be seen that the (i, j) entries of matrix R are $R_{ij} = \operatorname{cov}(X_i, X_j)$, which are the covariances between the random variables X_i and X_j.

The covariance matrix can be defined in cases where the components of X are complex valued random variables replacing in definition (1.4) $(X - \mathbf{E}(X))^{\mathrm{T}}$ by $(X - \mathbf{E}(X))^{*\mathrm{T}}$ the complex composed of transpose.

An important property of a covariance matrix R is that it is nonnegative definite, i.e., for all real or complex k-dimensional column vectors $z = (z_1, \dots, z_k)^{\mathrm{T}}$ the inequality

$$zRz^{\mathrm{T}} \geq 0$$

holds.

The matrix $r = (r_{i,j})$ with components $r_{i,j} = \operatorname{corr}(X_i, X_j)$, $1 \leq i \leq k$, $1 \leq j \leq m$ is called a **correlation matrix** of random vector X.

Conditional Expectation and Its Properties The notion of conditional expectation is defined with the help of results of set and measure theories. We present the general concept and important properties and illustrate the important special cases.

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a fixed probability space, and let X be a random variable whose expected value exists. Let $\mathcal{C}$ be an arbitrary sub-σ-algebra of $\mathcal{A}$. We wish to define the conditional expectation $Z = \mathbf{E}(X|\mathcal{C})$ of X given $\mathcal{C}$ as a $\mathcal{C}$-measurable random variable for which the random variable satisfies the condition $\mathbf{E}\left(\mathbf{E}(X|\mathcal{C})\mathcal{I}_{\{C\}}\right) = \mathbf{E}\left(X\mathcal{I}_{\{C\}}\right)$ for all $C \in \mathcal{C}$. As a consequence of the Radon–Nikodym theorem, a random variable Z exists with probability 1 that satisfies the required conditions.

Definition 1.41. Random variable Z is called the **conditional expectation** of X given σ-algebra $\mathcal{C}$ if the following conditions hold:

(a) Z is a $\mathcal{C}$-measurable random variable.
(b) $\mathbf{E}\left(\mathbf{E}(X|\mathcal{C})\mathcal{I}_{\{C\}}\right) = \mathbf{E}\left(X\mathcal{I}_{\{C\}}\right)$ for all $C \in \mathcal{C}$.

Definition 1.42. Let $A \in \mathcal{A}$ be an event. The random variable $\mathbf{P}(A|\mathcal{C}) = \mathbf{E}\left(\mathcal{I}_{\{A\}}|\mathcal{C}\right)$ is called the **conditional expectation** of event A given σ-algebra $\mathcal{C}$.

Important Properties of Conditional Expectation Let $\mathcal{C}$, $\mathcal{C}_1$, and $\mathcal{C}_2$ be sub-σ-algebras of $\mathcal{A}$, and let X, X_1, and X_2 be random variables with finite expected values. Then the following relations hold with probability 1:

1. $\mathbf{E}(\mathbf{E}(X|\mathcal{C})) = \mathbf{E}(X)$.
2. $\mathbf{E}(cX|\mathcal{C}) = c\mathbf{E}(X|\mathcal{C})$ for all constant c.
3. If $\mathcal{C}_0 = \{\varnothing, \Omega\}$ is the trivial σ-algebra, then $\mathbf{E}(X|\mathcal{C}_0) = \mathbf{E}(X)$.
4. If $\mathcal{C}_1 \subset \mathcal{C}_2$, then $\mathbf{E}(\mathbf{E}(X|\mathcal{C}_1)|\mathcal{C}_2) = \mathbf{E}(\mathbf{E}(X|\mathcal{C}_2)|\mathcal{C}_1) = \mathbf{E}(X|\mathcal{C}_1)$.
5. If random variable X does not depend on the σ-algebra $\mathcal{C}$, i.e., if for all Borel sets $D \in \mathcal{B}$ and for all events $A \in \mathcal{C}$ the equality $\mathbf{P}(X \in D, A) = \mathbf{P}(X \in D)\mathbf{P}(A)$ holds, then $\mathbf{E}(X|\mathcal{C}) = \mathbf{E}(X)$.
6. $\mathbf{E}(X_1 + X_2|\mathcal{C}) = \mathbf{E}(X_1|\mathcal{C}) + \mathbf{E}(X_2|\mathcal{C})$.
7. If the random variable X_1 is $\mathcal{C}$-measurable, then $\mathbf{E}(X_1X_2|\mathcal{C}) = X_1\mathbf{E}(X_2|\mathcal{C})$.

Definition 1.43. Let Y be a random variable, and denote by $\mathcal{A}_Y$ the σ-algebra generated by random variable Y, i.e., let $\mathcal{A}_Y$ be the minimal sub-σ-algebra of $\mathcal{A}$ for which Y is $\mathcal{A}_Y$-measurable. The random variable $\mathbf{E}(X|Y) = \mathbf{E}(X|\mathcal{C}_Y)$ is called the **conditional expectation** of X given random variable Y.

Main Properties of Conditional Expectation Firstly, consider the case where random variable Y is discrete and takes values in the set $\mathcal{Y} = \{y_1, \dots, y_n\}$ and

$\mathbf{P}(Y = y_i) > 0$, $1 \le i \le n$. We then define the events $C_i = \{Y = y_i\}$, $1 \le i \le n$. It is clear that the collection of events $\{C_1, \ldots, C_n\}$ forms a complete system of events, i.e., they are mutually exclusive, $\mathbf{P}(C_i) > 0$, $\le i \le n$ and $\mathbf{P}(C_1) + \ldots + \mathbf{P}(C_n) = 1$. The σ-algebra $\mathcal{C}_Y = \sigma(C_1, \ldots, C_n) \subset \mathcal{A}$, which is generated by random variable Y, is the set of events consisting of all subsets of $\{C_1, \ldots, C_n\}$. Note that here we can write "algebra" instead of "σ-algebra" because the set $\{C_1, \ldots, C_n\}$ is finite. Since the events C_i have positive probability, the conditional probabilities

$$\mathbf{E}(X|C_i) = \frac{\mathbf{E}\left(X\mathcal{I}_{\{C_i\}}\right)}{\mathbf{P}(C_i)}$$

are well defined.

Theorem 1.44. *The conditional expectation $\boldsymbol{E}(X|\mathcal{C}_Y)$ satisfies the relation*

$$\boldsymbol{E}(X|\mathcal{C}_Y) = \boldsymbol{E}(X|\mathcal{C}_Y)(\omega) = \sum_{k=1}^{n} \boldsymbol{E}(X|C_k)\mathcal{I}_{\{C_k\}} \text{ with probability } 1. \tag{1.5}$$

Note that Eq. (1.5) can also be rewritten in the form

$$\mathbf{E}(X|Y) = \mathbf{E}(X|Y)(\omega) = \sum_{k=1}^{n} \mathbf{E}(X|Y = y_k)\mathcal{I}_{\{Y=y_k\}}. \tag{1.6}$$

Proof. Since the relation

$$\{\mathbf{E}(X|\mathcal{C}_Y) \le x\} = \cup\{C_i : \ \mathbf{E}(X|C_i) \le x\} \in \mathcal{C}_Y$$

holds for all $x \in \mathbb{R}$, then $\mathbf{E}(X|\mathcal{C}_Y)$ is a $\mathcal{C}_Y$-measurable random variable. On the other hand, if $C \in \mathcal{C}_Y$, $C \ne \{\varnothing\}$, then $C = \cup\{C_i : i \in K\}$ stands with an appropriately chosen set of indices $K \subset \{1, \ldots, n\}$, and we obtain

$$\begin{aligned}\mathbf{E}\left(\mathbf{E}(X|\mathcal{C}_Y)\mathcal{I}_{\{C\}}\right) &= \mathbf{E}\left(\sum_{k\in K} \mathbf{E}(X|C_k)\mathcal{I}_{\{C_k\}}\right) \\ &= \sum_{k\in K} \mathbf{E}(X|C_k)\mathbf{P}(C_k) = \sum_{k\in K} \mathbf{E}\left(X\mathcal{I}_{\{C_k\}}\right) = \mathbf{E}\left(X\mathcal{I}_{\{C\}}\right).\end{aligned}$$

If $C = \{\varnothing\}$, then $\mathbf{E}\left(\mathbf{E}(X|\mathcal{C}_Y)\mathcal{I}_{\{C\}}\right) = \mathbf{E}\left(X\mathcal{I}_{\{C\}}\right) = 0$. Thus we have proved that random variable (1.5) satisfies all the required properties of conditional expectation. □

Comment 1.45. *From expression (1.6) the following relation can be obtained:*

$$\boldsymbol{E}(X) = \boldsymbol{E}(\boldsymbol{E}(X|Y)) = \int_{-\infty}^{\infty} \boldsymbol{E}(X|Y = y)\mathrm{d}F_Y(y). \tag{1.7}$$

This relation remains valid if, instead of the finite set $\mathcal{Y} = \{y_1, \ldots, y_n\}$*, we choose the countable infinite set* $\mathcal{Y} = \{y_i,\ i \in I\}$ *for which* $\boldsymbol{P}(Y = y_i) > 0$*,* $i \in I$*.*

Comment 1.46. *Denote the function* g *by the relation*

$$g(y) = \begin{cases} \boldsymbol{E}(X|Y = y_k), & \text{if } y = y_k \text{ for an index } k, \\ 0, & \text{otherwise.} \end{cases} \tag{1.8}$$

Then, using formula (1.6), the conditional expectation of X *given* Y *can be obtained with the help of the function* g *as follows:*

$$\boldsymbol{E}(X|Y) = g(Y) \tag{1.9}$$

with probability 1.

Continuous Random Variables (X, Y) Consider a pair of random variables (X, Y) having joint density $f_{X,Y}(x, y)$ and marginal densities $f_X(x)$ and $f_Y(y)$, respectively. Then the conditional density $f_{X|Y}(x|y)$ exists and, according to (1.1), can be defined as

$$f_{X|Y}(x|y) = \begin{cases} \dfrac{f_{XY}(x, y)}{f_Y(y)}, & \text{if } f_Y(y) > 0, \\ 0, & \text{otherwise}. \end{cases}$$

Define $g(y) = \mathbf{E}(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\mathrm{d}x$. Then the conditional expectation of X given Y can be determined with probability 1 as follows:

$$\mathbf{E}(X|Y) = g(Y),$$

and so we can define

$$\mathbf{E}(X|Y = y) = g(y).$$

Proof. It is clear that $g(Y)$ is a $\mathcal{C}_Y$-measurable random variable; therefore, it is enough to prove that the equality

$$\mathbf{E}\left(\mathbf{E}(X|Y)\mathcal{I}_{\{Y \in D\}}\right) = \mathbf{E}\left(X\mathcal{I}_{\{Y \in D\}}\right)$$

holds for all Borel sets D of a real line. It is not difficult to see that

$$\mathbf{E}\left(\mathbf{E}(X|Y)\mathcal{I}_{\{Y \in D\}}\right) = \mathbf{E}\left(g(Y)\mathcal{I}_{\{Y \in D\}}\right)$$

$$= \int_D \int_{-\infty}^{\infty} x \, \frac{f_{XY}(x,y)}{f_Y(y)} \, f_Y(y) \mathrm{d}x \mathrm{d}y$$

$$= \int_D \int_{-\infty}^{\infty} x f_{XY}(x,y) \mathrm{d}x \mathrm{d}y$$

and, on other hand,

$$\mathbf{E}\left(X \mathcal{I}_{\{Y \in D\}}\right) = \int_D \int_{-\infty}^{\infty} x f_{XY}(x,y) \mathrm{d}x \mathrm{d}y.$$

□

Comment 1.47. *In the case where a pair of random variables has a joint normal distribution, the conditional expectation* $\boldsymbol{E}(X|Y)$ *is a linear function of random variable* Y *with probability* 1*, that is, the regression function* g *is a linear function and the relation*

$$\boldsymbol{E}(X|Y) = \boldsymbol{E}(X) + \frac{\operatorname{cov}(X,Y)}{\boldsymbol{D}(X)}(X - \boldsymbol{E}(X))$$

holds.

General Case By the definition of conditional expectation, $\mathbf{E}(X|Y)$ is $\mathcal{C}_Y$-measurable; therefore, there is a Borel-measurable function g such that $\mathbf{E}(X|Y)$ can be given with probability 1 in the form

$$\mathbf{E}(X|Y) = g(Y). \tag{1.10}$$

This relation makes it possible to give the conditional expectation $\mathbf{E}(X|Y = y)$ as the function

$$\mathbf{E}(X|Y = y) = g(y),$$

which is called a **regression function**. It is clear that the regression function is not necessarily unique and is determined on a Borel set of the real line D satisfying the condition $\mathbf{P}(Y \in D) = 1$.

Comment 1.48. *Let* X *and* Y *be two random variables. Assume that* X *has finite variation. Consider the quadratic distance* $\boldsymbol{E}\left([X - h(Y)]^2\right)$ *for the set* $\mathcal{H}_Y$ *of all Borel-measurable functions* h*, for which* $h(Y)$ *has finite variation. Then the assertion*

$$\min\left\{\boldsymbol{E}\left([X - h(Y)]^2\right) : \; h \in \mathcal{H}_Y\right\} = \boldsymbol{E}\left([X - g(Y)]^2\right)$$

holds. This relation implies that the best approximation of X by Borel-measurable functions of Y in a quadratic mean is the regression $\boldsymbol{E}(X|Y) = g(Y)$.

Formula of Total Expected Value A useful formula can be given to compute the expected value of random variable X if the regression function $\mathbf{E}(X|Y = y)$ can be determined.

Making use of relation 1 given as a general property of conditional expectation and Eq. (1.10), it is clear that

$$\begin{aligned}\mathbf{E}(X) &= \mathbf{E}(\mathbf{E}(X|Y)) = \mathbf{E}(g(Y)) \\ &= \int_{-\infty}^{\infty} g(y)\mathrm{d}F_Y(y) = \int_{-\infty}^{\infty} \mathbf{E}(X|Y = y)\mathrm{d}F_Y(y).\end{aligned}$$

From this relation we have the so-called **formula of total expected value**. If random variable Y has discrete or continuous distributions, then we have the formulas

$$\mathbf{E}(X) = \sum_{i \in I} \mathbf{E}(X|Y = y_i)\mathbf{P}(Y = y_i)$$

and

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} \mathbf{E}(X|Y = y) f_Y(y)\mathrm{d}y.$$

1.2 Frequently Used Discrete and Continuous Distributions

In this part we consider some frequently used distributions and give their definitions and important characteristics. In addition to the formal description of the distributions, we will give appropriate mathematical models that lead to a given distribution. If the distribution function of a random variable is given as a function $F_X(x; a_1, \ldots, a_n)$ depending on a positive integer n and constants $a_1, \ldots, a_n$, then $a_1, \ldots, a_n$ are called the parameters of the density function F_X.

1.2.1 Discrete Distributions

Bernoulli Distribution $Be(p)$, $0 \leq p \leq 1$. The PDF of random variable X with values $\{0, 1\}$ is called a Bernoulli distribution if

$$p_k = \mathbf{P}(X = k) = \begin{cases} p, & \text{if } k = 1, \\ 1 - p, & \text{if } k = 0. \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = p,\ \mathbf{D}^2(X) = p(1-p)$;
Generating function: $1 - p + pz$;
Characteristic function: $1 - p + pe^{it}$.

Example. Let X be the number of heads appearing in one toss of a coin, where

$$p = \mathbf{P}(\text{head appearing in a toss}).$$

Then X has a $Be(p)$ distribution.

Binomial Distribution $B(n, p)$. The distribution of a discrete random variable X with values $\{0, 1, \ldots, n\}$ is called binomial with the parameters n and p, $0 < p < 1$, if its PDF is

$$p_k = \mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k},\ \ k = 0, 1, \ldots, n.$$

Expected value and variance: $\mathbf{E}(X) = np,\ \ \mathbf{D}^2(X) = np(1-p)$;
Generating function: $G(z) = (pz + (1-p))^n$;
Characteristic function: $\varphi(t) = (1 + p(e^{it} - 1))^n$.

Example. Consider an experiment in which we observe that an event A with probability $p = \mathbf{P}(A)$, $0 < p < 1$, occurs (success) or not (failure). Repeating the experiment n times independently, define random variable X by the frequency of event A. Then the random variable has a $B(n, p)$ PDF.

Note that if the $Be(n, p)$ random variables $X_1, \ldots, X_n$ are independent, then the random variable $X = X_1 + \ldots + X_n$ has a $B(n, p)$ distribution.

Polynomial Distribution The PDF of a random vector $X = (X_1, \ldots, X_k)^{\mathrm{T}}$ taking values in the set $\{(n_1, \ldots, n_k) : n_i \geq 0,\ n_1 + \ldots + n_k = n\}$ is called polynomial with the parameters n and $p_1, \ldots, p_k$ ($p_i > 0,\ p_1 + \ldots + p_k = 1$) if X has a PDF

$$p_{n_1,\ldots,n_k} = \mathbf{P}(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k}.$$

Note that each coordinate variable X_i of random vector X has a $B(p_i, n)$ binomial distribution whose expected value and variance are np_i and $np_i(1-p_i)$.

Expected value $\mathbf{E}(X) = (np_1, \ldots, np_n)^{\mathrm{T}}$;
Covariance matrix $R = (R_{ij})_{1 \leq i,j \leq k}$, where $R_{ij} = \begin{cases} np_i(1-p_i), & \text{if } i = j, \\ np_i p_j, & \text{if } i \neq j; \end{cases}$
Characteristic function: $\varphi(t_1, \ldots, t_k) = (p_1 e^{it_1} + \ldots + p_k e^{it_k})^n$.

Example. Let $A_1, \ldots, A_k$ be k disjoint events for which $p_i = \mathbf{P}(A_i) > 0,\ p_1 + \ldots + p_k = 1$. Consider an experiment with possible outcomes $A_1, \ldots, A_k$ and repeat it n times independently. Denote by X_i the frequency of event A_i in the series of n

observations. Then the distribution of X is polynomial with the parameters n and $p_1, \ldots, p_k$.

Geometric Distribution The PDF of random variable X taking values in $\{1, 2, \ldots\}$ is called a geometric distribution with the parameter p, $0 < p < 1$, if its PDF is

$$p_k = \mathbf{P}(X = k) = (1-p)^{k-1} p,\ k = 1, 2, \ldots .$$

Expected value and variance:	$\mathbf{E}(X) = \frac{1}{p}$, $\mathbf{D}^2(X) = \frac{1-p}{p^2}$;
Generating function:	$G(z) = \frac{pz}{1-(1-p)z}$;
Characteristic function:	$\varphi(t) = \frac{p}{1-(1-p)e^{it}}$.

Theorem 1.49. *If X has a geometric distribution, then X has a so-called **memoryless property**, that is, for all nonnegative integer numbers i, j the following relation holds:*

$$\boldsymbol{P}(X \geq i + j \,|\, X \geq i) = \boldsymbol{P}(X \geq j).$$

Proof. It is easy to verify that for $k \geq 1$

$$\begin{aligned}\mathbf{P}(X \geq k) &= \sum_{\ell=k}^{\infty} \mathbf{P}(X = \ell) = \sum_{\ell=k}^{\infty} (1-p)^{\ell-1} p \\ &= (1-p)^{k-1} p \sum_{\ell=0}^{\infty} (1-p)^{\ell} = (1-p)^{k-1};\end{aligned}$$

therefore,

$$\begin{aligned}\mathbf{P}(X \geq i + j \,|\, X \geq i) &= \frac{\mathbf{P}(X \geq i + j, X \geq i)}{\mathbf{P}(X \geq i)} \\ &= \frac{\mathbf{P}(X \geq i + j)}{\mathbf{P}(X \geq i)} \\ &= \frac{(1-p)^{i+j-1}}{(1-p)^{i-1}} = (1-p)^j,\ \ j = 0, 1, \ldots .\end{aligned}$$

□

Note that a geometric distribution is sometimes defined on the set $\{0, 1, 2, \ldots\}$ instead of $\{1, 2, \ldots, \}$; in this case, the PDF is determined by

$$p_k = (1-p)^k p,\ k = 0, 1, 2, \ldots .$$

Example. Consider a sequence of experiments and observe whether an event A, $p = \mathbf{P}(A) > 0$, occurs (success) or does not (failure) in each step. If the event

occurs in the kth step first, then define the random variable as $X = k$. In other words, let X be the number of Bernoulli trials of the first success. Then random variable X has a geometric distribution with the parameter p.

Negative Binomial Distribution The distribution of random variable X taking values in $\{0, 1, \ldots\}$ is called a negative binomial distribution with the parameter p, $0 < p < 1$, if

$$p_k = \mathbf{P}(X = k + r) = \binom{r + k - 1}{k} (1 - p)^k p^r, \quad k = 0, 1, \ldots .$$

Expected value and variance: $\mathbf{E}(X) = r\frac{1}{p}, \quad \mathbf{D}^2(X) = r\frac{1-p}{p^2}$;

Generating function: $G(z) = \left(\frac{pz}{1-(1-p)z}\right)^r$;

Characteristic function: $\varphi(t) = p^r \left(1 - (1 - p)\mathrm{e}^{it}\right)^{-r}$.

Example. Let p, $0 < p < 1$, and the positive integer r be two given constants. Suppose that we are given a coin that has a probability p of coming up heads. Toss the coin repeatedly until the rth head appears and define by X the number of tosses. Then random variable X has a negative binomial distribution with parameters (p, r).

Note that from this example it immediately follows that X has a geometric distribution with the parameter p when $r = 1$.

Poisson Distribution The PDF of a random variable X is called a Poisson distribution with the parameter λ ($\lambda > 0$) if X takes values in $\{0, 1, \ldots\}$ and

$$p_k = \mathbf{P}(X = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \quad k = 0, 1, \ldots .$$

Expected value and variance: $\mathbf{E}(X) = \lambda, \quad \mathbf{D}^2(X) = \lambda$;

Generating function: $G(z) = \mathrm{e}^{\lambda(z-1)}$;

Characteristic function: $\varphi(t) = \mathrm{e}^{\lambda(\mathrm{e}^{it}-1)}$.

The following theorem establishes that a binomial distribution can be approximated with a Poisson distribution with the parameter λ when the parameters (p, n) of the binomial distribution satisfy the condition $np \to \lambda$, $n \to \infty$.

Theorem 1.50. *Consider a binomial distribution with the parameter (p, n). Assume that for a fixed constant λ, $\lambda > 0$, the convergence $np \to \lambda$, $n \to \infty$, holds; then the limit of probabilities satisfies the relation*

$$\binom{n}{k} p^k (1 - p)^{n-k} \to \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \quad k = 0, 1, \ldots .$$

Proof. For any fixed $k \geq 0$ integer number we have

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{(np)((n-1)p)\dots((n-k+1)p)}{k!} \mathrm{e}^{(n-k)\log(1-p)}.$$

Since $np \to \lambda$, $n \to \infty$, therefore $p \to 0$, and we obtain

$$\frac{(np)((n-1)p)\dots((n-k+1)p)}{1 \cdot 2 \cdot \dots \cdot k} \to \frac{\lambda^k}{k!}, \quad np \to \lambda.$$

On the other hand, if $p \to 0$, then we get the asymptotic relation $\log(1-p) = -p + o(p)$. Consequently,

$$(n-k)\log(1-p) = -(n-k)(p+o(p)) \to -\lambda, \quad np \to \lambda,\ n \to \infty;$$

therefore, using the last two asymptotic relations, the assertion of the theorem immediately follows. □

1.2.2 Continuous Distributions

Uniform Distribution Let a, b $(a < b)$ be two real numbers. The distribution of random variable X is called uniform on the interval (a, b) if its PDF is given by

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

Expected value and variance: $\mathbf{E}(X) = \frac{a+b}{2}$, $\mathbf{D}^2(X) = \frac{(b-a)^2}{12}$;
Characteristic function: $\varphi(t) = \frac{1}{b-a} \frac{\mathrm{e}^{itb} - \mathrm{e}^{ita}}{it}$.

Note that if X has a uniform distribution on the interval (a, b), then the random variable $Y = \frac{X-a}{b-a}$ is distributed uniformly on the interval $(0, 1)$.

Exponential Distribution $\mathrm{Exp}(\lambda)$, $\lambda > 0$. The distribution of a random variable X is called exponential with the parameter λ, $\lambda > 0$, if its PDF

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0. \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = \frac{1}{\lambda}$, $\mathbf{D}^2(X) = \frac{1}{\lambda^2}$;
Characteristic function: $\varphi(t) = \frac{\lambda}{\lambda - it}$.

The Laplace and Laplace–Stieltjes transforms of the density and distribution function of an $\mathrm{Exp}(\lambda)$ distribution are determined as

$$\mathbf{E}\left(\mathrm{e}^{-sX}\right) = f^*(s) = F^{\sim}(s) = \frac{\lambda}{s+\lambda}.$$

The exponential distribution, similarly to the geometric distribution, has the memoryless property.

Theorem 1.51. *For arbitrary constants $t, s > 0$ the relation*

$$\boldsymbol{P}(X > t+s | X > t) = \boldsymbol{P}(X > s)$$

holds.

Proof. It is clear that

$$\mathbf{P}(X > t+s | X > t) = \frac{\mathbf{P}(X > t+s, X > t)}{\mathbf{P}(X > t)} = \\ = \frac{\mathbf{P}(X > t+s)}{\mathbf{P}(X > t)} = \frac{\mathrm{e}^{-\lambda(t+s)}}{\mathrm{e}^{-\lambda t}} = \mathrm{e}^{-\lambda s}.$$

□

Hyperexponential Distribution Let the PDF of random variable X be a mixture of exponential distributions with the parameters $\lambda_1, \ldots, \lambda_n$ and with weights $a_1, \ldots, a_n$ ($a_k > 0, a_1 + \ldots + a_n = 1$). Then the PDF

$$f(x) = \begin{cases} \sum\limits_{k=1}^{n} a_k \lambda_k \mathrm{e}^{-\lambda_k x} & \text{if } x > 0, \\ 0, & \text{if } x \le 0, \end{cases}$$

of random variable X is called hyperexponential.

Expected value and variance: $\mathbf{E}(X) = \sum\limits_{k=1}^{n} \frac{a_k}{\lambda_k}$, $\mathbf{D}^2(X) = 2\sum\limits_{k=1}^{n} \frac{a_k}{\lambda_k^2} - \left(\sum\limits_{k=1}^{n} \frac{a_k}{\lambda_k}\right)^2$;

Characteristic function: $\varphi(t) = \sum\limits_{k=1}^{n} a_k \frac{\lambda_k}{\lambda_k - it}$.

Denote by $\Gamma(x) = \int\limits_0^{\infty} y^{x-1}\mathrm{e}^{-y}\mathrm{d}y$, $x > -1$ the well-known **gamma function** Γ in analysis, which is necessary for the definition of the gamma distribution.

Gamma Distribution Gamma(α, λ), $\alpha, \lambda > 0$.

The distribution of a random variable X is called a **gamma distribution** with the parameters $\alpha, \lambda > 0$, if its PDF is

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0. \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = \frac{\alpha}{\lambda}, \quad \mathbf{D}^2(X) = \frac{\alpha}{\lambda^2}$;
Characteristic function: $\varphi(t) = \left(\frac{\lambda}{\lambda - it}\right)^\alpha$.

Comment 1.52. *A gamma distribution with the parameters $\alpha = n$, $\lambda = n\mu$ is called an **Erlang distribution**.*

Comment 1.53. *If the independent identically distributed random variables $X_1, X_2, \ldots$ have an exponential distribution with the parameter λ, then the distribution of the sum $Z = X_1 + \ldots + X_n$ is a gamma distribution with the parameter (n, λ). This relation is easy to see because the characteristic function of an exponential distribution with the parameter λ is $(1 - it/\lambda)^{-1}$; then the characteristic function of its nth convolution power is $(1 - it/\lambda)^{-n}$, which equals the characteristic function of a Gamma(n, λ) distribution.*

Beta Distribution Beta(a, b), $a, b > 0$. The distribution of random variable X is called a beta distribution if its PDF is

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

Expected value and variance: $\mathbf{E}(X) = \frac{a}{a+b}, \quad \mathbf{D}^2(X) = \frac{ab}{(a+b)^2(a+b+1)}$;
Characteristic function in the form of power series: $\varphi(t) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha+\beta+k)}$.

Gaussian (Also Called Normal) Distribution $N(\mu, \lambda)$, $-\infty < \mu < \infty$, $0 < \sigma < \infty$. The distribution of random variable X is called Gaussian with the parameters (μ,σ) if it has a PDF

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty.$$

Expected value and variance: $\mu = \mathbf{E}(X)$ and $\sigma^2 = \mathbf{D}^2(X)$;
Characteristic function: $\varphi(t) = \exp\left\{i\mu t - \frac{\sigma^2}{2} t^2\right\}$.

The $N(0, 1)$ distribution is usually called a standard Gaussian or standard normal distribution, and its PDF is equal to

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}, \quad -\infty < x < \infty.$$

It is easy to verify that if a random variable has an $N(\mu, \sigma)$ distribution, then the centered and linearly normed random variable $Y = (X - \mu)/\sigma$ has a standard Gaussian distribution.

Multidimensional Gaussian (Normal) Distribution $N(\mu, \mathbf{R})$ Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ be an n-dimensional random vector whose coordinates $Z_1, \ldots, Z_n$ are independent and have a standard $N(0, 1)$ Gaussian distribution. Let $\mathbf{V} \in \mathbb{R}^{m \times n}$ be an $(m \times n)$ matrix and $\mu = (\mu_1, \ldots, \mu_m)^{\mathrm{T}} \in \mathbb{R}^m$ an m-dimensional vector. Then the distribution of the m-dimensional random vector X defined by the equation $\mathbf{X} = \mathbf{VZ} + \mu$ is called an m-dimensional Gaussian distribution.

Expected value and variance matrix:

$$\mathbf{E}(\mathbf{X}) = \mu_{\mathbf{X}} = \mu \qquad \text{and} \qquad \mathbf{D}^2(\mathbf{X}) = \mathbf{R}_{\mathbf{X}} = \mathbf{E}\left((\mathbf{X} - \mu)(\mathbf{X} - \lambda)^{\mathrm{T}}\right) = \mathbf{V}\mathbf{V}^{\mathrm{T}};$$

Characteristic function:

$$\varphi(\mathbf{t}) = \exp\left\{ i\mathbf{t}^{\mathrm{T}}\mu - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{R}_{\mathbf{X}}\mathbf{t} \right\}, \text{ where } \mathbf{t} = (t_1, \ldots, t_m)^{\mathrm{T}} \in \mathbb{R}^m.$$

If $\mathbf{V}$ is a nonsingular quadratic matrix ($m = n$ and $\det \mathbf{V} \neq 0$), then the random vector $\mathbf{X}$ has a density in the form

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi \det \mathbf{R}_X)^{n/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\mathbf{R}_X^{-1}(\mathbf{x}-\mu) \right\}, \quad \mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n.$$

Example. If the random vector $\mathbf{X} = (X_1, X_2)^{\mathrm{T}}$ has a two-dimensional Gaussian distribution with expected value $\mu = (\mu_1, \mu_2)^{\mathrm{T}}$ and covariance matrix

$$R_X = \begin{bmatrix} a & b \\ b & c \end{bmatrix},$$

then its PDF has the form

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\sqrt{ac - b^2}}{2\pi} \exp\left\{ -\frac{1}{2}[a(x_1 - \mu_1)^2 + 2b(x_1 - \mu_1)(x_2 - \mu_2) + c(x_2 - \mu_2)^2] \right\},$$

where a, b, c, μ_1, μ_2 are constants satisfying the conditions $a > 0$, $c > 0$, and $b^2 < ac$.

Note that the marginal distributions of random variables X_1 and X_2 are $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ Gaussian, respectively, where

$$\sigma_1 = \sqrt{\frac{a}{ac - b^2}}, \quad \sigma_2 = \sqrt{\frac{c}{ac - b^2}} \quad \text{and} \quad b = \operatorname{cov}(X_1, X_2).$$

Distribution Functions Associated with Gaussian Distributions Let $Z, Z_1, Z_2, \ldots$ be independent random variables whose distributions are standard Gaussian, i.e., with the parameters $(0, 1)$. There are many distributions, for example the χ^2 and the logarithmically normal distributions defined subsequently (further examples are the frequently used t, F, and Wishart distributions in statistics [46]), that can be given as distributions of appropriately chosen functions of random variables $Z, Z_1, Z_2, \ldots$.

χ^2 Distribution The distribution of the random variable $X = Z_1^2 + \ldots + Z_n^2$ is called a χ^2 distribution with parameter n. The PDF is

$$f_n(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2a)} x^{n/2-1} \mathrm{e}^{-x/2}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0. \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = n, \quad \mathbf{D}^2(X) = 2n;$
Characteristic function: $\varphi(t) = (1 - 2it)^{-n/2}.$

Logarithmic Gaussian (Normal) Distribution If random variable Z has an $N(\mu, \sigma)$ Gaussian distribution, then the distribution of the random variable $X = \mathrm{e}^Z$ is called a logarithmic Gaussian (normal) distribution. The PDF is

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{\frac{(\log x - \mu)^2}{2\sigma^2}\right\}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0. \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = \mathrm{e}^{\sigma^2/2+\mu}, \quad \mathbf{D}^2(X) = \mathrm{e}^{\sigma^2/2+\mu}\left(\mathrm{e}^{\mathrm{e}^2} - 1\right).$

Weibull Distribution The Weibull distribution is a generalization of the exponential distribution for which the behavior of the tail distribution is modified by a positive constant k as follows:

$$F(x) = \begin{cases} 1 - \mathrm{e}^{-(x/\lambda)^k}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0; \end{cases}$$

$$f(x) = \begin{cases} \left(\frac{k}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{k-1} \mathrm{e}^{-(x/\lambda)^k}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0. \end{cases}$$

Expected value and variance:

$$\mathbf{E}(X) = \lambda\Gamma(1 + 1/k), \quad \mathbf{D}^2(X) = \lambda^2\left(\Gamma(1 + 2/k) - \Gamma^2(1 + 1/k)\right).$$

Pareto Distribution Let c and λ be positive numbers. The density function and the PDF of a Pareto distribution are defined as follows:

$$F(x) = \begin{cases} 1 - (\frac{x}{c})^{-\lambda}, & \text{if } x > c, \\ 0, & \text{if } x \le 0; \end{cases}$$

$$f(x) = \begin{cases} \left(\frac{\lambda}{c}\right)\left(\frac{x}{c}\right)^{-\lambda-1} & \text{if } x > c, \\ 0, & \text{if } x \le c. \end{cases}$$

Since the PDF of the Pareto distribution is a simple power function in consequence of this property, it tends to zero with polynomial order as x goes to infinity and the nth moment exists if and only if $n < \lambda$.
Expected value (if $k > 1$) and variance (if $k > 2$):

$$\mathbf{E}(X) = \frac{ck}{k-1}, \quad \mathbf{D}(X) = \frac{c^2 k}{(k-1)^2(k-2)}.$$

1.3 Limit Theorems

1.3.1 Convergence Notions

There are many convergence notions in the theory of analysis, for example, pointwise convergence, uniform convergence, and convergences defined by various metrics. In the theory of probability, several kinds of convergences are also used that are related to the sequences of random variables or to their sequence of distribution functions. The following notion is the so-called weak convergence of distribution functions.

Definition 1.54. The sequence of distribution functions $F_n,\ n = 1, 2, \ldots$ **weakly converges** to the distribution function F (abbreviated $F_n \xrightarrow{w} F,\ n \to \infty$) if the convergence $F_n(x) \to F(x),\ n \to \infty$, holds in all continuity points of F.

If the distribution function F is continuous, then the convergence $F_n \xrightarrow{w} F,\ n \to \infty$ holds if and only if $F_n(x) \to F(x),\ n \to \infty$ for all $x \in \mathbb{R}$. The weak convergence of the sequence $F_n,\ n = 1, 2, \ldots$ is equivalent to the condition that the convergence

$$\int_{-\infty}^{\infty} g(x)\mathrm{d}F_n(x) \to \int_{-\infty}^{\infty} g(x)\mathrm{d}F(x)$$

is true for all bounded and continuous functions g.

In addition, the weak convergence of a distribution function can be given with the help of an appropriate metric in the space $\mathbb{F} = \{F\}$ of all distribution functions.

Let G and H be two distribution functions (i.e., $G, H \in \mathbb{F}$), and define the **Levy metric** [96] as follows:

$$L(G,H) = \inf\{\varepsilon :\ G(x) \le H(x+\varepsilon) + \varepsilon,\ H(x) \le G(x+\varepsilon) + \varepsilon, \text{ for all } x \in \mathbb{R}\}.$$

Then it can be proved that the weak convergence $F_n \overset{w}{\to} F,\ n \to \infty$, of the distribution functions $F,\ F_n,\ n = 1, 2, \ldots$, holds if and only if $\lim\limits_{n\to\infty} L(F_n, F) = 0$.

The most frequently used convergence notions in probability theory for a sequence of random variables are the convergence in distribution, convergence in probability, convergence with probability 1, or almost surely (a.s.), and convergence in mean square (convergence in L_2), which will be introduced subsequently. In cases of the last three convergences, it is assumed that the random variables are defined on a common probability space $(\Omega, \mathcal{A}, \mathbf{P}())$.

Definition 1.55. The sequence of random variables $X_1, X_2, \ldots$ **converges in distribution** to a random variable X (abbreviated $X_n \overset{d}{\to} X,\ n \to \infty$) if their distribution functions satisfy the weak convergence

$$F_{X_n} \overset{w}{\to} F_X,\ n = 1, 2, \ldots .$$

Definition 1.56. The sequence of random variables $X_1, X_2, \ldots$ **converges in probability** to a random variable X ($X_n \overset{P}{\to} X,\ n \to \infty$) if the convergence

$$\lim_{n\to\infty} \mathbf{P}(|X_n - X| > \varepsilon) = 0$$

holds for all positive constants ε.

Definition 1.57. The random variables $X_1, X_2, \ldots$ **converge with probability** 1 (or **almost surely**) to a random variable X (abbreviated $X_n \overset{\text{a.s.}}{\to} X,\ n \to \infty$) if the condition

$$\mathbf{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$$

holds.

The limit $\lim\limits_{n\to\infty} X_n = X$ exists if there are defined random variables with probability 1 $X' = \limsup\limits_{n\to\infty} X_n$ and $X''(\omega) = \liminf\limits_{n\to\infty} X_n$ for which the relation

$$\mathbf{P}\left(X'(\omega) = X''(\omega) = X(\omega)\right) = 1$$

is true. This means that there is an event $A \in \mathcal{A}$, $\mathbf{P}(A) = 0$, such that the equality

$$X'(\omega) = X''(\omega) = X(\omega),\ \ \omega \in \Omega \setminus A$$

holds.

Theorem 1.58 (*[84]*). *The convergence* $\lim_{n\to\infty} X_n = X$ *with probability 1 is true if and only if for all* $\varepsilon > 0$

$$\boldsymbol{P}\left(\sup_{k\geq n}|X_k - X| > \varepsilon\right) = 0.$$

Definition 1.59. Let $X_n,\ n \geq 1$ and X be random variables with finite variance. The sequence $X_1, X_2, \ldots$ **converges in mean square** to random variable X (abbreviated $X_n \xrightarrow{L_2} X,\ n \to \infty$) if

$$\mathbf{E}\left(|X_n - X|^2\right) \to 0,\ n \to \infty.$$

This type of convergence is often called an L_2 convergence of random variables.

The enumerated convergence notions are not equivalent to each other, but we can mention several connections between them. The convergence in distribution follows from all the others. The convergence in probability follows from the convergence with probability 1 and from the convergence in mean square. It can be proved that if the sequence $X_1, X_2, \ldots$ is convergent in probability to the random variable X, then there exists a subsequence $X_{n_1}, X_{n_2}, \ldots$ such that it converges with probability 1 to random variable X.

1.3.2 Laws of Large Numbers

The intuitive introduction of probability implicitly uses the limit behavior of the average

$$\overline{S}_n = \frac{X_1 + \ldots + X_n}{n},\ n = 1, 2, \ldots,$$

of independent identically distributed random variables $X_1, X_2, \ldots$. The main question is: under what condition does the sequence $\overline{S}_n$ converge to a constant μ in probability (weak law of large numbers) or with probability 1 (strong law of large numbers) as n goes to infinity?

Consider an experiment in which we observe that an event A occurs or not. Repeating the experiment n times independently, define the frequency of event A by $S_n(A)$ and the relative frequency by $\overline{S}_n(A)$.

Theorem 1.60 (*Bernoulli*). *The relative frequency of an event A tends in probability to the probability of the event* $p = \mathbf{P}(A)$, *that is, for all* $\varepsilon > 0$ *the relation*

$$\lim_{n\to\infty} \boldsymbol{P}\left(\left|\overline{S}_n(A) - p\right| > \varepsilon\right) = 0$$

holds.

If we introduce the notation

$$X_i = \begin{cases} 1, \text{ if the } i\text{-th outcome in } A, \\ 0, \text{ otherwise,} \end{cases}$$

then the assertion of the last theorem can be formulated as follows:

$$\overline{S}_n = \frac{X_1 + \ldots + X_n}{n} \xrightarrow{p} p,\ n \to \infty,$$

which is a simple consequence of the Chebyshev inequality because the X_i are independent and identically distributed and $\mathbf{E}(X_i) = p = \mathbf{P}(A)$, $\mathbf{D}^2(X_i) = p(1-p)$, $i = 1, 2, \ldots$. This result can be generalized without any difficulties as follows.

Theorem 1.61. *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with common expected value μ and finite variance σ^2. Then the convergence in probability*

$$\overline{S}_n = \frac{X_1 + \ldots + X_n}{n} \xrightarrow{p} \mu,\ n \to \infty,$$

is true.

Proof. Example 1.38, which is given after the proof of the Chebyshev inequality, shows that for all $\varepsilon > 0$ the inequality

$$\mathbf{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

is valid. From this the convergence in probability $\overline{S}_n \xrightarrow{p} \mu$, $n \to \infty$ follows. It is not difficult to see that the convergence in L_2 is also true, i.e., $\overline{S}_n \xrightarrow{L_2} \mu$, $n \to \infty$. □

It should be noted that the inequality $\mathbf{P}\left(\left|\frac{X_1+\ldots+X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$, which guarantees the convergence in probability, gives an upper bound for the probability $\mathbf{P}\left(\left|\frac{X_1+\ldots+X_n}{n} - \mu\right| \geq \varepsilon\right)$ also.

The Kolmogorov strong law of large numbers gives a necessary and sufficient condition for convergence with probability 1.

Theorem 1.62 (*Kolmogorov*). *If the sequence of random variables $X_1, X_2, \ldots$ is independent and identically distributed, then the convergence*

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{a.s.} \mu,\ n \to \infty$$

holds for a constant μ if and only if the random variables X_i have finite expected value and $\boldsymbol{E}(X_i) = \mu$.

Corollary 1.63. *If $\overline{S}_n(A)$ defines the relative frequency of an event A occurring in n independent experiments, then the Bernoulli law of large numbers*

$$\overline{S}_n(A) \xrightarrow{p} p = \boldsymbol{P}(A),\ n \to \infty,$$

is valid. By the Kolmogorov law of large numbers, this convergence is true with probability 1 *also, that is,*

$$\overline{S}_n(A) \xrightarrow{a.s.} p = \boldsymbol{P}(A),\ n \to \infty.$$

1.3.3 Central Limit Theorem, Lindeberg–Feller Theorem

The basic problem of central limit theorems is as follows. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with a common distribution function $F_X(x)$. The question is, under what conditions does a sequence of constants μ_n and σ_n, $\sigma_n \neq 0$, $n = 1, 2, \ldots$ exist such that the sequence of centered and linearly normed sums

$$\overline{S}_n = \frac{X_1 + \ldots + X_n - \mu_n}{\sigma_n},\ n = 1, 2, \ldots \tag{1.11}$$

converges in the distributions

$$F_{\overline{S}_n} \xrightarrow{w} F,\ n \to \infty$$

and have a nondegenerate limit distribution function F? A distribution function $F(x)$ is nondegenerate if there is no point $x_0 \in \mathbb{R}$ satisfying the condition $F(x_0) - F(x_0-) = 1$, that is, the distribution does not concentrate at one point.

Theorem 1.64. *If the random variables $X_1, X_2, \ldots$ are independent and identically distributed with finite expected value $\mu = \boldsymbol{E}(X_1)$ and variance $\sigma^2 = \mathbf{D}^2(X_1)$, then*

$$\boldsymbol{P}\left(\frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}\sigma} \leq x\right) \to \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-u^2/2} \mathrm{d}u$$

holds for all $x \in \mathbb{R}$, where the function $\Phi(x)$ denotes the distribution function of standard normal random variables.

If the random variables $X_1, X_2, \ldots$ are independent but not necessarily identically distributed, then a general, so-called Lindeberg–Feller theorem is valid.

Theorem 1.65. *Let $X_1, X_2, \ldots$ be independent random variables whose variances are finite. Denote*

$$\mu_n = \boldsymbol{E}(X_1) + \ldots + \boldsymbol{E}(X_n), \quad \sigma_n = \sqrt{\boldsymbol{D}^2(X_1) + \ldots + \boldsymbol{D}^2(X_n)}, \quad n = 1, 2, \ldots .$$

The limit

$$\boldsymbol{P}\left(\frac{X_1 + \ldots + X_n - \mu_n}{\sigma_n} \le x\right) \to \Phi(x), \quad n \to \infty,$$

is true for all $x \in \mathbb{R}$ if and only if the Lindeberg–Feller condition holds:

$$\lim_{n\to\infty} \max_{1\le j\le n} \frac{1}{\sigma_j^2} \boldsymbol{E}\left(X_j^2 \mathcal{I}_{\{|X_j| > \varepsilon\sigma_n\}}\right) = 0, \ x \in \mathbb{R}, \ \varepsilon > 0,$$

where $\mathcal{I}_{\{\}}$ denotes the indicator variable.

1.3.4 Infinitely Divisible Distributions and Convergence to the Poisson Distribution

There are many practical problems for which model (1.11) and results related to it are not satisfactory. The reason is that the class of possible limit distributions is insufficiently large; for instance, it does not consist of discrete distributions. An example of this is a Poisson distribution, which is an often-used distribution in queueing theory.

As a generalization of model (1.11), consider the sequence of series of random variables (sometimes called a sequence of random variables of triangular arrays)

$$\{X_{n,1}, \ldots, X_{n,k_n}\}, \ n = 1, 2, \ldots, \ k_n \to \infty,$$

satisfying the following conditions for all fixed positive integers n:

1. The random variables $X_{n,1}, \ldots, X_{n,k_n}$ are independent.
2. The random variables $X_{n,1}, \ldots, X_{n,k_n}$ are **infinitesimal** (in other words, **asymptotically negligible**) if the limit for all $\varepsilon > 0$

$$\lim_{n\to\infty} \max_{1\le j\le k_n} \mathbf{P}\left(\left|X_{n,j}\right| > \varepsilon\right) = 0$$

holds.

Considering the sums of series of random variables

$$S_n = X_{n,1} + \ldots + X_{n,k_n}, \ n = 1, 2, \ldots,$$

the class of possible limit distributions (so-called infinitely divisible distributions) is already a sufficiently large class containing, for example, a Poisson distribution.

Definition 1.66. A random variable X is called **infinitely divisible** if it can be given in the form

$$X \stackrel{d}{=} X_{n,1} + \ldots + X_{n,n}$$

for every $n = 1, 2, \ldots$, where the random variables $X_{n,1}, \ldots, X_{n,n}$ are independent and identically distributed.

Infinitely divisible distributions (to which, for example, the normal and Poisson distributions belong) can be given with the help of their characteristic functions.

Theorem 1.67. *If random variable X is infinitely divisible, then its characteristic function has the form (**Lévy–Khinchin canonical form**)*

$$\log f(t) = i\mu t - \frac{\sigma^2}{2}t^2 + \int_{-\infty}^{0} \left(\mathrm{e}^{itx} - 1 - \frac{itx}{1+x^2}\right) \mathrm{d}L(x)$$

$$+ \int_{0}^{\infty} \left(\mathrm{e}^{itx} - 1 - \frac{itx}{1+x^2}\right) \mathrm{d}R(x),$$

where the functions L and R satisfy the following conditions:

(a) μ and σ $(\sigma \geq 0)$ are real constants.
(b) $L(x)$, $x \in (-\infty, 0)$ and $R(x)$, $x \in (0, \infty)$ are monotonically increasing functions on the intervals $(-\infty, 0)$ and $(0, \infty)$, respectively.
(c) $L(-\infty) = R(\infty) = 0$ and the inequality condition

$$\int_{-\infty}^{0} x^2 \mathrm{d}L(x) + \int_{0}^{\infty} x^2 \mathrm{d}R(x) < \infty$$

holds.

If an infinitely divisible distribution has finite variation, then its characteristic function can be given in a more simple form (**Kolmogorov formula**):

$$\log f(t) = i\mu t + \int_{-\infty}^{\infty} \left(\mathrm{e}^{itx} - 1 - itx\right) \frac{1}{x^2} \mathrm{d}K(x),$$

where μ is a constant and $K(x)$ $(K(-\infty) = 0)$ is a monotonically nondecreasing function.

As special cases of the Kolmogorov formula, we get the normal and Poisson distributions.

(a) An infinitely divisible distribution is normal with the parameters (μ,σ) if the function $K(x)$ is defined as

$$K(x) = \begin{cases} 0, & \text{if } x \le 0, \\ \sigma^2, & \text{if } x > 0. \end{cases}$$

Then the characteristic function is

$$f(t) = i\mu t - \frac{\sigma^2}{2}t^2.$$

(b) An infinitely divisible distribution is Poisson with the parameter λ ($\lambda > 0$) if $\mu = \lambda$ and the function $K(x)$ is defined as

$$K(x) = \begin{cases} 0, & \text{if } x \le 1, \\ \lambda, & \text{if } x > 1. \end{cases}$$

In this case the characteristic function can be given as follows:

$$f(t) = i\mu t + \int_{-\infty}^{\infty} \left(e^{itx} - 1 - itx\right) \frac{1}{x^2} dK(x) = \lambda(e^{it} - 1).$$

The following theorem gives an answer to the question of the conditions under which the limit distribution of sums of independent infinitesimal random variables is Poisson. This result will be used later when considering sums of independent arrival processes of queues.

Theorem 1.68 (*Gnedenko, Marczinkiewicz*). *Let* $\{X_{1,n}, \ldots, X_{k_n,n}\}$, $n = 1, 2, \ldots$, *be a sequence of series of independent infinitesimal random variables. The sequence of distributions of sums*

$$X_n = X_{n1} + \ldots + X_{n,k_n}, \quad n \ge 1,$$

converges weakly to a Poisson distribution with the parameter λ ($\lambda > 0$) *as* $n \to \infty$ *if and only if the following conditions hold for all* ε ($0 < \varepsilon < 1$)*:*

(A) $$\sum_{j=1}^{k_n} \int_{\mathbb{R}_\varepsilon} dF_{nj}(x) \to 0.$$

(B) $$\sum_{j=1}^{k_n} \int_{|x-1|<\varepsilon} dF_{nj}(x) \to \lambda.$$

(C) $\sum_{j=1}^{k_n} \int\limits_{|x|<\varepsilon} \mathrm{d}F_{nj}(x) \to 0.$

(D) $\sum_{j=1}^{k_n} \left[\int\limits_{|x|<\varepsilon} x^2 \mathrm{d}F_{nj}(x) - \left(\int\limits_{|x|<\varepsilon} x \mathrm{d}F_{nj}(x) \right)^2 \right] \to 0,$

where $F_{nj}(x) = \boldsymbol{P}\left(X_{nj} \le x\right)$ *and* $\mathbb{R}_\varepsilon = \mathbb{R} \setminus (\{|x| < \varepsilon\} \cup \{|x-1| < \varepsilon\})$.

Note that conditions (A) and (B) guarantee the convergence of the Poisson part to the appropriate Poisson distribution of the limit, (C) means that there is no centralization, and from (D) it follows that the limit distribution does not contain a Gaussian part.

1.4 Exercises

Exercise 1.1. Let X be a nonnegative random variable with CDF F_X. Given $0 \le t \le X$ $[\mathbf{P}(X > t) \ne 0]$, find the CDF of residual lifetime X.

Exercise 1.2. Let X and Y be independent random variables with a Poisson distribution of parameters λ and μ, respectively. Verify that

(a) The sum $X + Y$ has a Poisson distribution with the parameter $\lambda + \mu$;
(b) For any nonnegative integers $m \le n$ the conditional distribution $\mathbf{P}(X = m \mid X + Y = n)$ is binomial with the parameter $(n, \frac{\lambda}{\lambda+\mu})$, i.e.,

$$\mathbf{P}(X = m \mid X + Y = n) = \binom{m}{n} \left(\frac{\lambda}{\lambda+\mu}\right)^m \left(1 - \frac{\lambda}{\lambda+\mu}\right)^{n-m}.$$

Exercise 1.3. Let X and Y be independent random variables having a uniform distribution on the interval $(0, 1)$ and an exponential distribution with the parameter 1, respectively. Find the probability (concrete number) that $X < Y$.

Exercise 1.4. Divide the interval $(0, 1)$ into three parts with two independently and randomly chosen points U_1 and U_2 of the interval $(0, 1)$. Find the probability of event A that the three parts can determine a triangle.

Exercise 1.5. Show that for a nonnegative random variable X with a finite nth ($n \ge 1$) moment it is true that $\mathbf{E}(X^n) = \int\limits_0^\infty \mathbf{P}(x < X) n x^{n-1} \mathrm{d}x$.

Exercise 1.6. Let X and Y be independent random variables with a uniform distribution on the interval $(0, 1)$. Find the quantities

(a) $\mathbf{E}(|X - Y|)$, $\mathbf{D}^2(|X - Y|)$,
(b) $\mathbf{P}(|X - Y|) > \frac{1}{2}$.

Exercise 1.7. Let X and Y be independent random variables having an exponential distribution with the parameters λ and μ, respectively.

(a) Determine the density function of the random variable $Z = X + Y$.
(b) Find the density function of the random variable $W = \min(X, Y)$.

Exercise 1.8. Let $X_1, \ldots, X_n$ be independent random variables having an exponential distribution with the parameter λ.
Find the expected values of the random variables $V_n = \max(X_1, \ldots, X_n)$, and $W_n = \min(X_1, \ldots, X_n)$.

Exercise 1.9. Let X and Y be independent random variables with density functions $f_X(x)$ and $f_Y(x)$, respectively. Determine the conditional expected value $E(X \mid X < Y)$.

Exercise 1.10. Determine the conditional expectations $\mathbf{E}(X \,|Y = y)$ and $\mathbf{E}(X \,|Y)$ if the joint PDF of the random variables X and Y has the form

(a) $f_{X,Y}(x, y) = \begin{cases} 2, \text{ if } 0 < x, y \text{ and } x + y < 1, \\ 0, \text{ otherwise;} \end{cases}$

(b) $f_{X,Y}(x, y) = \begin{cases} 3(x + y), \text{ if } 0 < x, y \text{ and } x + y < 1, \\ 0, \qquad\quad \text{otherwise.} \end{cases}$

Exercise 1.11. Let $X_1, X_2, \ldots$ be independent random variables with an exponential distribution of the parameter λ. Let N be a geometrically distributed random variable with the parameter p [$p_k = \mathbf{P}(N = k) = p(1 - p)^k$, $k = 1, 2, \ldots$], which does not depend on random variables $(X_1, X_2, \ldots)$. Prove that the sum $Y = X_1 + \ldots + X_N$ has an exponential distribution with the parameter $p\lambda$.

Exercise 1.12. Consider the distribution function of the sum Y_{40} of independent random variables $X_1, \ldots, X_{40}$ having an exponential distribution with the parameter 1. Give an estimate for the probability $p = \mathbf{P}\left(\frac{|Y_{40} - \mathbf{E}(Y_{40})|}{\mathbf{D}(Y_{40})} > 0.05\right)$ calculated with the help of the central limit theorem. We can numerically calculate this probability because the random variable Y_{40} has a gamma distribution with the parameter $(40, 1)$. Using this fact, what result can we obtain for the considered probability? (On the numerical calculation of the gamma distribution see, for example, [72] or [63].)