UNIVERSITY OF TARTU
INSTITUTE OF MATHEMATICS AND STATISTICS

Stochastic Models

Lecture notes

Fall 2020

Kalev Pärna

J. Liivi 2, 50409 Tartu,
Email: kalev.parna@ut.ee

These lecture notes present some concepts and results in probability and random processes (and also some selected topics of statistics) needed in forthcoming courses of the master program in Actuarial and Financial Mathematics. We assume that students have taken a basic course in probability before. The course starts with general (axiomatic) definition of probability, followed by random variables (as measurable functions) and their expected values (as Lebesgue integrals). Then the concept of conditional expectation is explained. The characteristic functions and two important limit theorems, the Law of Large Numbers and the Central Limit Theorem, are presented. Among random processes we will briefly cover Markov chains, Poisson processes and Brownian motion - the models that can be used to study various random phenomena in insurance and financial markets. Finally, multiple linear regression is covered using matrix notation.

There are many textbooks of different levels in probability theory. As introductory courses the following books can be recommended, for example:
G.R. Grimmett, D. Welsh. *Probability. An Introduction.* Oxford Science Publications, 1998.
S.M. Ross. *Introduction to Probability Models.* Academic Press, 1985. Chapters 4, 5, 10.

More detailed an more extensive presentation is given in:
G.R. Grimmett, D.R. Stirzaker. *Probability and Random Processes.* Oxford Science Publications, 1993.

Books for systematic study of measure-theoretic probability theory:
J. Jacod, P. Protter. *Probability Essentials.*Springer, 2004.
J.S. Rosenthal. *A First Look at Rigorous Probability Theory.* World Scientific, 2006.

English - Estonian dictionary

stochastic (random) - juhuslik
probability space $(\Omega, \mathcal{F}, \mathbf{P})$ - tõenäosusruum
conditional probability - tinglik tõenäosus
random variable - juhuslik suurus
expectation (mean value) - keskväärtus
variance - dispersioon
conditional expectation - tinglik keskväärtus
measurable - mõõtuv
Brownian motion - Browni liikumine (= Wieneri protsess)
random walk - juhuslik ekslemine
central limit theorem - tsentraalne piirteoreem
trajectory (path) - trajektoor

Basic concepts of probability

1 Events and probabilities

Random experiment (trial) is an action whose consequence is not predetermined.

Space of elementary events (sample space) Ω is the set of all possible outcomes of the random trial. Each element ω in Ω is called an elementary event (sample point).

σ-algebra of events (event space) $\mathcal{F}$ is a collection of subsets of Ω satisfying:

1. $\emptyset, \Omega \in \mathcal{F}$
2. if $A_1, A_2, \ldots \in \mathcal{F}$, then also $\cup_i A_i \in \mathcal{F}$
3. if $A \in \mathcal{F}$, then also $\bar{A} \in \mathcal{F}$.

Remark: From 1.-3. it follows that $\mathcal{F}$ is also closed w.r.t. intersections of its elements: if $A_i \in \mathcal{F}$ then also $\cap_i A_i \in \mathcal{F}$.

All elements of the event space $\mathcal{F}$ are called **events**.

Example 1. Let us throw a dice. Then $\Omega = \{1, 2, \ldots, 6\}$ and we can define $\mathcal{F}$ as a collection of **all** possible subsets of Ω, i.e. $\mathcal{F} = 2^{\Omega}$, which contains $2^6 = 64$ elements.

Example 2. In the previous example, suppose you are betting on the result of a dice throw: you win or lose 1 euro depending on whether the outcome is an odd number $(1, 3, 5)$ or even number $(2, 4, 6)$. Then you can use a simpler event space, namely $\mathcal{F} = \{\emptyset, \Omega, \{1, 3, 5\}, \{2, 4, 6\}\}$.

In the latter example, we say that event space $\mathcal{F}$ is generated by the event $A = \{1, 3, 5\}$, and we write $\mathcal{F} = \sigma(A)$. This is the smallest σ- algebra which contains the subset A. σ-algebras can also be generated in a more sophisticated way (see below σ-algebras induced by random variables).

Example 3. (Borel sets and Borel σ-algebra).
Consider a special case where Ω is the real line $\mathcal{R}$ (imagine that we are throwing a random point on the real line, or we measure yearly profit or loss of an enterprise). Consider all possible intervals $(a, b],\ a < b$. The collection of all such intervals is not a σ-algebra by itself (why?), and we have to add other necessary subsets in order to get the requirements 1-3 fulfilled. For example, we have to add all countable unions of intervals, their complements etc. We see that also closed intervals $[a, b]$ and open intervals (a, b) must be included, together with their unions etc. The smallest σ-algebra which contains all the intervals above is called **Borel σ-algebra**, and it is denoted by $\mathcal{B}$. Each element B of $\mathcal{B}$ is called a **Borel set**. We can say that the Borel σ- algebra $\mathcal{B}$ is generated by the class K of all intervals, and we write $\mathcal{B} = \sigma(K)$.

Now, after we have defined events, each of them will get its probability - a number showing the possibility of its occurrence.

Probability (probability measure) $\mathbf{P}$ is a function on $\mathcal{F}$ satisfying:

1. $\mathbf{P}(A) \geq 0$ for each $A \in \mathcal{F}$
2. $\mathbf{P}(\emptyset) = 0,\ \mathbf{P}(\Omega) = 1$
3. if $A_1, A_2, \ldots$ do not intersect and each $A_i \in \mathcal{F}$, then $\mathbf{P}(\cup_i A_i) = \sum_i \mathbf{P}(A_i)$.

The last property is called σ-additivity, or countable additivity.
The triple $(\Omega, \mathcal{F}, \mathbf{P})$ is called a **probability space**.

Let us consider some typical examples of probability spaces.

Example 4. (Classical probability) Consider a simple special case of probability space where Ω is a *finite* set, $\Omega = \{\omega_1, \omega_2, \dots \omega_N\}$, and all N possible outcomes are *equiprobable*. Then put $\mathcal{F} = 2^\Omega$ - the set of all subsets of Ω, and define $\mathbf{P}(A) = N_A/N$, where N_A is the number of elements in A.

Exercise. Two dice are tossed. What is the probability that the sum is 10 or more.

Example 5. (Discrete probability space). Let Ω be a *countable* set, $\Omega = \{\omega_1, \omega_2, \dots\}$. Then, again, one can take $\mathcal{F} = 2^\Omega$, and define $\mathbf{P}$ by indicating the probability of each elementary event $\omega_i \in \Omega$, i.e. $p_i = \mathbf{P}(\omega_i)$, $i = 1, 2, \dots$, satisfying $p_i \geq 0$, $\sum_i p_i = 1$.

For example, an insurance company is interested in statistics of car accidents among its clients. The number of car accidents per day can take a value from the set $\Omega = \{0, 1, 2, \dots\}$. A reasonable way to define probability $\mathbf{P}$ in this case is to put

$$\mathbf{P}(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \dots$$

where λ is the average number of car accidents per day. This is a typical example of using *Poisson distribution* to model real data. (Note that some other distributions can also be used for this purpose.)

Example 6. (Uncountable probability space). Let us throw a random point into the interval $[0, 1]$ (for example, computer generates a random number). Then $\Omega = [0, 1]$ and $\mathcal{F}$ can be taken as the collection all Borel sets within this interval, i.e. $\mathcal{F} = \mathcal{B}([0, 1])$. Define the probability $\mathbf{P}$ of a simple interval $(a, b] \in \mathcal{B}([0, 1])$ as its length, i.e. $P(a, b] = b - a$. Further on, if B is a finite union of intervals $(a_i, b_i]$, then use its total length

$$P(B) = \sum_{i=1}^{n} (b_i - a_i).$$

Finally, for an arbitrary event $B \in \mathcal{B}([0, 1])$, the probability will again be defined by its 'length', but now it needs more general formula of called *Lebesgue measure*

$\mu(B)$:

$$\mathbf{P}(B) = \mu(B) = \inf_{B \in \cup_i (a_i, b_i]} \sum (b_i - a_i).$$

The infimum is taken over all possible finite and countable covers of B. Such a probability measure $\mathbf{P}$ is called *uniform* distribution on $[0, 1]$.

In the last example we used, in fact, an important principle that the probability can be defined starting from a rather simple subclass of events (intervals $(a, b]$ in our case). It is the idea behind the following extension theorem.

Properties of probability

All the following properties follow from the definition.

1. If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.
2. $P(A) \leq 1$ for any $A \in \mathcal{F}$.

   Proof. Use $P(A) + P(\bar{A}) = P(\Omega) = 1$.
3. $P(\bar{A}) = 1 - P(A)$.
4. Monotonicity: if $A \subset B$, then $P(A) \leq P(B)$.
5. Difference: $P(A \setminus B) = P(A) - P(A \cap B)$.

   Special case: if $B \subset A$, then $P(A \setminus B) = P(A) - P(B)$.
6. For arbitrary A and B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
7. Boole's inequality (subadditivity):

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

Theorem 1 (Continuity of probability)**.**

(i) If $A_1 \subset A_2 \subset ...$ *and* $A = \bigcup_{i=1}^{\infty} A_i$*, then* $\lim_{n\to\infty} P(A_n) = P(A)$.

(ii) If $B_1 \supset B_2 \supset ...$ *and* $B = \bigcap_{i=1}^{\infty} B_i$*, then* $\lim_{n\to\infty} P(B_n) = P(B)$.

The first is called the **continuity from below**, the second - **continuity from above.**

Proof. To prove (i) denote $C_1 = A_1,\ C_i = A_i \setminus A_{i-1},\ i = 2, 3, \ldots$ and notice that these events do not intersect. Since

$$A_n = C_1 \cup C_2 \cup \ldots \cup C_n, \quad A = C_1 \cup C_2 \cup \ldots$$

we have

$$P(A_n) = \sum_{i=1}^{n} P(C_i), \quad P(A) = \sum_{i=1}^{\infty} P(C_i).$$

However, the series on the RHS converges, hence the sequence of its partial sums converges to $P(A)$, or $\lim_{n\to\infty} P(A_n) = P(A)$. This proves (i). ◀

Extension theorem It turns out that in order to define the probability $\mathbf{P}$ on an σ-algebra $\mathcal{F}$, it is not necessary to show the probability $\mathbf{P}(A)$ for all $A \in \mathcal{F}$ – it is enough to show the probability only for a simpler subclass of events. One of such simpler subclasses is called 'algebra'. Algebra differs from σ-algebra only by one aspect: it is closed w.r.t *finite* unions of its elements. **Algebra** $\mathcal{A}$ is a collection of subsets of Ω satisfying:

1. $\emptyset, \Omega \in \mathcal{A}$
2. if $A_1, \ldots A_n \in \mathcal{A}$, then also $\cup_{i=1}^{n} A_i \in \mathcal{A}$
3. if $A \in \mathcal{A}$, then also $\bar{A} \in \mathcal{A}$.

It is clear that each σ-algebra is an algebra. However, in general, the algebra is not σ-algebra. In order to turn the algebra $\mathcal{A}$ into a σ-algebra, it is necessary to add more subsets from Ω until the property 2 of σ-algebra gets fulfilled. Let $\sigma(\mathcal{A})$ be the a smallest σ-algebra containing algebra $\mathcal{A}$ (we say that $\sigma(\mathcal{A})$ is generated by the algebra $\mathcal{A}$). Next theorem says that is enough to know probabilities of elements of $\mathcal{A}$ only, in order to define the probability on $\sigma(\mathcal{A})$.

Theorem. (Carathéodory's Extension Theorem). Let Ω be a set and let $\mathcal{A}$ be an algebra on Ω, and let $\mathcal{F} = \sigma(\mathcal{A})$. If $\mathbf{P}_0$ is countably additive probability on $\mathcal{A}$,

then there exists a unique probability measure $\mathbf{P}$ on $\mathcal{F}$ such that

$$\mathbf{P}(A) = \mathbf{P}_0(A) \text{ for each } A \in \mathcal{A}.$$

This theorem (which we do not prove here) will be used later when considering distribution functions of random variables.

Corollary. In order to define a probability on the real line $\mathcal{R}$, it is enough to show probabilities of the intervals of type $(a, b]$.

Proof. Suppose we know probabilities $P(a, b]$ of all simple intervals $(a, b]$, $a < b$. Then, considering algebra $\mathcal{A}$ consisting of finite unions of non-intersecting intervals $A = \cup_{i=1}^{n}(a_i, b_i]$, we put $P(A) = \sum_{i=1}^{n} P((a_i, b_i])$. It can be shown that such P is countably additive on $\mathcal{A}$. Now the Extension Theorem applies, giving us a (unique) probability measure on Borel σ- algebra $\mathcal{B} = \sigma(\mathcal{A})$.

Exercises on probability space.

Ex.1. Let A and B be two non-overlapping subsets of a sample space Ω. Build the smallest σ-algebra containing these two sets (denote it by $\sigma(A, B)$).

Ex.2. (Good properties of inverse images). Let Ω and S be any two sets and let $X : \Omega \to S$ be a function on Ω. Consider inverse mapping $X^{-1}(B) = \{\omega : X(\omega) \in B\}$ where $B \subset S$. Show that X^{-1} has following properties:

- $X^{-1}(B_1 \cup B_2) = X^{-1}(B_1) \cup X^{-1}(B_2)$
- $X^{-1}(B_1 \cap B_2) = X^{-1}(B_1) \cap X^{-1}(B_2)$
- $X^{-1}(\overline{B}) = \overline{X^{-1}(B)}$

  (recall that $\overline{B}$ means the complement, $\overline{B} = \Omega \setminus B$.).

Ex.3. Let $\mathcal{F}$ and $\mathcal{G}$ be two σ-algebras defined on a sample space Ω. Show that then the intersection of these σ-algebras, $\mathcal{H} = \mathcal{F} \cap \mathcal{G}$ is also a σ-algebra.

HW 1 Ex.4. Let $\Omega = \{1, 2, 3, 4\}$. Determine whether or not each of the following is a σ-algebra.
(a) $\mathcal{F}_1 = \{\emptyset,\ \{1, 2\},\ \{3, 4\},\ \{1, 2, 3, 4\}\}$.
(b) $\mathcal{F}_2 = \{\emptyset,\ \{3\},\ \{4\},\ \{1, 2\},\ \{3, 4\},\ \{1, 2, 3\},\ \{1, 2, 4\},\ \{1, 2, 3, 4\}\}$.
(c) $\mathcal{F}_3 = \{\emptyset,\ \{1, 2\},\ \{1, 3\},\ \{1, 4\},\ \{2, 3\},\ \{2, 4\},\ \{3, 4\},\ \{1, 2, 3, 4\}\}$.

HW 2 Ex.5. Consider the intervals of the form $(a, b]$, $a \leq b$ (we allow for $a = -\infty$ and $b = +\infty$ and it is assumed that the real line $\mathcal{R} = (-\infty, +\infty]$). Note that $(a, a] = \emptyset$. Let $\mathcal{A}$ be the class of all finite unions of such intervals, i.e. each $A \in \mathcal{A}$ can be written as $A = \cup_{i=1}^{n}(a_i, b_i]$ where $a_i \leq b_i$ and n is a finite number. Show that $\mathcal{A}$ is an algebra on $\mathcal{R}$. Is it also a σ-algebra?

2 Random variables

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a given probability space. Let X be a function $X : \Omega \to \mathcal{R}$. For a Borel set B we can consider its *inverse image* $X^{-1}(B) = \{\omega : X(\omega) \in B\}$. Depending on the function X such an inverse set can be an event (i.e. $X^{-1}(B) \in \mathcal{F}$), or not. The function X is called **measurable** (w.r.t. $\mathcal{F}$) if for *each* Borel set $B \in \mathcal{B}$ its inverse set $X^{-1}(B) \in \mathcal{F}$. Measurable functions are also called **random variables**.

Example. The function X defined by

$$X(\omega) = \begin{cases} +1, \text{ if } \omega = 1, 3, 5 \\ -1, \text{ if } \omega = 2, 4, 6 \end{cases}$$

is measurable w.r.t. the σ-algebra $\mathcal{F}_1 = \{\emptyset, \Omega, \{1, 3, 5\}, \{2, 4, 6\}\}$ but it is not measurable w.r.t. $\mathcal{F}_2 = \{\emptyset, \Omega, \{1, 2, 3\}, \{4, 5, 6\}\}$. (Why?)
Remark: We see that X is measurable w.r.t. $\mathcal{F}$ if X changes its values only on the borders of subsets of $\mathcal{F}$.

Exercises on random variables

HW 3 Ex.1. Let $\Omega = [0, 1]$ and let us have a σ-algebra $\mathcal{F} = \{\emptyset,\ [0, \frac{1}{3}],\ (\frac{1}{3}, 1],\ [0, 1]\}$.

The function X is defined as

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \in [0, \frac{1}{2}] \\ 2, & \text{if } \omega \in (\frac{1}{2}, 1] \end{cases}$$

(a) Is measurable w.r.t $\mathcal{F}$? If not, then modify X to make it measurable.
(b) Can a random variable Y have more than two different values and still be measurable w.r.t $\mathcal{F}$ above? (Reason!)

σ-algebra generated by a random variable

Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Each random variable X generates its σ-algebra $\sigma(X)$ defined as the collection of all subsets of the form $X^{-1}(B)$ where B is an arbitrary Borel subset:

$$\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}\}.$$

Note that always $\sigma(X) \subset \mathcal{F}$ (sub-σ–algebra).

Example: $\Omega = \{$all students in the classroom$\}$.
Let $X(\omega) = 0$, if $\omega =$male student, and $X(\omega) = 1$, if $\omega =$female student.
Then $\sigma(X)$ consists of 4 subsets. (Which ones?)

Each random variable X induces in a natural way a probability measure on $\mathcal{R}$. Indeed, since for all Borel sets $B \in \mathcal{B}$ we have $X^{-1}(B) \in \mathcal{F}$, one can speak about the probabilities $\mathbf{P}(X^{-1}(B))$. Using these probabilities, define a new probability $\mathbf{P}_X$ on the real line by

$$\mathbf{P}_X(B) = \mathbf{P}(X^{-1}(B)), \quad B \in \mathcal{B}.$$

The measure $\mathbf{P}_X$ is called the *distribution of* X. The distribution $\mathbf{P}_X$ gives full information about the probabilistic behavior of X, however, it is somewhat difficult to work with set-functions (as $\mathbf{P}_X$ is). It is easier to work with functions having numeric argument. Next we introduce such a function.

HOMEWORK NB! To solve next problems, you can use only the three properties given in the definition of probability P. (It is assumed that the events $A, B \in \mathcal{F}$.)

HW 4. Show that, in general, $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.

HW 5. Show that if $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.

HW 6. Suppose $\mathbf{P}(A) = 4/5$ and $\mathbf{P}(B) = 1/3$. Show that always $1/15 \leq \mathbf{P}(A \cap B) \leq 1/3$.

HW 7. (Subadditivity) Let $A_i \in \mathcal{F}$ be a sequence of events. Show that

$$\mathbf{P}(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} \mathbf{P}(A_i).$$

Distribution functions

Let X be a random variable and let t be any real number, $t \in \mathcal{R}$. We are interested in the probability that X takes a value which is less than t.

Definition. *Distribution function* of a random variable X is defined as

$$F(t) = \mathbf{P}\{X \leq t\} \equiv \mathbf{P}\{\omega : X(\omega) \leq t\}, \quad -\infty < t < \infty.$$

Remark: The set $\{\omega : X(\omega) \leq t\}$, written shortly as $\{X \leq t\}$, can be presented in the form $X^{-1}(B)$ where the Borel set $B = (-\infty, t]$. However, since X is measurable, all sets of the form $X^{-1}(B) \in \mathcal{F}$, and thus the probability $\mathbf{P}\{X \leq t\}$ is well defined.

The main properties of a distribution function are:

1. $0 \leq F(t) \leq 1$,
2. $F(t)$ is a monotonic function,
3. $\lim_{t \to -\infty} F(t) = 0, \quad \lim_{t \to +\infty} F(t) = 1$.

It is important to note that the distribution function F of the random variable X uniquely defines its distribution $\mathbf{P}_X$. Indeed, the probabilities of simple intervals $(a, b]$ can be expressed using the distribution function F:

$$\mathbf{P}_X(a, b] = P\{a < X \leq b\} = F(b) - F(a).$$

It remains to apply the Corollary (following Caratheodory's Extension Theorem) to see that the measure $\mathbf{P}_X$ can uniquely be extended to the whole Borel σ-algebra $\mathcal{B}$.

The probability that the r.v. X takes a value $a \in \mathcal{R}$ can be expressed via its d.f. F in the following way:

$$\mathbf{P}(X = a) = F(a) - F(a-),$$

where the left limit is defined as $F(a-) = lim_{x \to a-} F(x)$.

Two important subclasses of random variables

The random variable X (and its distribution F) is called **discrete** if X takes on only finite or countable number of different values $x_1, x_2, \ldots$. The distribution function of a discrete r.v. X is a jump function with jumps equal to the probabilities $p_i = \mathbf{P}(X = x_i)$.

Examples of discrete distributions : Bernoulli (0-1), binomial, Poisson, geometric,...

The random variable X is called **continuous** if its distribution function can be expressed in the form of integral

$$F(t) = \int_{-\infty}^{t} f(x)dx.$$

The function $f(x)$ is called density function (or simply density) of X and it can be calculated as $f(x) = F'(x)$.

Examples of continuous distributions: uniform, exponential, normal, lognormal, Pareto, gamma,..., and many others.

NB! There are random variables that are neither discrete nor continuous. For example, mixtures of discrete and continuous distributions. Also, theoretically there are so called singular distributions.

Definition: A distribution is called *singular*, if its distribution function $F(t)$ is continuous but the set of its growth points has Lebesgue measure 0. (A point t is called growth point of $F(t)$ if for each $\epsilon > 0$ $F(t-\epsilon) - F(t+\epsilon) > 0$.)

An ***example*** of a singular distribution is Cantor curve.

The *Theorem of Lebesgue* says that each distribution function $F(t)$ can be represented as a sum of three components: discrete, continuous, and singular.

3 Expectation (expected value)

The reader probably knows the formulas of expected values for discrete and continuous r.v. The expectation of a discrete RV is calculated as

$$\mathbf{E}X = \sum_i x_i p_i.$$

Expected value of a continuous random variable with density function $f(x)$ is calculated as

$$\mathbf{E}X = \int_{-\infty}^{\infty} x f(x) dx.$$

These two formulas are special cases of the **general formula of expectation**:

$$\mathbf{E}X = \int_\Omega X(\omega)\mathbf{P}(dw) = \int_\Omega X d\mathbf{P}$$

which is called *Lebesgue integral* (here X is an arbitrary RV).
Alternative notation for the same is

$$\mathbf{E}X = \int_{-\infty}^{\infty} x dF(x)$$

- so called *Stiltjes integral* (here $F(x)$ is distribution function of X).

We now explain how this general expression for the expected value is defined. There are three steps to define expected value for a general random variable X :
1) simple random variable

2) a general non-negative random variable

3) a general random variable X.

Consider these three steps in details.

1. Random variable X is called simple, if it takes on a finite number values $x_1, x_2, \ldots, x_n$ whose probabilities are $p_i = \mathbf{P}(X = x_i)$. Then the expected value (expectation) of X is defined as

$$\mathbf{E}X = \sum_{i=1}^{n} x_i p_i.$$

Note that for simple r.v. $\mathbf{E}X < \infty$.

2. Let X be arbitrary non-negative random variable. Then it can be expressed as a limit of an increasing sequence of non-negative simple random variables X_n defined by

$$X_n(\omega) = \begin{cases} \frac{k-1}{2^n}, \text{ if } \frac{k-1}{2^n} \le X(\omega) < \frac{k}{2^n}, \ k = 1, \ldots, n2^n \\ n, \text{ if } X(\omega) \ge n. \end{cases}$$

Since $\mathbf{E}X_n$ is also an increasing sequence, it has some limit $\lim_{n\to\infty} \mathbf{E}X_n$ (possibly equal to $+\infty$. We now define $\mathbf{E}X = \lim_{n\to\infty} \mathbf{E}X_n$.

3. Let X be arbitrary r.v. Then X can be represented as $X = X^+ - X^-$, where X^+ and X^- are positive and negative parts of X:

$$X^+ = \begin{cases} X, \text{ if } X \ge 0 \\ 0, \text{ if } X < 0, \end{cases}$$

and

$$X^- = \begin{cases} 0, \text{ if } X \ge 0 \\ -X, \text{ if } X < 0. \end{cases}$$

Define now $\mathbf{E}X = \mathbf{E}X^+ - \mathbf{E}X^-$. Since according to step 2, both $\mathbf{E}X^+$ and $\mathbf{E}X^-$ can also take values $+\infty$, we have four cases:

1) If both $\mathbf{E}X+ < \infty$ and $\mathbf{E}X^- < \infty$, then $\mathbf{E}X < \infty$ and we say that X has *finite expectation* and call X *integrable.* Note that this is equivalent to $\mathbf{E}|X| = \mathbf{E}X^+ + \mathbf{E}X^- < \infty$.

2) If $\mathbf{E}X^+ = \infty$ and $\mathbf{E}X^- < \infty$, then $\mathbf{E}X = \infty$.
3) If $\mathbf{E}X^+ < \infty$ and $\mathbf{E}X^- = \infty$, then $\mathbf{E}X = -\infty$.
4) If $\mathbf{E}X^+ = \infty$ and $\mathbf{E}X^- = \infty$, then $\mathbf{E}X$ can not be defined (as we get $\infty - \infty$).

The expected value $\mathbf{E}X$ is often written as $\int_\Omega X(\omega)\mathbf{P}(d\omega)$ or, shortly, $\int_\Omega X d\mathbf{P}$. We call this the **Lebesgue integral** of X with respect to the probability measure $\mathbf{P}$.

Some important **properties** of $\mathbf{E}X$:

1) linearity: $\mathbf{E}(aX + bY) = a\mathbf{E}X + b\mathbf{E}Y$
2) monotonicity: if $X \leq Y$ then $\mathbf{E}X \leq \mathbf{E}Y$

3) $|\mathbf{E}X| \leq \mathbf{E}|X|$ (follows from triangle inequality $|a - b| \leq |a| + |b|$

4) Monotone convergence theorem (MON): If $X_1, X_2, \ldots$ is an increasing sequence of random variables, $X_n \uparrow X$ with $\mathbf{E}X_1 > -\infty$, then $\lim_{n\to\infty} \mathbf{E}X_n = \mathbf{E}X$.

5) Dominated convergence theorem (DOM): if $X_1, X_2, \ldots$ is a sequence of random variables and $X_n \xrightarrow{a.s.} X$, and $\forall n \ X_n \leq Y$ where $\mathbf{E}|Y| < \infty$, then $\lim_{n\to\infty} \mathbf{E}X_n = \mathbf{E}X$.

Proof of 2 (monotonicity): It is easy to see that if $X \geq 0$, then $\mathbf{E}X \geq 0$.
Now, if $X \leq Y$ then $(Y - X) \geq 0$. Hence $\mathbf{E}(Y - X) \geq 0$. Using linearity property $0 \leq \mathbf{E}(Y - X) = \mathbf{E}Y - \mathbf{E}X$ which leads to $\mathbf{E}X \leq \mathbf{E}Y$.

Proof of 3: We know $X = X^+ - X^-$, where both X^+ and X^- are non-negative random variables. Hence, $|X| = X^+ + X^-$, and for linearity property we have $\mathbf{E}|X| = \mathbf{E}X^+ + \mathbf{E}X^-$. Using the triangle inequality $|a - b| \leq |a| + |b|$ we can estimate

$$|\mathbf{E}X| = |\mathbf{E}X^+ - \mathbf{E}X^-| \leq |\mathbf{E}X^+| + |\mathbf{E}X^-|.$$

Now, since $\mathbf{E}X^+ \geq 0$ and $\mathbf{E}X^- \geq 0$, we have

$$|\mathbf{E}X| \leq \mathbf{E}X^+ + \mathbf{E}X^- = \mathbf{E}|X|.$$

We now show an alternative formula for the expected value of non-negative random variables.

Lemma: If $X \geq 0$ and $\mathbf{E}X < \infty$, then

$$\mathbf{E}X = \int_0^\infty [1 - F(x)]dx.$$

Proof. We know, by definition, $\mathbf{E}X = \int_0^\infty xdF(x)$

$$= -\int_0^\infty xd[1 - F(x)]$$

(using integration by parts)

$$= -[x(1 - F(x))]_0^\infty + \int_0^\infty [1 - F(x)]dx.$$

Now, $[x(1 - F(x))]_0^\infty = \lim_{x\to\infty}[x(1 - F(x))] - 0 \cdot [1 - F(0)]$. Estimate the first expression

$$x[1 - F(x)] = x\int_x^\infty dF(y) \leq \int_x^\infty ydF(y)$$

which tends to zero as $x \to \infty$ (for the finitness of the integral $\int_0^\infty ydF(y) < \infty$). Therefore,

$$0 \leq x[1 - F(x)] \to 0,$$

and we have

$$\mathbf{E}X = \int_0^\infty [1 - F(x)]dx.$$

□

Special case of lemma: Let X takes only non-negative integer values 0,1,2,... Then

$$\mathbf{E}X = \sum_{k=1}^\infty P(X \geq k).$$

Proof. By definition

$$\begin{aligned}\mathbf{E}X &= \sum_{k=0}^{\infty} k \cdot P(X=k) \\ &= P(X=1)+ \\ &+ P(X=2) + P(X=2)+ \\ &+ P(X=3) + P(X=3) + P(X=3)+ \\ &... \text{ (by adding terms vertically)} \\ &= P(X\geq 1) + P(X\geq 2) + P(X\geq 3) + ... \\ &= \sum_{k=1}^{\infty} P(X\geq k).\end{aligned}$$

□

Integral over subset

We need often to integrate over subsets $A \in \mathcal{F}$. This is defined as follows:

$$\int_A X dP = \int_\Omega X \cdot I_A \, dP$$

where

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

By taking $X = 1$, we get

$$\begin{aligned}\int_A dP = \int_\Omega I_A \, dP = E(I_A) \\ = 1 \cdot P(A) + 0 \cdot P(A) = P(A).\end{aligned}$$

Thus we have obtained a formula showing that probabilities can always be written as integrals:

$$P(A) = \int_A dP.$$

Expected value of function of random variable

Often it is necessary to calculate the mean value of some function of r.v., for example $E(X^2)$. More generally, let $g(x)$ a function, $g : \mathcal{R} \to \mathcal{R}$. It can be

showed (following the definition of the Lebesgue integral) that the expectation of $g(X)$ can be calculated as

$$\mathbf{E}g(X) = \int_{-\infty}^{\infty} g(x)dF_X(x),$$

where $F_X(x)$ is the distribution function of X. In case when X is a continuous r.v. having density $f_X(x) = F'_X(x)$, we have $dF_X(x) = f_X(x)dx$ and thus

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

Example. Let X have uniform distribution on $[0, 1]$ and let $g(x) = x^2$. Since the density of X is 1 within the interval [0,1] and 0 elsewhere, the mean value of X^2 is

$$\mathbf{E}(X^2) = \int_0^1 x^2 dx = \frac{1}{3}.$$

Definition: The number $\mathbf{E}(X^k)$ is called *k-th moment* of X,
the number $\mathbf{E}|X|^k$ is called *k-th absolute moment* of X.

Definition: The number $DX = \mathbf{E}(X - \mathbf{E}X)^2$ is called the **variance** of X.

Remark: Quite often, the variance is also denoted as $var(X)$.

The variance is an important measure of fluctuation of the values of X around its expected value $\mathbf{E}X$. The properties of variance can easily be derived from the properties of expectation:

$$D(cX) = c^2 DX$$
$$D(X + c) = DX$$
$$D(X + Y) = DX + DY + 2cov(X, Y).$$

HOMEWORK

HW8. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability triple where $\Omega = [0, 1]$, $\mathcal{F}$ is Borel σ-algebra on $[0, 1]$ and $\mathbf{P}$ is Lebesgue measure. Give an example of random variables X and Y defined on $(\Omega, \mathcal{F}, \mathbf{P})$ such that $\mathbf{P}(X < Y) > 1/2$, but $\mathbf{E}X > \mathbf{E}Y$.

HW9. Give (with proof) an example of two discrete random variables having the same mean and the same variance, but which are not identically distributed (i.e. have different distributions).

4 Inequalities

Next simple inequalities are widely used in probability theory.

Markov inequality: If $X \geq 0$, then for all $\alpha > 0$, one has

$$\mathbf{P}(X \geq \alpha) \leq \frac{\mathbf{E}X}{\alpha}.$$

Proof: Define

$$X' = \begin{cases} \alpha, & \text{if} \quad X \geq \alpha \\ 0, & \text{if} \quad X < \alpha. \end{cases}$$

Then, clearly, $X' \leq X$ and using monotonicity property, $\mathbf{E}X' \leq \mathbf{E}X$.
Now, $\mathbf{E}X' = \alpha \cdot \mathbf{P}(X \geq \alpha) + 0 \cdot \mathbf{P}(X < \alpha) = \alpha \cdot \mathbf{P}(X \geq \alpha)$. That is

$$\alpha \cdot \mathbf{P}(X \geq \alpha) \leq \mathbf{E}X$$

$$\Rightarrow \mathbf{P}(X \geq \alpha) \leq \frac{\mathbf{E}X}{\alpha}.$$

Extended Markov inequality: If $f(x)$ is a monotonically increasing non-negative function, then for all $\alpha > 0$ (where $f(\alpha) > 0$), one has

$$\mathbf{P}(X \geq \alpha) \leq \frac{\mathbf{E}(f(X))}{f(\alpha)}.$$

Chebyshev inequality: Let Y be an arbitrary r.v. with finite mean $\mathbf{E}Y$, then for all $\alpha > 0$, one has

$$\mathbf{P}(|Y - \mathbf{E}Y| \geq \alpha) \leq \frac{DY}{\alpha^2}.$$

Proof: Since $|Y - \mathbf{E}Y| \geq \epsilon$ and $|Y - \mathbf{E}Y|^2 \geq \epsilon^2$ are equivalent events, we have

$$\mathbf{P}(|Y - \mathbf{E}Y| \geq \alpha) = \mathbf{P}(|Y - \mathbf{E}Y|^2 \geq \alpha^2).$$

Now, using Markov inequality, we get,

$$\mathbf{P}(|Y - \mathbf{E}Y|^2 \geq \alpha^2) \leq \frac{\mathbf{E}|Y - \mathbf{E}Y|^2}{\alpha^2} = \frac{DY}{\alpha^2}.$$

Cauchy-Schwarz inequality: Let X and Y be r.v. with finite 2-nd moments. Then

$$\mathbf{E}|XY| \leq \sqrt{\mathbf{E}(X^2)\mathbf{E}(Y^2)}.$$

Jensen inequality: Let $\mathbf{E}X$ be a r.v. with finite mean, and let $g(x)$ be a convex function, i.e. a function such that $\lambda g(x) + (1-\lambda)g(y) \geq g(\lambda x + (1-\lambda)y)$ for all $x,\ y \in \mathcal{R}$ and $0 \leq \lambda \leq 1$. Then

$$\mathbf{E}(g(X)) \geq g(\mathbf{E}X)).$$

Proof (in special case): Consider a special case where $g(X) = X^2$, then we have to show that $\mathbf{E}(X^2) \geq (\mathbf{E}X)^2$. But this we know already before because

$$0 \leq DX = \mathbf{E}(X - \mathbf{E}X)^2 = \mathbf{E}(X^2) - (\mathbf{E}X)^2$$

$$\Rightarrow \mathbf{E}(X^2) \geq (\mathbf{E}X)^2.$$

Using a bit more complicated method, one can also obtain

$$\mathbf{E}(g(X)) \geq g(\mathbf{E}X)).$$

HOMEWORK

HW10. Suppose $\mathbf{E}(3^X) = 9$. Prove that $\mathbf{P}(X \geq 4) \leq 1/9$.

HW11. Show that no more than 1/10 of the population can have more than 10 times the average income.

5 Conditional probabilities and expectations

The *conditional probability* of event A given the event B has occured is defined as

$$P(A|B) = \frac{P(AB)}{P(B)}. \tag{1}$$

Here we assume that $P(B) > 0$.

Example 1. *A family has two children. One child is a boy. What is the probability that the other child is a girl?*

Solution: The event space $\Omega = \{bb, bg, gb, gg\}$, where b ja g denote boy and girl (older first). We know that the event $B = \{bb, bg, gb\}$ has occured and we are finding the conditional probability of $A = \{gb, bg\}$ given B. According to the definition

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{\frac{2}{4}}{\frac{3}{4}} = \frac{2}{3}.$$

Remarque: The formula (1) defines a new probability measure on $\mathcal{F}$: for each $A \in \mathcal{F}$ we define

$$P_B(A) = P(A|B). \tag{2}$$

Multiplication rule

From the formula of conditional probability it follows that

$$P(AB) = P(B)P(A|B) = P(A)P(B|A)$$

Formula of total probability

Let the event space Ω be divided into disjoint events $B_1, \ldots, B_n$ so that $\cup_{j=1}^n B_j = \Omega$. Such a system $\{B_1, \ldots, B_n\}$ is called a **partition** of Ω. Then for any event A we have the following formula (assuming that $P(B_j > 0)$):

$$P(A) = \sum_{j=1}^{n} P(A|B_j)P(B_j). \tag{3}$$

Proof. Event A can be represented as a union of disjoint subsets $A = \cup_{j=1}^n (A \cap B_j)$. Applying additivity of P and the multiplication rule, one can write

$$P(A) = P(\bigcup_{i=j}^{n}(A \cap B_j) = \sum_{j=1}^{n} P(A \cap B_j) = \sum_{j=1}^{n} P(B_i)P(A|B_j).$$

◀

Conditional expectation

Let X be a random variable and assume that an event B has occured. This information affects on the mean value of X because now only elementary events $\omega \in B$ should be taken into account when finding average of the values $X(\omega)$. We define conditional expectation of X given the event B as its expected value with respect to the conditional probability measure $P_B(A)$ defined by (2):

$$E(X|B) = \int_\Omega X(\omega)\mathbf{P}_B(d\omega). \tag{4}$$

We know that if X is discrete r.v. with values $x_1, x_2, ...$ then the last mean value writes as

$$\mathbf{E}(X|B) = \sum_i x_i P(X = x_i|B). \tag{5}$$

Now, let us have a partition $\{B_1, \ldots, B_n\}$, satisfying $P(B_j > 0)$. If we write down the conditional expectation $\mathbf{E}(X|B_j)$ for each B_j, and on the right hand side of (5) replace $P(X = x_i|B_j) = \frac{P((X=x_i)\cap B_j)}{P(B_j)}$, we reach a formula, similar to the formula of total probability:

$$\mathbf{E}X = \sum_{j=1}^{n} \mathbf{E}(X|B_j)P(B_j). \tag{6}$$

In a similar way it can be shown that the formula (6) remains true for arbitrary random variable X (not only discrete ones).[1]

The formula (6) represents $\mathbf{E}X$ as weighted average of n conditional expectations $\mathbf{E}(X|B_j)$, each weighted by its probability $P(B_j)$.

The principle of using conditional expectations for the calculation of the (total) expectation $\mathbf{E}X$ is called *conditioning.*

Exercise: How many coin tosses are needed (in average) in order to get the first 'head'?

Exercise: How many tosses of dice are needed (in average) in order to get all numbers?

[1]For general X we have $\mathbf{P}_{B_j}(d\omega) = \frac{P(d\omega\cap B_j)}{P(B_j)}$ in (4), which leads to $\mathbf{E}(X|B_j) = \frac{1}{P(B_j)}\int_{B_j} X(\omega)P(d\omega)$. Now (6) follows immediately.

In particular, if the partition is induced by a discrete r.v. Y with values y_j, that is $B_j = \{Y = y_j\}$, we get from (6) that

$$\mathbf{E}X = \sum_j \mathbf{E}(X|Y = y_j)P(Y = y_j).$$

This is, in fact, a special case of a more general formula of 'total' expectation. Let Y be an arbitrary random variable with distribution function $G(y)$. Then we have

$$\mathbf{E}X = \int_{-\infty}^{\infty} \mathbf{E}(X|Y = y)dG(y), \tag{7}$$

where $dG(y)$ acts as weight function of conditional means $\mathbf{E}(X|Y = y)$. Since probability $P(A)$ can always be written as expected value (indeed $P(A) = \mathbf{E}I_A$), we have, as a special case of (7):

$$P(A) = \int_{-\infty}^{\infty} P(A|Y = y)dG(y). \tag{8}$$

Exercise. Let X, Y be independent exponentially distributed RV's: $X \sim \mathcal{E}xp(\lambda_1)$, $Y \sim \mathcal{E}xp(\lambda_2)$. Find the probability $\mathbf{P}(X < Y)$.
Hint: Condition on Y using $\mathbf{P}(X < Y|Y = y) = \mathbf{P}(X < y) = 1 - e^{-\lambda_1 y}$.
Answer: $\mathbf{P}(X < Y) = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

The formula (8) can also be used, for example, to study the distribution of the sum of two random variables X and Y.

HW12. Let X and Y be independent exponentially distributed random variables, $X \sim \mathcal{E}xp(\lambda)$, $Y \sim \mathcal{E}xp(\lambda)$. Find the density function of the sum $S = X+Y$. Sketch the graph of the density function obtained.
Hints: 1) First find $F_S(t) = P(X + Y \leq t)$ - the distribution function of S, using conditioning on Y. Note that the equality $P(X \leq t-y) = 1-e^{-\lambda(t-y)}$ holds only for $y \leq t$.
2) Calculate the density of S as derivative of the distribution function obtained.

6 Independence. Laws of large numbers.

Definition: Two events A and B are called *independent*, if

$$P(AB) = P(A) \cdot P(B).$$

Now let us have three events A, B, C. If

$$P(AB) = P(A) \cdot P(B)$$
$$P(BC) = P(B) \cdot P(C)$$
$$P(AC) = P(A) \cdot P(C),$$

then A, B, C are called *pairwise independent*. If, in addition, also

$$P(ABC) = P(A) \cdot P(B) \cdot P(C),$$

then A, B, C are called *completely independent*.
These concepts can easily be extended to more than three events.

Definition: Random variables X and Y are called independent if

$$P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y) \quad \text{for all} \quad x, y \in \mathcal{R}.$$

Comment: The last requirement can alternatively be written as:

$$P(X \in B_1, Y \in B_2) = P(X \in B_1) \cdot P(Y \in B_2) \text{ for any Borel sets } B_1, B_2 \in \mathcal{B}.$$

Expected value of the product of independent r.v. X and Y:

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

Variance of the sum of independent r.v. X and Y:

$$D(X + Y) = DX + DY.$$

Comment: The last formula remains true if (instead of independence) X and $Y$are uncorrelated, i.e. $cov(X, Y) = 0$.

Convergence in probability

Definition. We say that the sequence of random variables $X_1, X_2, \ldots$ *converges* to a random variable X *in probability* if for any $\epsilon > 0$

$$\mathbf{P}(|X_n - X| < \epsilon) \to 1, \quad n \to \infty.$$

Convergence in probability is denoted by $X_n \xrightarrow{P} X$.

Convergence almost surely (with probability one)

Definition. We say that the sequence of random variables $X_1, X_2, \ldots$ *converges* to a random variable X *almost surely* if

$$\mathbf{P}(\lim_n X_n = X) = 1.$$

Convergence almost surely is also called *convergence with probability 1.* Convergence almost surely is denoted by $X_n \xrightarrow{a.s.} X$.

Lemma 1. Convergence almost surely implies convergence in probability, i.e. if $X_n \xrightarrow{a.s.} X$, then also $X_n \xrightarrow{P} X$.
(NB! The converse is not true.)

Proof of Lemma 1. We show that the almost sure convergence is equivalent to the following condition: for any $\epsilon > 0$

$$\mathbf{P}(\bigcap_{k=n}^{\infty} |X_k - X| < \epsilon) \to 1, \quad n \to \infty.$$

Weak Law of Large Numbers: Let $X_1, X_2, \ldots$ be independent identically distributed (IID) random variables with finite mean a and variance σ^2. Then partial arithmetic means converge in probability to a:

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \xrightarrow{P} a, \quad n \to \infty.$$

Example: A student throws a dice 400 times. Without knowing the results, what is the arithmetic mean (approximately) of these 400 throws ? (Answer: ≈3,5)

HW13. Prove the Weak Law of the Large Numbers (see above).
Hints: Denote $S_n = X_1 + X_2 + \ldots + X_n$.
1) Show that the arithmetic mean S_n/n has expected value a and variance σ^2/n.
2) Apply Chebyshev inequality for the random variable S_n/n to obtain required convergence $S_n/n \xrightarrow{P} a$ (in probability).

Strong Law of Large Numbers: Let $X_1, X_2, \ldots$ be IID random variables with finite mean a. Then partial arithmetic means converge almost surely to a:

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \xrightarrow{a.s.} a, \quad n \to \infty.$$

Proof: The proof of SLLN is more involved and we omit it.

7 Sums of random variables. Central Limit Theorem

7.1 Distribution of the sum

One of the important problems in probablity theory is to find the distribution of the sum of two or more random variables. Let X and Y be independent random variables, where $X \sim F$ and $Y \sim G$. We are interested in finding the distribution of $X + Y$ i.e. $F_{X+Y}(t) = P(X + Y \leq t)$. To find this, we can use conditioning on Y, i.e. the formula (8) where $A = \{X + Y \leq t\}$. We 1) let Y to take a fixed value $Y = y$, 2) calculate the conditional probability $P(X + Y \leq t | Y = y)$, and 3) find weighted average of conditional probabilities using the distribution G of

r.v. Y. Formally, it can be written as:

$$\begin{aligned} F_{X+Y}(t) = P(X+Y \le t) &= \int_{-\infty}^{\infty} P(X+Y \le t | Y = y)) dG(y) \\ &= \int_{-\infty}^{\infty} P(X+y \le t) dG(y) \\ &= \int_{-\infty}^{\infty} P(X \le t-y) dG(y) \\ &= \int_{-\infty}^{\infty} F(t-y) dG(y). \end{aligned}$$

Note that the second equality is due to the independence of X and Y. The last expression is called *the convolution* of $F$and G and denoted by $F * G(t)$. Thus, the distribution of the sum is given by

$$F_{X+Y}(t) = \int_{-\infty}^{\infty} F(t-y) dG(y) \equiv F * G(t).$$

Often the distribution of IID r.v. $X_1, X_2, \ldots, X_n$, each having distribution F, is of interest. Then the convolution operation can be applied recursively, e.g. in case of $n = 3$ we have

$$F_{X_1+X_2+X_3}(t) = F * (F * F)(t) := F^{*3}(t),$$

etc. Calculation of convolution by integration (as above) is not very convenient tool to be used in practice. However, there are more advanced methods to obtain the distribution of sums: moment generating functions and characteristic functions.

7.2 Moment generating functions

Definition: The moment generating function (mgf) of a random variable X is defined as

$$m_X(t) = \mathbf{E}(e^{tX}),\ t \in \mathcal{R}.$$

Ex.1. (mgf of exponential distribution). Let $X \sim Exp(\lambda)$. Then

$$m_X(t) = \mathbf{E}(e^{tX}) = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = ... = \frac{\lambda}{\lambda - t},$$

for $t < \lambda$. However, $m(t) = \infty$ for $t \geq \lambda$. **Ex.2.** (mgf of normal distribution). If X has standard normal distribution, $X \sim \mathcal{N}(0,1)$, then

$$m_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \tag{9}$$

$$= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} dx \tag{10}$$

$$= e^{\frac{1}{2}t^2}, \tag{11}$$

since the integrand in the latter integral is the density function of the normal distribution with mean t and variance 1, and thus has integral 1. This mgf exists for all $t \in \mathcal{R}$.

From this example we see that the existence of mgf (finitness of respective expectation) is not always ensured which makes the range of applications of mgf somewhat limited.

However, moment generating functions share many useful properties.

1. Mgf and moments

Knowing the mgf of a distribution (random variable), it is easy for example to calculate its moments. Let us show that

1. $m'(0) = \mathbf{E}X$.
2. $m''(0) = \mathbf{E}(X^2)$, *etc.*

Proof: We differentiate w.r.t. the argument t:

$$m'(t) = [\mathbf{E}(e^{tX})]' = \mathbf{E}[(e^{tX})'] = \mathbf{E}[Xe^{tX}],$$

therefore

$$m'(0) = \mathbf{E}(Xe^0) = EX.$$

Similarly,

$$m''(0) = \mathbf{E}[X^2 e^{tX}]|_{t=0} = \mathbf{E}X^2.$$

Ex.3. Show, by using mgf, that if $X \sim Exp(\lambda)$, then $EX = \frac{1}{\lambda}$ and $DX = \frac{1}{\lambda^2}$.

2. Mgf of sum

Let X and Y be independent random variables with mgf $m_X(t)$ and $m_Y(t)$ respectively. Then the mgf of their sum can be calculated as

$$m_{X+Y}(t) = m_X(t) \cdot m_Y(t).$$

Also, knowing the mgf, it is often possible to recognize the type (shape) of the distribution. This is based on the following

3. Uniqueness Theorem of mgf. If there exists $\delta > 0$ such that $m_X(t) = m_Y(t) < \infty$ for all $t \in (-\delta, \delta)$, then the distributions of X and Y are identical (coincide), $F_X(t) = F_Y(t)$ for all $t \in \mathcal{R}$.

Ex.4. Let X and Y be independent Poisson distributed random variables, $X \sim \texttt{Pois}(\lambda_1)$ and $Y \sim \texttt{Pois}(\lambda_2)$. Show that then $X + Y$ is also Poisson distributed, $X + Y \sim \texttt{Pois}(\lambda)$, where $\lambda = \lambda_1 + \lambda_2$.

Hint: First show that $m_X(t) = e^{\lambda_1(e^t-1)}$ and $m_Y(t) = e^{\lambda_2(e^t-1)}$. Then use property 2 to calculate $m_{X+Y}(t) = e^{(\lambda_1+\lambda_2)(e^t-1)}$. We recognize this as being the mgf of the Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2$, and by the uniqueness theorem, we deduce that $X + Y$ has this distribution.

HW 14. Let X and Y be independent normally distributed random variables, $X \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2)$. Show that then $X + Y$ is also normally distributed, $X + Y \sim \mathcal{N}(\mu, \sigma)$, where $\mu = \mu_1 + \mu_2$ and $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$.

Hint: Show first that if $Z \sim \mathcal{N}(0, 1)$ then $m_Z(t) = e^{t^2/2}$. From this and from $X = \sigma_1 Z + \mu_1$ deduce that $m_X(t) = e^{(\mu_1 t + \frac{1}{2}\sigma_1^2 t^2)}$. Now continue as in the previous example.

HW 15. If X has the normal distribution with mean μ and variance σ^2, find $E(X^3)$.

Ex.7. Show that if X has normal distribution then so does $aX + b$ for any $a, b \in \mathcal{R}$ with $a \neq 0$.

7.3 Characteristic functions

We introduce another type of functions for studying the distribution of sums (and other problems in probability), called characteristic functions. Their advantage is that they are, in contrast to mgf, defined for all distributions.

Consider imaginary unit $i = \sqrt{-1}$. Euler formula says that $e^{ix} = \cos x + i \sin x$. Note that the module $|e^{ix}| = \sqrt{\cos^2 x + \sin^2 x} = 1$.

Definition. The *characteristic function* of a r.v. X is defined as

$$\phi_X(t) = \mathbf{E}(e^{itX}), \quad t \in \mathcal{R}.$$

According to Euler's formula, we have

$$\phi_X(t) = \mathbf{E}(e^{itX}) = \mathbf{E}\cos(tX) + i\,\mathbf{E}\sin(tX).$$

Example 2. ***Coin.*** *Let X take values 1 and -1 with probability $1/2$ and $1/2$.*

$$\phi_X(t) = \frac{1}{2}(e^{it} + e^{-it}) = cost$$

Example 3. ***Bernoulli.*** *Let X take values 1 and 0 with probability p and $1-p$.*

$$\phi_X(t) = e^{it\cdot 1} \cdot p + e^{-it\cdot 0} \cdot (1-p) = e^{it}p + 1 - p = 1 + p(e^{it} - 1)$$

Example 4. ***Standard normal distribution.*** *It can be shown that if $X \sim N(0,1)$, then its c.f.*

$$\phi(t) = e^{-t^2/2}.$$

The last result will be used to prove one of the most outstanding achievements of the probability theory - the Central Limit Theorem.

Properties of c.f.:

1. $\phi_X(0) = 1$. Indeed, $\phi_X(0) = \mathbf{E}(e^{i0X}) = \mathbf{E}(e^0)) = \mathbf{E}(1) = 1$.
2. $|\phi_X(t)| \leq 1, \quad \forall t$. Indeed, $|\phi_X(t)| = |\mathbf{E}(e^{itX})| \leq \mathbf{E}|e^{itX}| = \mathbf{E}(1) = 1$.

3. $\phi_{aX+b}(t) = e^{itb} \cdot \phi(at)$

4. The characteristic function of the sum of independent r.v. X and Y

$$\phi_{X+Y}(t) = \phi_X(t) \cdot \phi_Y(t).$$

Indeed, $\phi_{X+Y}(t) = \mathbf{E}(e^{it(X+Y)}) = \mathbf{E}(e^{itX} \cdot e^{itY}) = \mathbf{E}(e^{itX}) \cdot \mathbf{E}(e^{itY}) = \phi_X(t) \cdot \phi_Y(t)$.

5. *The uniqueness theorem.* There is one-to-one correspondence between distributions and characteristic functions: $X \stackrel{D}{=} Y \Leftrightarrow \phi_X(t) = \phi_Y(t)$, $t \in \mathcal{R}$.

6. *The theorem of convergence*: Let F_n and F be distributions functions and $\phi_n(t)$, $\phi(t)$ their respective c.f. If $\phi_n(t) \to \phi(t) \ \ \forall t$, then $F_n(x) \to F(x)$ for each $x \in C(F)$ (the set of continuity points of F).

7.4 Central Limit Theorem

Why the heights of men (or women) are normally distributed? Why the normal distribution is so much used in practice? The answer is given by the following Central Limit Theorem, stating that if a random variable can be regarded as a sum of a big number of small and independent addends X_i, then it is (at least approximately) normally distributed. As to the height of people, it results from a very big number of genetic and environmental (including nutrition) random factors X_i, each having only a small effect on the height, but taken together the total result follows normal distribution.

Theorem. Let $X_1, X_2, \ldots$ be iid with mean a and variance $\sigma^2 < \infty$. Then for any $x \in \mathcal{R}$

$$P\left(\frac{S_n - na}{\sqrt{n}\,\sigma} \le x\right) \to \Phi(x), \quad n \to \infty,$$

where $\Phi(x)$ is the distribution function of the standard normal distribution,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Proof of CLT: Let us write $S_n = X_1 + X_2 + \cdots + X_n$, then $\mathbf{E}(S_n) = na$ and, by independence, $D(S_n) = n\sigma^2$. So the standard deviation of S_n is equal to $\sqrt{DS_n} = \sqrt{n}\sigma$. Let us denote

$$Y_n = \frac{S_n - na}{\sqrt{n}\sigma}.$$

Since $\mathbf{E}Y_n = 0$ and $DY_n = D(\frac{S_n}{\sqrt{n}\sigma}) = \frac{D(S_n)}{n\sigma^2} = 1$ the term *standardized sums* is used for Y_n. To prove CLT, it is sufficient to show that the characteristic function of Y_n converges to characteristic function $N(0,1)$. i.e. $\phi_{Y_n}(t) \to e^{\frac{-t^2}{2}}$.
Assume (without loss of generality) $a = 0$, $\sigma = 1$ (otherwise, we can introduce random variables $X_i' = \frac{X_i - a}{\sigma}$, which would have same standardized sums Y_n.) Now, using properties 3 and 4,

$$\phi_{Y_n}(t) = \phi_{\frac{S_n}{\sqrt{n}}}(t) = \phi_{S_n}\left(\frac{t}{\sqrt{n}}\right) = \left[\phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

Now, using Taylor expansion $e^{ix} = 1 + ix + \frac{i^2t^2}{2!} + o(t^2)$, where $o(t^2)$ is an infinitesimal which converges to 0 faster than t^2 (when $t \to 0$), we calculate

$$\phi_{X_1}\left(\frac{t}{\sqrt{n}}\right) = E\left(1 + i\frac{t}{\sqrt{n}}X_1 - \frac{t^2}{2n}X_1^2 + \ldots\right)$$

$$= 1 + i\frac{t}{\sqrt{n}}\mathbf{E}X_1 - \frac{t^2}{2n}\mathbf{E}X_1^2 + o\left(\frac{t^2}{n}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right),$$

since according to our assumption $\mathbf{E}X_1 = a = 0$, $\mathbf{E}X_1^2 = \sigma^2 = 1$. Therefore, we obtain

$$\phi_{Y_n}(t) = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n.$$

However, from calculus we know that if $x_n \to x$, then $\lim\left(1 + \frac{x_n}{n}\right)^n = e^x$, which gives in our case

$$\phi_{Y_n}(t) \to e^{-\frac{t^2}{2}}.$$

The proof is completed. ◀

HW 16. Let $\{X_n\}$ be IID, each having mean 2 and variance 9. Let us denote $S = X_1 + X_2 + \ldots + X_{10,000}$. Give an approximate value for the probability $\mathbf{P}(19\ 500 < S \leq 20\ 500)$.

HW Open and enjoy the applet https://www.geogebra.org/m/F9qSxXwP which demonstrates graphically how the Poisson distribution converges to the normal distribution (when the mean value m increases). Try to give theoretical explanation to this phenomena. (Hint: First use an earlier home exercise to present the Poisson variable X of mean m as the sum of m independent Poisson variables each having unit mean (assume m in integer).)

Class exercise on application of CLT. Throw 6 dice and add up the results (do it repeatedly). Sketch the histogram of totals of six throws. Is it similar to the normal density curve?

Random processes

Introduction

What is a random process? Many random variables vary in time, for example the water level in the river, the price of a stock on stock exchange. Then it is useful to add one more argument to random variables - time. It gives many new opportunities for studying different real processes like economic, physical, social, biological etc processes.

A random process is (or stochastic process) is a family of random variables $\{X(t) : t \in T\}$, where each $X(t)$ is a random variable is usual sense. We assume here that all r.v. $X(t),\ t \in T$, are defined on a same probability space $(\Omega, \mathcal{F}, P)$.

The parameter t is usually a real-valued variable, often interpreted as time. The set T is called the **index-set** of the process. If T is a countable set, then we say that the random process has *discrete time*. In that case we denote $T = \{0, 1, 2, \ldots\}$, although it does not necessarily mean physically equal time intervals. If T is an interval of the real line, e.g. $T = [0, \infty)$, then we say that the random process has *continuous time*.

The set S consisting of all possible values of $X(t)$ for a given t is called the **state space** and $X(t)$ is the **state** of the process at time t. The state space can be the real line $\mathcal{R}$, or its subset e.g. $\mathcal{N}$ (natural numbers) or a finite set as $\{0, 1, \ldots, k\}$. More generally, the state space can be a subset of $\mathcal{R}^d$.

Le us consider some simple examples of random processes.

Example 5. *The number of cars $X(t)$ on a parking lot at time t is a process with continuous time with state space $S = \{0, 1, 2, \ldots, M\}$. An interesting question here is: what is the probability that there is at least one parking space available at time t_0?*

Example 6. *Let $X(t)$ be water level behind the dam at time t. This is a random process that is continuous in time with the state space equal to certain interval $S = [0, a]$. One natural question here is: what should be the dam height H so that the risk of flood would be virtually eliminated? How long can be the time to the first flood given the dam height H?*

Example 7. *Let X_n be the number of bicycles left in the store at the end of week n. Then $\{X_n; n = 0, 1, 2, \ldots\}$ is a discrete time random process. A possible question is: for which values of X_n it is necessary to order new bicycles?*

The random process $X(t)$ is, in fact, a function of two arguments $X(\omega, t)$. If we fix $\omega = \omega_0$, then we obtain a function of one variable $X(\omega_0, t)$ which is called a *trajectory* of the random process. One can say that a random process is a collective of its possible trajectories each corresponding to an ω.

From the other side, if we fix the time, $t = t_0$, then we obtain the cross-section of the trajectories at time t_0, or a random variable $X(\omega, t_0)$.

For two time instants $t_1 < t_2$, the difference $X(t_2) - X(t_1)$ is called the increment of the process in the time interval (t_1, t_2). If for all non-overlapping time intervals (t_1, t_2) and (t_3, t_4) their respective increments $X(t_2) - X(t_1)$ and $X(t_4) - X(t_3)$ are independent random variables, then $X(t)$ is said to be a process with **independent increments**. If the distribution of $X(s + t) - X(s)$ depends on t, but not on s, then the process is said to have **stationary increments**.

Several processes we will study (random walk, Poisson process, Brownian motion) have independent and stationary increments. However, we start with a broader class of processes called Markov chains.

8 Markov chains

It happens often that the outcome (and their probabilities) of each next random trial depend on the outcome of the previous trial. For example, tomorrow's

weather depends on the weather today. Markov chains are a proper tool to handle such processes.

8.1 Definition of MC

Let a random trial $\mathcal{K}$ have at most countably many possible outcomes $E_0, E_1, E_2, \ldots$. Repeating the trial one gets a random sequence, e.g. $E_4, E_1, E_3, E_7, E_1, E_5$ jne. For convenience, we will simply write j instead of E_j. Then the outcome of the n-th trial is a random variable X_n with possible values $0, 1, 2, \ldots$.

If X_n depends on the past values only through X_{n-1} then we have a Markov chain.

Definition 1 (Markov chain)**.** *The sequence of r.v.* $\{X_n\}$*, where* $n = 0, 1, 2, \ldots$*, is called* ***Markov chain****, if for any values* $j, k_0, k_1, \ldots, k_{n-1}$ *the conditional probability*

$$P\{\underbrace{X_n = j}_{future} \mid \underbrace{X_0 = k_0, \ldots, X_{n-2} = k_{n-2}}_{past}, \underbrace{X_{n-1} = k_{n-1}}_{present}\} =$$
$$= P\{X_n = j \mid X_{n-1} = k_{n-1}\}.$$

The same in short: Given the present, the future does not depend on the past. In other words, in order to predict X_n, it is enough to know X_{n-1}, since it contains all the necessary information about the past.

The time n takes values $0, 1, 2...$, and thus we have discrete time MC's here. (In principle, there are also continuous time MC's.)

Possible values of X_n (i.e. $0, 1, 2, \ldots$) are called the **states** of MC. Conditional probability

$$P\{X_n = j \mid X_{n-1} = i\} =: {p_{ij}}^{(n)}$$

is called transition probability from state i to state j on the step n. The initial state X_0 is usually given by its probability distribution (*initial distribution*)

$$P\{X_0 = j\} = {p_j}^0, \quad j = 1, 2, \ldots, \quad \sum_j {p_j}^0 = 1.$$

Definition 2 (Homogeneous Markov chain)**.** *If the transition probabilities do not depend on the time n, i.e. for each n we have ${p_{ij}}^{(n)} =: p_{ij}$, then the MC is called* ***homogeneous****.*

The matrix

$$\mathbf{P} = (p_{ij}),\ i, j = 0, 1, 2, \ldots$$

is called **transition matrix** of the MC. The i-th row of P is the conditional distribution of states j on the next step, given that the current state is i. Therefore, the rows of P always sum up to one, $\sum_j p_{ij} = 1$.

Example 8. *(Simple model for weather) Suppose tomorrow's weather only depends on today's and earlier weather data does not help to improve the forecast for tomorrow. For simplicity, we distinguish only between two states: 0 – it rains, 1 – it does not rain. Suppose if it rains today, then it will rain tomorrow with probability α, and it is does not rain today then it will rain tomorrow with probability β. Then we have a MC with transition matrix*

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{pmatrix} \end{array} \begin{array}{c} \sum \\ 1 \\ 1 \end{array},$$

where the last column Σ is for checking if the row totals are 1.

Example 9. *(Communication model) Consider transmission of digital information (i.e. sequence zeros and of ones) by multiple steps. At each step an error can occur of types $0 \to 1$ or $1 \to 0$ with probability p. The transition matrix will be*

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \end{array}.$$

Example 10. *(Gambling model) Consider a player who starts the game with k euros and who can earn in each game 1 euro with probability p vor lose 1 euro with probability $1 - p$. The games are independent. The player quits the game if he has ruined (i.e. his wealth is zero) or if his wealth has reached N euros ($N \geq k$). Let X_n be player's wealth after the game n. Since the value of X_n*

depends only on the value of X_{n-1}, and the outcome of the game n, the process $\{X_n\}$ is MC with possible states $0, 1, 2, \ldots, N$. The transition matrix in this case will be 'three-diagonal':

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & 3 & \cdots & N & \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 \\ 0 & 1-p & 0 & p & \cdots & 0 \\ 0 & 0 & 1-p & 0 & \cdots & 0 \\ \vdots & \vdots & & & \ddots & \\ 0 & 0 & \ldots & & 0 & 1 \end{pmatrix} \end{array} \begin{array}{c} \sum \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{array},$$

where for $i = 1, \ldots, N-1$ on has $p_{i,i+1} = p$ (wins 1 euro), $p_{i,i-1} = 1-p$ (loses 1 euro), $p_{0,0} = 1$ (ruin) and $p_{N,N} = 1$ (reaching a 'fortune').

Definition 3. *The state i with $p_{ii} = 1$ is called **absorbing state**.*

In this example, we have 2 absorbing states: 0 and N – the MC never leaves these states.

Example 11. *(Transforming a process into a Markov chain) Suppose that tomorrow's weather depends on not only on today's wether but also on yesterday's weather (longer memory). That is, suppose that*
if it has rained for the past two days, then it will rain tomorrow with probability 0,7;
if it rained today but not yesterday, then it will rain tomorrow with probability 0,5;
if it rained yesterday but not today, then it will rain tomorrow with probability 0,4;
if it has not rained in the past two days, then it will rain tomorrow with probability 0,2.

If we let the state at time n be just 'rains =0' or 'does not rain=1', then the model above is not a MC. However, we can transform the model into a MC by redefining states as follows:
(it rained yesterday and today)= state 0,

(it rained today but not yesterday)= state 1,

(it rained yesterday but not today)= state 2,

(it did not rain either yesterday or today)= state 3.

The above model then would represent a four-state Markov chain having a transition probability matrix

$$\boldsymbol{P} = \begin{pmatrix} .7 & 0 & .3 & 0 \\ .5 & 0 & .5 & 0 \\ 0 & .4 & 0 & .6 \\ 0 & .2 & 0 & .8 \end{pmatrix}.$$

Example 12. *(Random walk on the line) A particle starts moving form an integer point X_0 and at each step it moves either 1 unit to the right (with probability p) or 1 unit to the left (with probability $q = 1 - p$). Each step is independent from all previous steps.*

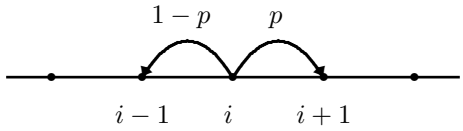

Figure 1: Random walk on the line

Here we have a MC where $p_{i,i+1} = p$, $p_{i,i-1} = 1 - p$ and $p_{i,i} = 0$, $i = \pm 1, \pm 2, \ldots$. A random walk is called ***symmetric****, if $p = q = \frac{1}{2}$.*

Definition 4. *A state i satisfying $p_{i,i-1} = 1$ or $p_{i,i+1} = 1$ is called* ***reflecting*** *(****from above*** *or* ***from below****, respectively).*

Example 13. *(Random walk with reflection) A particle starts from k (≥ 0) and at each step it moves either 1 unit to the right (with probability p) or 1 unit to the left (with probability $q = 1 - p$). Let 0 and N ($N \geq k$) be reflecting states. Then the transition matrix is:*

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & \cdots & N \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 \\ 0 & 1-p & 0 & p & \cdots & 0 \\ 0 & 0 & 1-p & 0 & \cdots & 0 \\ \vdots & \vdots & & & \ddots & \\ 0 & 0 & \ldots & & 1 & 0 \end{pmatrix} \end{array} \begin{array}{c} \sum \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{array}$$

8.2 k-step transition

Let us have a homogeneous MC with transition matrix $\mathbf{P} = (p_{ij})$. Let's find the probability $p_{ij}(k)$ that starting from the state i the process will be in state j after k additional transitions (intermediate visits of j are also allowed). For homogeneity we have

$$p_{ij}(k) = P\{X_k = j | X_0 = i\}. \tag{12}$$

Let m be an intermediate number of steps, $0 < m < k$. Since the events $\{X_m = l\}, l = 0, 1, \ldots$ are non-overlapping, the probability $p_{ij}(k)$ can be expressed as the sum

$$p_{ij}(k) = \sum_{l=0}^{\infty} P\{X_m = l, X_k = j | X_0 = i\},$$

from where (using $P(BC|A) = P(B|A) \cdot P(C|AB)$) we have

$$p_{ij}(k) = \sum_{l=0}^{\infty} P\{X_m = l | X_0 = i\} \cdot P\{X_k = j | X_0 = i, X_m = l\}.$$

In the last probability the event $\{X_0 = i\}$ can be ignored, hence

$$p_{ij}(k) = \sum_{l=0}^{\infty} P\{X_m = l | X_0 = i\} \cdot P\{X_k = j | X_m = l\}$$

or, shortly,

$$p_{ij}(k) = \sum_{l=0}^{\infty} p_{il}(m) \cdot p_{lj}(k - m). \tag{13}$$

This is called **Chapman–Kolmogorov equation**.

We present it in matrix form. Denoting the matrix $\mathbf{P}(k) = (p_{ij}(k))$, the equation (13) can be written as:

$$\mathbf{P}(k) = \mathbf{P}(m) \cdot \mathbf{P}(k - m). \tag{14}$$

Now take $m = 1$ and, since $\mathbf{P}(1) = \mathbf{P}$, we obtain, by repeating,

$$\begin{aligned}\mathbf{P}(k) = \mathbf{P}(1) \cdot \mathbf{P}(k-1) &= \mathbf{P} \cdot \mathbf{P}(k-1) \\ &= \mathbf{P}^2 \cdot \mathbf{P}(k-2) = \dots \\ &= \mathbf{P}^{k-1} \cdot \mathbf{P}(1) \\ &= \mathbf{P}^k.\end{aligned}$$

Thus, in case of k steps the transition matrix is

$$\mathbf{P}(k) = \mathbf{P}^k.$$

Example 14. *(Weather model continued) Consider again the simple weather model, where we take $\alpha = 0.7$ and $\beta = 0.4$, i.e. the transition matrix is*

$$\boldsymbol{P} = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

Suppose it is Monday today and we want to know what will be the the weather on Wednesday. For that, we calculate 2-step transition matrix:

$$\boldsymbol{P}(2) = \boldsymbol{P}^2 = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix} \cdot \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{pmatrix}.$$

We see now that if it rains (0) today, then on Wednesday it will rain with probability 0.61. However, if there is no rain today then it will rain on Wednesday with probability 0.52.

HW17. (Weather model continued) Consider again the simple weather model with transition matrix as before:

$$\mathbf{P} = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

Find the probability that there will be no rain next Thursday if it rains today (Thursday).

HW18. Consider a symmetric random walk $\{X_n\}$ starting from a random point X_0 between 0 and 6. Find the 2-step transition matrix, if the states 0 and 6 are absorbing. What is the probability that the particle will be absorbed within 2 steps if the current state is 4?

8.3 Classification of states

Definition 5. *State j is said to be* ***accessible*** *from state i if $p_{ij}(n) > 0$ for some $n \geq 0$.*

This is denoted by $i \to j$ (otherwise we write $i \nrightarrow j$). Note that this implies that state j is accessible from state i if and only if, starting in i, it is possible that the process will ever (sooner or later) enter state j.

Definition 6 (Inessential state)**.** *State i is called* ***inessential*** *if for some state j it holds that $i \to j$ but $i \not\leftarrow j$.*

Thus it is possible to leave the inessential state i forever. Inessential state can appear only a finite number of times, sooner or later it disappears.

If the state is not inessential, it is called **essential**. If i is essential, $i \to j$ implies $i \leftarrow j$.

Example 15. *Consider the Markov chain consisting of four states $1, 2, 3, 4$ and having transition matrix*

$$\boldsymbol{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Here the state 1 is inessential, since leaving 1 for 3, it is not possible to return to 1. Also 2 is inessential, since leaving 2 for 4, it is not possible to return. However, the states 3 and 4 are essential.

Example 16. *Consider the random walk where 0 and N are absorbing states, so that the transition matrix is*

$$\begin{array}{c} \\ \boldsymbol{P} = \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & 3 & \cdots & N & \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 \\ 0 & 1-p & 0 & p & \cdots & 0 \\ 0 & 0 & 1-p & 0 & \cdots & 0 \\ \vdots & \vdots & & & \ddots & 0 \\ 0 & 0 & \ldots & & 0 & 1 \end{pmatrix} \end{array} \begin{array}{c} \sum \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{array}$$

Here the absorbing states 0 *and* N *are essential.*
However, the state 1 *is inessential, since* $p_{10} > 0$ *but* $p_{01}(t) = 0\ \forall t$ *korral.*
Also, 2 *is inessential, since* $p_{20}(2) = p_{21} \cdot p_{10} = (1-p)^2 > 0$*, but* $p_{02}(t) = 0\ \forall t = 1, 2, \ldots$ *korral.*
Similarly, it is possible to show that the states $3, \ldots, N-1$ *are inessential.*

Example 17. *Let us modify the previous example by replacing the absorbtion property of states* 0 *and* N *by reflection property. The we obtain a random walk with reflection described in Example 13. It is easy to see that now all states are essential.*

Definition 7. *Two states* i *and* j *that are accessible to each other are said to* ***communicate*** *(also concurrent), and we write* $i \longleftrightarrow j$ *.*

Note that any state communicates with itself since, by definition,

$$p_{ii}(0) = P(X_0 = i | X_0 = i) = 1.$$

The relation of communication satisfies the following three properties:
(i) State i communicates with state i, for all i.
(ii) If state i communicates with state j, then state j communicates with state i.
(iii) If state i communicates with state j, and state j communicates with state k, then state i communicates with state k.

Properties (i) and (ii) follow immediately from the definition of communication. To prove (iii) suppose that i communicates with j, and j communicates with k. Thus, there exist integers n and m such that $p_{ij}(n) > 0, p_{jk}(m) > 0$. Now by the Chapman–Kolmogorov equations, we have

$$p_{ik}(n+m) = \sum_{r=o}^{\infty} p_{ir}(n) \cdot p_{rk}(m) \geq p_{ij}(n) \cdot p_{jk}(m) > 0.$$

Hence, state k is accessible from state i. Similarly, we can show that state i is accessible from state k. Hence, states i and k communicate. Two states that communicate are said to be in the same *class*.

It is an easy consequence of (i), (ii), and (iii) that the concept of communication divides the set of all essential states into a number of disjoint classes $S^1, S^2, \ldots$ where in each class all states communicate with each other. The whole state space S can thus be represented as

$$S = S^0 \cup S^1 \cup S^2 \cup S^3 \cup \ldots,$$

where S^0 is the class of inessential states.

The work of a MC can now be described as follows. First, MC can start in an inessential state $i \in S^0$, or in an essential state $i \in S^k,\ k \geq 1$. In the first case, MC leaves (sooner or later) the class S^0, enters one of the classes $S^k,\ k \geq 1$ and stays there forever. In the second case, the process never leaves initial class S^k -all transitions will be made within that class.

Definition 8. *The Markov chain is said to be* ***irreducible*** *(or indivisible) if there is only one class, $S = S^1$, that is, if all states communicate with each other.*

8.4 Recurrence of Markov chain

For any state j we let F_j denote the probability that, starting in state j, the process will ever reenter state j. This probability can be 1 or less than 1.

Definition 9 (Recurrent state)**.** *State j is said to be* ***recurrent*** *if $F_j = 1$. If $F_j < 1$, then the state j is called* ***transient.***

Suppose that the process starts in state j and j is recurrent. Hence, with probability 1, the process will eventually reenter the state j. However, by the definition of Markov chain, it follows that the process will be starting over again when it renters state j and, therefore, state j will eventually be visited again. Continual repetition of this argument leads to the conclusion that *if state j is recurrent then, starting in state j, the process will reenter state j again and again and again - in fact, infinitely often.* More formally, if N_{jj} denotes the total number

of returns to j, starting from j, then we have $N_{jj} = \infty$ (with probability 1) which gives that also the expected number of returns to j is infinity:

$$EN_{jj} = \infty. \tag{15}$$

On the other hand, suppose that state j is transient. Hence, each time the process enters state j there will be a positive probability, namely $1 - F_j$, that it will never again enter that state. Therefore, starting in state j, the probability that the process will be in state j for exactly n time periods equals $F_j^{n-1}(1 - F_j),\ n \geq 1$. In other words, if state j is transient then, starting in state j, the number of time periods that the process will be in state j has a geometric distribution with finite mean

$$EN_{jj} = \frac{1}{1 - F_j} \quad . \tag{16}$$

From the preceding two paragraphs, it follows that state j is recurrent if and only if, starting in state j, the expected number of time periods that the process is in state j is infinite. But, letting

$$I_n = \begin{cases} 1, \text{ if } X_n = j, \\ 0, \text{ if } X_n \neq j \end{cases}$$

we have that $\sum_{n=1}^{\infty} I_n$ represents the the total number of time periods that the process is in state j. Therefore also

$$\begin{aligned} EN_{jj} &= E(\sum_{n=1}^{\infty} I_n | X_0 = j) = \sum_{n=1}^{\infty} E(I_n | X_0 = j) \\ &= \sum_{n=1}^{\infty} \left[1 \cdot P\{X_n = j | X_0 = j\} + 0 \cdot P\{X_n \neq j | X_0 = j\}\right] \\ &= \sum_{n=1}^{\infty} p_{jj}(n). \end{aligned}$$

We have thus proved the following.

Theorem 2. *State j is*

$$\textit{recurrent if} \quad \sum_{n=1}^{\infty} p_{jj}(n) = \infty, \tag{17}$$

$$\textit{transient if} \quad \sum_{n=1}^{\infty} p_{jj}(n) < \infty. \tag{18}$$

Example 18. *(Symmetric random walk, continued)*
Let us show, using Theorem 2, that all states of SRW are recurrent. Since, starting from j, it is possible to return to j only by even number of steps (with half of steps made to the right and half of steps to the left), we can write, using binomial distribution, that

if n is even number then $p_{jj}(n) = p_{jj}(2k) = \binom{2k}{k}(\frac{1}{2})^k(\frac{1}{2})^k$,

if n is odd number then $p_{jj}(n) = 0$.

To check the convergence-divergence of the series $\sum_{n=1}^{\infty} p_{jj}(n)$ we use Stirling's formula

$$k! \sim \sqrt{2\pi k}\left(\frac{k}{e}\right)^k,$$

where '$\sim$' means asymptotic equivalence.[2] *Applying Stirling's formula, we have*

$$\binom{2k}{k} = \frac{(2k)!}{k!\ k!} \sim \frac{\sqrt{2\pi 2k}\cdot 2k^{2k}\cdot e^k \cdot e^k}{e^{2k}\cdot\sqrt{2\pi k}\cdot k^k \cdot \sqrt{2\pi k}\cdot k^k} = \frac{4^k}{\sqrt{\pi k}}. \tag{19}$$

Thus

$$p_{jj}(2k) = \binom{2k}{k}\cdot\frac{1}{4^k} \sim \frac{1}{\sqrt{\pi k}}.$$

Now it is easy to check the convergence. Since the harmonic series

$$\sum_{k=1}^{\infty}\frac{1}{\sqrt{k}} = \infty,$$

then also the equivalent series diverges:

$$\sum_{k=1}^{\infty} p_{jj}(2k) = \infty.$$

[2]Two sequences are called asymptotically equivalent, $a_n \sim b_n$, if $\frac{a_n}{b_n} \to 1$, $n \to \infty$. It is known that asymptotically equivalent series converge and diverge simultaneously, i.e. if $a_n \sim b_n$, then $\sum_n a_n < \infty \Longleftrightarrow \sum_n b_n < \infty$.

thus, by Theorem 2, the state j is recurrent. Since j was arbitrary, we can say that ***all states of SRW on the real line $\mathcal{R}$ are recurrent.***

Comment: It is possible to show that the result above is also true for symmetric random walk in the space $\mathcal{R}^2$. However, in case of dimension 3 SRW stops to be recurrent. Therefore, it has been said that the wording 'All roads lead to Rome' is valid only in case of $\mathcal{R}$ and $\mathcal{R}^2$.

HW. Consider an asymmetric random walk on integer points of the real line, given by unequal transition probabilities: $p_{j,j+1} = 0.6$ and $p_{j,j-1} = 0.4$. Show that the states of this random walk are transient.

Hints: 1. In general, follow the symmetric case covered above (i.e. use the Strirling's formula). 2. At the end, note that $\sum_k 0.96^k = 24 < \infty$.

Recurrence of finite Markov chain

Consider a Markov chain with only a finite number of states (as in most of the examples above). Show that then at least one of the states is recurrent. To see this, let the states are $0, 1, \ldots, k$ and let N_j be the number or periods when the process is in state j. Then the total number of periods (which is infinity) can be divided between k states, giving $N_0 + N_1 + \ldots + N_k = \infty$. After taking expectations, one gets $EN_0 + EN_1 + \ldots + EN_k = \infty$. However, then at least one of the addends on the left hand side, say EN_l, must also be infinity, but this means that the state l is recurrent.

Recurrence is a class property

Consider a class of communicating states, say S^1. Then it is enough to check recurrence for one single state only.

Solidarity theorem. If any of the states of a class of communicating states is recurrent, then the same is true for all other states in this class.

Proof: Let the state $j \in S^1$ be recurrent. Consider any other state $k \in S^1$ and

show this is also recurrent. First, since j and k communicate, there exist M, N such that $p_{kj}(N) =: \alpha > 0$ ja $p_{jk}(M) =: \beta > 0$. By using Chapman-Kolmogorov equation, we have

$$\begin{aligned} p_{kk}(N+M+n) &= \sum_l \sum_m p_{kl}(N) p_{lm}(n) p_{mk}(M) \\ &\geq p_{kj}(N) p_{jj}(n) p_{jk}(M) \\ &= \alpha \cdot \beta \cdot p_{jj}(n). \end{aligned}$$

To show k is recurrent, we have to show that $P_k = \sum_{n=1}^{\infty} p_{kk}(n) = \infty$. Using the inequality above, we have

$$P_k \geq \sum_{n=1}^{\infty} p_{kk}(N+M+n) \geq \alpha\beta \sum_{n=1}^{\infty} p_{jj}(n) = \infty,$$

because j is recurrent. Thus k is also recurrent. The proof is completed.

Time to return

Let j be a recurrent state of a Markov chain. Then, starting from j, the process will reenter j at some later period with probability 1. Or, equivalently, if T_j denotes the number of steps necessary for returning to j, then $P(T_j < \infty) = 1$. However, it does not follow that the expected time ET_j is also finite.

Definition 10. *A state j is said to be **positive recurrent** if, starting in j, the expected time until the process returns to state j is finite, $ET_j < \infty$. In case of $ET_j = \infty$ the state j is called **null recurrent**.*

It is possible to show that positive recurrence is a class property. It can also be shown that in a finite-state Markov chain all recurrent states are positive recurrent. However, if MC has infinitely many states, then null-recurrence is an option. For example, all states of symmetric random walk on the real line (i.e. integers) are null recurrent - it takes (in average) infinitely much time to wait

until the process returns to the initial state. (A bit strange - isn't it? But try to show this - see the exercise below.)

Exercise*. Show that all states of symmetric random walk on the real line are null recurrent.
Hint: First show that for each state j it holds $P(T_j = 2n) = \frac{\frac{1}{2n-1}\binom{2n}{n}}{2^{2n}}$. Then show $ET_j = \infty$.

8.5 The gambler's ruin problem

This is an application of Markov chains to solve a classical problem. Consider a gambler who at each play of the game has probability p of winning one unit and probability $q = 1 - p$ of losing one unit. Assuming that successive plays of the game are independent, what is the probability that, starting with i units, the gambler's fortune will reach N before reaching 0? If we let X_n denote the player's fortune at time n, then the process $X_n, n = 0, 1, 2, ...$ is a Markov chain with absorbing states 0 and N that we have studied before.

Denote the unknown probability
$\Pi_i = P\{$ starting from i one attains N before $0\}$.

Note that $\Pi_0 = 0$. In order to find Π_i we condition on the outcome of the initial play of the game or, in other words, apply the formula of total probability $P(A) = \sum P(B_i)P(A|B_i)$, wher $A =$ 'N is attained before the ruin', $B_1 =$ 'win in the first play' and $B_2 =$ 'loss in the first play'. Since $P(B_1) = p$ and after the win in the first play the fortune is $i + 1$, we obtain

$$\Pi_i = p \cdot \Pi_{i+1} + q \cdot \Pi_{i-1}, \quad i = 1, \ldots, N-1.$$

Since $p + q = 1$, it follows that $q\Pi_i + p\Pi_i = p\Pi_{i+1} + q\Pi_{i-1}$ or

$$\Pi_{i+1} - \Pi_i = \frac{q}{p}(\Pi_i - \Pi_{i-1}), \quad i = 1, 2, \ldots, N-1.$$

The quantity $\Pi_{i+1} - \Pi_i$ is called the *difference* and the relationship obtained the

difference equation. Repeated application of the equation gives:

$$\begin{aligned}
&\Pi_2 - \Pi_1 = \tfrac{q}{p}(\Pi_1 - \Pi_0) = \tfrac{q}{p}\Pi_1 \\
&\Pi_3 - \Pi_2 = \tfrac{q}{p}(\Pi_2 - \Pi_1) = \left(\tfrac{q}{p}\right)^2 \Pi_1 \\
&\vdots \\
&\Pi_i - \Pi_{i-1} = \tfrac{q}{p}(\Pi_{i-1} - \Pi_{i-2}) = \left(\tfrac{q}{p}\right)^{i-1} \Pi_1 \\
&\vdots \\
&\Pi_N - \Pi_{N-1} = \tfrac{q}{p}(\Pi_{N-1} - \Pi_{N-2}) = \left(\tfrac{q}{p}\right)^{N-1} \Pi_1.
\end{aligned}$$

Adding up all the rows and taking into account that $\Pi_N = 1$, we obtain

$$1 - \Pi_1 = \left[\frac{q}{p} + \ldots + \left(\frac{q}{p}\right)^{N-1}\right] \Pi_1,$$

which gives

$$\Pi_1 = \left[1 + \frac{q}{p} + \ldots + \left(\frac{q}{p}\right)^{N-1}\right]^{-1}.$$

Since for $x < 1$ the sum of geometric progression $1 + x + x^2 + \ldots = \frac{1}{1-x}$, we have

$$\Pi_1 = \begin{cases} \frac{1-\frac{q}{p}}{1-\left(\frac{q}{p}\right)^N} & , \text{ if } q \neq p, \\ \frac{1}{N} & , \text{ if } q = p. \end{cases} \tag{20}$$

To obtain Π_i we add up first $i - 1$ lines which gives

$$\Pi_i = \left[1 + \frac{q}{p} + \ldots + \left(\frac{q}{p}\right)^{i-1}\right] \Pi_1.$$

By substituting Π_i from (20), we obtain the final result

$$\Pi_i = \begin{cases} \frac{1-\left(\frac{q}{p}\right)^i}{1-\left(\frac{q}{p}\right)^N} & , \text{ if } q \neq p, \\ \frac{i}{N} & , \text{ if } q = p = \frac{1}{2}. \end{cases} \tag{21}$$

Hence, for symmetric random walk the probability of attaining the upper barrier N is proportional to the initial capital i.

8.6 Limiting probabilities (Ergodic theorem)

Consider the situation when the MC has been working for a long time already, i.e. the number of steps n is big. It turns out that under certain circumstances the

state probabilities $P\{X_n = j\}$ (and also transition probabilities $p_{ij}(n)$) stabilize i.e. become independent both of n and i.

Denote the probability

$$P\{X_n = j\} =: p_j(n).$$

Then, by the formula of total probability

$$p_j(n+1) = \sum_i p_i(n) \cdot p_{ij}(1). \tag{22}$$

It is more convenient to write the same in matrix notation. Denote the vector $\pi_n = (p_1(n), p_2(n), \ldots)$ which represents the probability distribution of different states at time n. Then the formula (22) writes as

$$\pi_{n+1} = \pi_n \mathbf{P}, \tag{23}$$

where $\mathbf{P}$ is the transition matrix $\mathbf{P} = (p_{ij})$. Repeated application of the formula gives $\pi_{n+1} = \pi_n \mathbf{P} = \pi_{n-1}\mathbf{P}^2 = \ldots = \pi_0 \mathbf{P}^{n+1}$, where $\pi_0 = (p_1(0), p_2(0), \ldots)$ is initial distribution of states. It turns out that (under certain conditions) in a long run the distribution π_n stabilizes, i.e. for large n one has $\pi_{n+1} \approx \pi_n$.

Example 19. *(Weather forecast) Once more, consider the simple weather model with two states (*0 *– it rains,* 1 *– it does not rain), with transition matrix*

$$\boldsymbol{P} = \begin{pmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{pmatrix}.$$

By choosing $\alpha = 0.7$ *and* $\beta = 0.4$ *we have*

$$\boldsymbol{P}^2 = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix} \cdot \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{pmatrix},$$

$$\boldsymbol{P}^4 = \boldsymbol{P}^2 \cdot \boldsymbol{P}^2 = \begin{pmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{pmatrix},$$

$$\boldsymbol{P}^9 = \begin{pmatrix} 0.571\ldots & 0.429\ldots \\ 0.571\ldots & 0.429\ldots \end{pmatrix},$$

i.e. as n *increases, the rows of transition matrix become equal.*

Also the probabilities of states 0 and 1 stabilize, and the limits do not depend on the initial distribution π_0. Indeed, if we take e.g. $\pi_0 = (0.2, 0.8)$, we obtain, after 9 steps, that $\pi_9 = \pi_0 \cdot \boldsymbol{P}^9 = (0.571, 0.429)$. It is east to see that any other – initial distribution gives the same result – 57% of days are rainy and 43% are clear. Thus, this Markov chain 'forgets' its past.

We now explain under which conditions the limiting probabilities above exist. Let us define a characteristic of a MC, by considering its n-step transition matrix $\mathbf{P}^n$.

Definition 11. *The quantity*

$$K(n_0) = 1 - \frac{1}{2} \sup_{i,j} \sum_m |p_{im}(n_0) - p_{jm}(n_0)|)$$

is called the coefficient of ergodicity.

Example 20. *This is an example where $K(n_0) = 0$. Consider a periodic MC with two states (1 ja 2) such that the states alternates at each step. Then the transition matrix is*

$$\boldsymbol{P}(2n) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

ja

$$\boldsymbol{P}(2n+1) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

We see that the probabilities do not converge when $n \to \infty$. Calculate the coefficient of ergodicity: for each n_0 we have $K(n_0) = 1 - \frac{1}{2}(|1-0| + |0-1|) = 0$.

As it will be seen from the theorem below, it is important to have $K(n_0) > 0$, in order to have the probabilities $p_j(n)$ convergent. This condition is fulfilled, if e.g. all elements of a column m of the transition matrix $\mathbf{P}^{n_0}$ exceed some $\delta > 0$. This condition is fulfilled in our weather example.

Theorem 3 (Ergodic theorem)**.** *If $K(n_0) > 0$, then the following convergence takes place*

$$\lim_{n\to\infty} p_{ij}(n) = \lim_{n\to\infty} p_j(n) = p_j^*,$$

where the limit probabilities p_j^* can be calculated from the following system of equations:

$$p_j^* = \sum_i p_i^* \cdot p_{ij}, \quad j = 0, 1, \ldots \tag{24}$$

HW: What is the percentage of rainy days (clear days) in a long run? Or, in other words, what is the probability that it rains (it is clear) without having any information about the weather in the past. Let the transition matrix be as before (0 = it rains, 1=it does not rain):

$$\mathbf{P} = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

(Hint: Build and solve the system of equations given in the Ergodic Theorem. Take into account that $p_1^* + p_2^* = 1$.)

HW. Consider the random walk with reflecting states 0 and N (meaning $p_{01} = 1, p_{N,N-1} = 1$). Show that all states $0, 1, \ldots, N$ are essential and communicating with each other (such a MC is called irreducible - see the definition of irreducible MC above).

9 Poisson processes

In this chapter we make acquaintance with the simplest class of random processes in continuous time - Poisson processes. Such processes appear when we account occurrences of an event A in time.

9.1 Definition of the Poisson process

Consider an event A that occurs at random time points - for example, claims arrivals in insurance, clients entering a department store. Let $N(t)$ be the number of events A which have taken place in the time interval $[0, t]$. $N(t)$ is called a counting process.

Definition 12. *A counting process* $\{N(t),\ t \geq 0\}$ *is called (homogeneous)* ***Poisson process****, if*

(1) $N(0) = 0$,

(2) *the increments* $N(t) - N(s)$ *are independent and stationary,*

(3) *the number of events occurred in a time interval of length* t *is a random variable having Poisson distribution with mean* λt*, i.e. for any* $s, t \geq 0$

$$P\{N(t+s) - N(s) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t},\ n = 0, 1, 2, \ldots.$$

The condition (3) implies that the average number of events in the time interval $[0, t]$ is proportional to its length: $EN(t) = \lambda t$. The number λ is called the **intensity** (or rate) of the Poisson process.

Two typical examples where the Poisson process is a suitable model are:

1. $N(t)$ is the number of traffic accidents in a town in time interval [0,t].
2. $N(t)$ is the number of calls arrived at a call center in time interval [0,t].

It is possible to show that the condition (3) in the definition can be replaced by:

(3a) $P\{N(t) = 1\} = \lambda t + o(t)$, if $t \to 0$,

(3b) $P\{N(t) \geq 2\} = o(t)$, if $t \to 0$.

The property (3a) says that the probability of exactly one event in a short interval is proportional to the length of the interval. The property (3b) says that the probability of occurrence of two or more events in a short interval is practically impossible (the events do not occur in groups). Note also that due to (1) and the stationarity of increments, the probability that an event occurs at a fixed time t_0 is equal to zero. We denote $P_k(t) = P(N(t) = k)$.

9.2 Waiting times distributions

Let $\{N(t), t \geq 0\}$ be a Poisson process with intensity λ. Denote the time of the first event by T_1. Further, for $n > 1$, let T_n denote the elapsed time between the $(n-1)$st and the nth event. The random variables $T_1, T_2, \ldots$ are called **waiting times** (or interarrival times).

Let us determine the distribution of the T_n. First note that the event $\{T_1 > t\}$ takes place if and only if no events occur in the interval $[0, t]$ and thus,

$$P\{T_1 > t\} = P\{N(t) = 0\} = P_0(t) = e^{-\lambda t}.$$

Hence, T_1 has an exponential distribution with parameter λ, or $T_1 \sim \mathcal{E}(\lambda)$.

To obtain the distribution of T_2, we first calculate its conditional distribution given that $T_1 = s$:

$$\begin{aligned} P\{T_2 > t | T_1 = s\} &= P\{0 \text{ events in } (s, s+t] \mid T_1 = s\} \\ &= P\{0 \text{ events in } (s, s+t]\} \\ &= P\{0 \text{ events in } (0, t]\} \\ &= P_0(t) = e^{-\lambda t}, \end{aligned}$$

where the second equality comes from the independence and third from stationarity. We see that conditional distribution of T_2 does not depend on T_1, hence T_2 and T_1 are independent. Thus

$$P\{T_2 > t\} = P\{T_2 > t \mid T_1 = s\} = e^{-\lambda t}$$

i.e. T_2 has the same exponential distribution, $T_2 \sim \mathcal{E}(\lambda)$. Repeating the same argument yields

Theorem 4. *The waiting times $T_1, T_2, \ldots$ of a homogeneous Poisson process are independent identically distributed exponential random variables, $T_i \sim \mathcal{E}(\lambda)$.*

The time of n-th event

Another quantity of interest is the **arrival time** of the n-th event which can be represented as

$$S_n = T_1 + T_2 + \ldots + T_n.$$

In order to find the distribution of S_n, note that the event $S_n > t$ is equivalent to the event $N(t) < n$, hence

$$P\{S_n > t\} = P\{N(t) < n\} = \sum_{k=0}^{n-1} P_n(t) = \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

We now obtain the distribution function

$$F_{S_n}(t) = P\{S_n \le t\} = 1 - P\{S_n > t\} = 1 - \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t},$$

and the differentiation gives

$$f_{S_n}(x) = \lambda \frac{(\lambda x)^{n-1}}{(n-1)!} e^{-\lambda x}, \quad x \ge 0$$

which is the density of so called *gamma-distribution* $\Gamma(n, \lambda)$. Thus we have proved next

Lemma 1. *In a Poisson process, the arrival time S_n of the n-th event is gamma distributed with parameters n and λ.*

Computer simulation of a Poisson process

In order to simulate a Poisson process, we only need to generate waiting times $T_1, T_2, \ldots, T_N$. Since the waiting times must be exponentially distributed, one can use plug-in generators available in most programming languages. However, one can also start from uniformly distributed r.v. $U_1, U_2, \ldots$ and use the transformation $T_i = -\ln U_i/\lambda$, which is exponentially distributed, $T_i \sim \mathcal{E}(\lambda)$.

HW21. Suppose that people enter a bank office in accordance with a Poisson process with rate $\lambda = 30$ per hour. The office opens at 9 o'clock. (a) What is the expected time until the tenth customer arrives?
(b) What is the probability that the elapsed time between the tenth and the eleventh arrival exceeds three minutes?
(c) What is the probability that exactly 5 clients arrive in the time interval 12:20-12:25 and then nobody arrives during the next 3 minutes?

9.3 Conditional distribution of the arrival times

Suppose we are told that exactly one event of a Poisson process has taken place by time t, and we are asked when could it happen, i.e. what is the distribution of the time at which the event occurred. Now, since a Poisson process possesses stationary and independent increments it seems reasonable that each interval in $[0, t]$ of equal length should have the same probability of containing the event. In other words, the time of the event should be uniformly distributed on [0, t]. This can be easily checked since, for $s \leq t$

$$\begin{aligned} P\{S_1 \leq s \mid N(t) = 1\} &= \frac{P\{S_1 \leq s, N(t) = 1\}}{P\{N(t) = 1\}} = \\ &= \frac{P\{1 \text{ event in } [0, s] \text{ and } 0 \text{ events in } (s, t]\}}{P\{N(t) = 1\}} = \\ &= \frac{P\{1 \text{ event in } [0, s]\} \cdot P\{0 \text{ events in } (s, t]\}}{P\{N(t) = 1\}} = \\ &= \frac{P_1(s) \cdot P_0(t - s)}{P_1(t)} = \frac{\lambda s e^{-\lambda s} \cdot e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} = \frac{s}{t}, \end{aligned}$$

which is the distribution function of uniform distribution on the interval $[0, t]$. The result can be generalized to the case where we know that n events have taken place in $[0, t]$.

Theorem 5. *Given that $N(t) = n$, the n arrival times $S_1, S_2, \ldots, S_n$ have the same distribution as n ordered, uniformly distributed random variables.*

Comment. Intuitively, if we follow arrival times of a Poisson process in a long time, they look as being uniformly distributed, without grouping.

9.4 Properties of Poisson processes

9.4.1 Decomposition

Consider a Poisson process $\{N(t),\ t \geq 0)\}$ having rate (intensity) λ, and suppose that each time the event occurs it is classified as either type I or type II event.

Suppose further that each event is classified as a type I event with probability p or a type II event with probability $1-p$, independently of all other events. For example, suppose that customers arrive at a store in accordance with a Poisson process having rate λ; and suppose that each arrival is male with probability $1/2$ and female with probability $1/2$. Then a type I event would correspond to a male arrival and a type II event to a female arrival.

Let $\{N_1(t),\ t \geq 0\}$ and $\{N_2(t),\ t \geq 0\}$ denote respectively the number of type I and type II events occuring in $[0,t]$. Note that $N(t) = N_1(t) + N_2(t)$. The processes $N_1(t)$ and $N_2(t)$ are called *partial processes.*

Theorem 6. *The partial processes $N_1(t)\}$ and $N_2(t)\}$ are both Poisson processes having respective rates λp and $\lambda(1-p)$.*

9.5 Sum of independent Poisson processes

Consider now an inverse problem: given that $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are Poisson processes, what can be said about the sum $N(t) = N_1(t) + N_2(t)$. Is it also a Poisson process?

For example, if $N_1(t)$ is the number of calls within the company in time interval $[0,t]$ and $N_2(t)$ the number of calls made outside the company in the same time interval, then $N(t)$ is the total number of calls made in $[0,t]$.

Thye main result here is the following

Theorem 7. *If $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are independent Poisson processes with respective intensities λ_1 and λ_2, then their sum $N(t) = N_1(t) + N_2(t)$ is a Poisson process with intensity $\lambda = \lambda_1 + \lambda_2$.*

Proof. It is easy to see that $N(t)$ satisfies two first conditions of the definition 12. The last condition (3) is also fulfilled since the sum of two independent Poisson distributed r.v.-s $N(t+s) - N(t) = N_1(t+s) - N_1(t) + N_2(t+s) - N_2(t)$ has

Poisson distribution again. ◀

Example. (Fires) Let $\{N_1(t), t \geq 0\}$ be the number of fires in towns and $\{N_2(t), t \geq 0\}$ the number of fires in the countryside. Suppose they are independent Poisson processes with rates 4 and 3 (per month). Then the total number of fires is a Poisson process with intensity 7.

9.6 Some properties of exponential distribution

9.6.1 Memoryless property

A random variable X is said to be without memory, or **memoryless**, if

$$P(X > s + t | X > t) = P(X > s) \text{ for all } s, t \geq 0. \tag{25}$$

If we think of X as being the lifetime of some instrument, then equation (25) states that the probability that the instrument lives for at least s + t hours given that it has survived t hours is the same as the initial probability that it lives for at least s hours. In other words, if the instrument is alive at time t, then the distribution of the remaining amount of time that it survives is the same as the original lifetime distribution; that is, the instrument does not remember that it has already been in use for a time t. The condition (25) is equivalent to

$$\frac{P(X > s + t, X > t)}{P(X > t)} = P(X > s)$$

or, since $X > s + t$ implies $X > t$,

$$P(X > s + t) = P(X > s) \cdot P(X > t). \tag{26}$$

It easy to see that the equation (26) is satisfied when X is exponentially distributed, i.e. when the survival function is $P(X > t) = e^{-\lambda t}$. (Indeed, $e^{-\lambda(s+t)} = e^{-\lambda s} \cdot e^{-\lambda t}$.) Hence, *the exponential distribution is memoryless.*

Example. Suppose that the amount of time one spends in a bank is exponentially distributed with mean 10 minutes, that is $\lambda = 1/10$. What is the probability that

a customer will spend more than fifteen minutes in the bank? What is the probability that a customer will spend more than fifteen minutes in the bank given that she is still in the bank after ten minutes?
Solution: If X represents the amount of time that the customer spends in the bank, then the first probability is just

$$P(X > 15) = e^{-15\lambda} = e^{-3/2} \approx 0.220$$

The second question asks for the probability that a customer who has spent ten minutes in the bank will have to spend at least five more minutes. However, since the exponential distribution does not "remember" that the customer has already spent ten minutes in the bank, this must equal the probability that an entering customer spends at least five minutes in the bank. That is, the desired probability is just

$$P(X > 15|X > 10) = P(X > 5) = e^{-5\lambda} = e^{-1/2} \approx 0.604$$

9.6.2 Which comes first?

Another useful calculation is to determine the probability that one exponentially distributed r.v. is smaller than another. That is, suppose that X_1 and X_2 are independent exponential random variables with parameters λ_1 and λ_2; what is $P(X_1 < X_2)$? This probability is easily calculated by conditioning on X_1:

$$\begin{aligned} P(X_1 < X_2) &= \int_0^\infty P(X_1 < X_2|X_1 = x)\lambda_1 e^{-\lambda_1 x} dx && (27)\\ &= \int_0^\infty P(x < X_2)\lambda_1 e^{-\lambda_1 x} dx && (28)\\ &= \int_0^\infty e^{-\lambda_2 x}\lambda_1 e^{-\lambda_1 x} dx && (29)\\ &= \int_0^\infty \lambda_1 e^{-(\lambda_1+\lambda_2)x} dx && (30)\\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2}. && (31) \end{aligned}$$

Now apply this result.

Example. (Fires continued). What is the probability that first fire will break out in the countryside?
Solution: Using the result above the answer is

$$\frac{3}{3+4} = \frac{3}{7}.$$

HW22. Suppose one has stereo system consisting of two main parts, a radio and a speaker. If the lifetime of the radio is exponential with mean 1000 hours and the lifetime of the speaker is exponential with mean 500 hours independent of radio's life-time, then what is the probability that the system's failure (when it occurs) will be caused by the radio failing?

HW23. ('Memoryless' property of exponential distribution) The lifetime of a radio is exponentially distributed with a mean of ten years. Paul buys a new radio and Peter buys a ten-year-old radio. For both Paul and Peter, find the probability that his radio will be working after an additional ten years? Comment the result!

HW24. A shop is open from 8am to 20pm. Customers arrive in a shop according to Poisson process with rate $\lambda = 2$ per hour.
(a) What is the probability that no customer arrives between 9am and 10am?
(b) Starting at noon, what is the expected time at which the fourth customer arrives?
(c) What is the probability that two or more customers arrive between 6pm and 8pm?

9.7 Generalizations of the Poisson process

Consider two important generalizations of the Poisson process.

9.7.1 Nonhomogeneous Poisson process

Definition 13. *The counting process* $\{N(t), t \geq 0\}$ *is called* ***nonhomogeneous*** *Poisson process with intensity function* $\lambda(t),\ t \geq 0$*, if*

(1) $N(0) = 0$*,*

(2) $\{N(t), t \geq 0\}$ *increments are independent,*

(3) $P\{N(t+h) - N(t) = 1\} = \lambda(t)h + o(h),\ h \to 0\ \ \forall\ t$ *,*

(4) $P\{N(t+h) - N(t) \geq 2\} = o(h),\ h \to 0\ \ \forall\ t$ *.*

Now an important role is played by the function $m(t) = \int_0^t \lambda(s)ds$. Namely, it can be shown that the increment of $N(t)$ in time interval $[t, t+s)$ has Poisson distribution with mean $m(t+s) - m(t)$, i.e. for all $n \geq 0$

$$P\{N(t+s) - N(t) = n\} = \frac{(m(t+s) - m(t))^n}{n!} e^{-m(t+s)+m(t)}. \tag{32}$$

In particular, $N(t) \sim \mathcal{P}(m(t))$, hence $EN(t) = m(t)$ and therefore the function $m(t)$ is called the **mean value function** of the process. Note also that in case of homogeneous Poisson process we have $\lambda(t) \equiv \lambda$ and $m(t) = \lambda t$.

9.7.2 Compound Poisson process

Definition 14. *A random process* $\{X(t),\ t \geq 0\}$ *is called compound Poisson process, if it can be represented as* [3]

$$X(t) = \sum_{i=1}^{N(t)} Y_i, \quad t > 0,$$

where $\{N(t),\ t \geq 0\}$ *is a Poisson process and* $Y_1, Y_2, \ldots$ *are iid r.v. which are also independent of* $\{N(t), t \geq 0\}$*.*

[3] It is agreed that $\sum_{i=1}^{0} Y_i = 0$.

Note that if $Y_i = 1,\ i = 1, 2, \ldots$, then $X(t) = N(t)$ i.e. the compound Poisson process reduces to usual Poisson process $N(t)$.

Example. Assume that the customers leave a supermarket in accordance with a Poisson process. If Y_i, the amount spent by the customer i, are independent and identically distributed r.v., then the total amount of money $X(t)$ collected by time t is a compound Poisson process.

Let us calculate the expected value and the variance of $X(t)$. We use conditioning w.r.t. $N(t)$: 1) we first calculate conditional expectation $E[X(t)|N(t)]$ and then 2) we find weighted average (over $N(t)$) of the results obtained:

$$EX(t) = E\{E[X(t) \mid N(t)]\}.$$

Denote $EY_i = \mu$ and calculate

$$E[X(t) \mid N(t) = n] = E\left[\sum_{i=1}^{N(t)} Y_i \mid N(t) = n\right] =$$

$$= E\left[\sum_{i=n}^{n} Y_i \mid N(t) = n\right] = \sum_{i=1}^{n} E[Y_i \mid N(t) = n] =$$

$$= \sum_{i=1}^{n} EY_i = n \cdot \mu,$$

where we have used independence of Y_i and $\{N(t), t \geq 0\}$. Hence

$$E[X(t) \mid N(t)] = N(t) \cdot \mu,$$

and averaging over $N(t)$ gives a natural result

$$EX(t) = \lambda t \mu, \tag{33}$$

However, we do not get such a simple formula for the variance. The starting point here is

$$DX(t) = EX^2(t) - [EX(t)]^2 = EX^2(t) - \lambda^2 t^2 \mu^2,$$

and it remains to find

$$EX^2(t) = E\{E[X^2(t) \mid N(t)]\}.$$

First calculate

$$E[X^2(t) \mid N(t) = n] = E\left[\left(\sum_{i=1}^{N(t)} Y_i\right)^2 \mid N(t) = n\right] =$$

$$= E\left[\left(\sum_{i=1}^{n} Y_i\right)^2 \mid N(t) = n\right] = E\left(\sum_{i=1}^{n} Y_i\right)^2 =$$

$$= E\sum_{i=1}^{n} Y_i^2 + E\left(\sum_{i\neq j} Y_i Y_j\right) = n \cdot EY_1^2 + \sum_{i\neq j} EY_i EY_j =$$

$$= n \cdot EY_1^2 + n(n-1)\mu^2.$$

thus

$$E[X^2(t) \mid N(t)] = N(t) \cdot EY_1^2 + N(t)(N(t)-1)\mu^2,$$

and averaging over $N(t)$ gives

$$EX^2(t) = EN(t) \cdot EY_1^2 + E[N(t)(N(t)-1)]\ \mu^2.$$

By taking into account that $EN(t) = \lambda t$ ja $EN^2(t) = DN(t) + [EN(t)]^2 = \lambda t + \lambda^2 t^2$, we finally obtain

$$DX(t) = \lambda t E{Y_1}^2. \tag{34}$$

We see that the variance of the compound Poisson process is larger than average sum of individual variances equal to

$$\lambda t DY_1 = \lambda t[EY_1^2 - \mu^2].$$

Example. (Total claim process in insurance). Let $\{N(t),\ t \geq 0\}$, be the number of claims arrived on an insurance company in time interval $[0, t]$. Let $Z_i,\ i = 1, 2, \ldots$ be sizes of the claims. Assuming that $N(t)$ is a Poisson process and that the claim sizes are i.i.d. and also independent of $N(t)$, we see that the

total claim size by time t, the quantity $X(t) = \sum_{i=1}^{N(t)} Z_i$ is a compound Poisson process.

HW25. An insurance company receives in average 8 claims per day. The claim sizes are log-normally distributed with mean 780 EUR and standard deviation 420 EUR. Assume that the claim arrival process is a homogeneous Poisson process independent of claim sizes.
1) Find the expected value and the standard deviation of the daily total claim X.
2) Using Markov inequality and then its extension (by choosing $f(x) = x^2$), estimate the probability that the daily total claim exceeds 15000 EUR. Which result among these two is better?
3) Will the answers be different if the log-normal claim distribution is replaced by e.g. Pareto or Weibull distribution?

10 Brownian motion

10.1 Definition of Brownian motion.

Brownian motion is a simple continuous stochastic process that is widely used in physics and finance for modeling random behavior that evolves over time. Examples of such behavior are the random movements of a molecule of gas or fluctuations in an asset's price.

Brownian motion gets its name from the botanist Robert Brown (1828) who observed in 1827 that tiny particles of pollen suspended in water moved erratically on a microscopic scale; but he was not able to determine the mechanisms that caused this motion. The physicist Albert Einstein published a paper in 1905 explaining that the motion was caused by water molecules randomly bombarding the particle of pollen, and thus helping to firmly establish the atomic theory of matter. Later on, starting in 1918, American mathematician Norbert Wiener created a precise mathematical model for this phenomena. This is what we will study now.

In order to give precise definition of BM, we start from Simple Random Walk (SRW) $S_n = X_1 + X_2 + \ldots + X_n$, where the steps X_i are two-valued random variables

$$X_i = \begin{cases} +1, & \text{with probability } \frac{1}{2}, \\ -1, & \text{with probability } \frac{1}{2}, \end{cases}$$

Note that SRW has unit steps both in time ($n = 0, 1, 2, ..$) and in space (± 1). However, in many other areas (including finance), there are also processes with continuous time, with $t \in \mathbf{R}^+ = [0, \infty)$. For example, stock prices can change at any time instant within a business day. At the same time, the changes in the price can be very small. Therefore, we make both time step and space step of SRW shorter and shorter, by letting them to go to 0. Denote

$$S_t = \Delta x \cdot (X_1 + X_2 + \ldots + X_{[\frac{t}{\Delta t}]}),$$

where

$$X_i = \begin{cases} +1, & \text{with probability } \frac{1}{2}, \\ -1, & \text{with probability } \frac{1}{2}, \end{cases}$$

Δx - space step size,

Δt - time step size,

$[\frac{t}{\Delta t}\ ]$ - the number of time steps in the time interval $[0, t]$.

Since $X_1, X_2, \ldots, X_{[\frac{t}{\Delta t}]}$ are independent identically distributed (IID) random variables with mean value $EX_i = 0$ and variance $DX_i = 1$, then for each Δx and Δt we have $ES_t = 0$ and $DS_t = (\Delta x)^2[\frac{t}{\Delta t}]$. When $\Delta t \to 0, \Delta x \to 0$, then the number of summands $[\frac{t}{\Delta t}]$ increases unboundedly and according to the Central Limit Theorem

$$\frac{S_t}{\Delta x\sqrt{[\frac{t}{\Delta t}]}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

By choosing the relationship between Δx and Δt such that $\frac{(\Delta x)^2}{\Delta t} = \text{ const } =: C^2$ we get, on the limit, that $S_t \sim \mathcal{N}(0, C\sqrt{t})$. The limiting process S_t preserves some important features of SRW:

(i) The increments of S_t are independent i.e. for $0 \le s \le t \le u \le v$ the increments $S_t - S_s$ and $S_v - S_u$ are independent r.v. (the same is valid for any n time intervals).

(ii) The increments of S_t are stationary i.e. the distribution of $S_{s+t} - S_s$ only depends on t (and not on s).

These properties suggest the following definition.

Definition 15. *The random process* $\{W_t,\ t \ge 0\}$ *is called* ***Brownian motion*** *(Wiener process), if*

(i) $W(0) = 0$,

(ii) for all $t > 0$ *the r.v.* $W_t \sim \mathcal{N}(0, C\sqrt{t})$, *where* $C > 0$ *is a constant,*

(iii) increments of W_t are independent and stationary,

(iv) the paths of W_t are a.s. continuous (in t).

From the definition of BM it follows that also the increments of BM are normally distributed: by the stationarity of increments

$$W_t - W_s \stackrel{\mathcal{D}}{=} W_{t-s} - W_0 = W_{t-s} \sim \mathcal{N}(0, C\sqrt{t-s},) \tag{35}$$

where $\stackrel{\mathcal{D}}{=}$ is to be read as "has same distribution as" .

Note that, in fact, the property (iv) can be deduced from properties (i)-(iii).

BM is a mathematical model widely used in physics (diffusions), economics (price models) e.t.c. .
If $C = 1$, the BM is called **Standard Brownian Motion** (SBM). The process

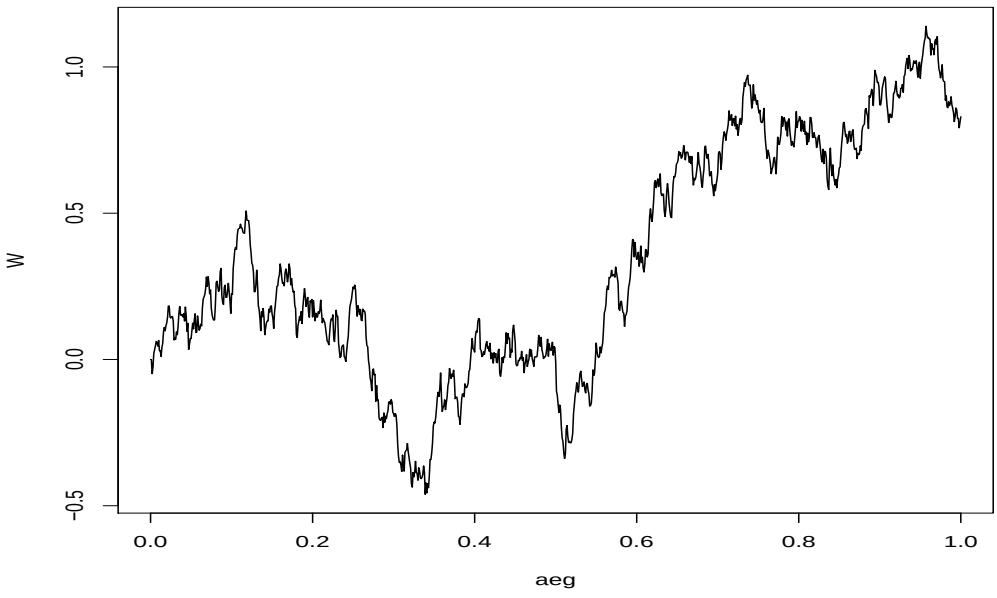

Figure 2: A trajectory of standard Brownian motion

$W_t + \mu t$, where μ is a real number, is called Brownian motion with *drift*(μ is called *drift coefficient*). The mean value of BM with positive (resp negative) drift increases (resp decreases) in time.

Exercise 1. *Let W_t be a standard Brownian motion.*
1) Find the probability that $W_5 < 0$.
2) Assume that $W_1 = 2$. Find now the (conditional) probability that $W_5 < 0$.

Solution. 1) Without any given condition we would have $P\{W_5 < 0\} = \frac{1}{2}$, since $W_t \sim N(0, \sqrt{t})$ and since the normal distribution is symmetric w.r.t its mean value (which is 0 here).

2) However, under the condition $W_1 = 2$ we make use of the stationarity of increments (see the formula (35)): the increment $W_5 - W_1 \stackrel{\mathcal{D}}{=} W_{5-1} - W_0 = W_4 \sim N(0, \sqrt{4})$ – this distribution only depends on the length of the time interval and not on its location. Therefore we have

$$P\{W_5 < 0 | W_1 = 2\} = P\{W_5 - W_1 < -2 | W_1 = 2\} =$$

$$= P\{W_5 - W_1 < -2\} = P\{W_4 < -2\} =$$

$$= P\{N(0, \sqrt{4}) < -2\} = \Phi\left(\frac{-2}{2}\right) = \Phi(-1) = 0.16 \ .$$

Note that $\Phi(t)$ is the distribution function of the standard normal distribution $N(0, 1)$:

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \, e^{-\frac{x^2}{2}} dx.$$

HW26. Let W_t be a standard Brownian motion.
1) Find the probability that $W_3 < 1$.
2) Assume that $W_2 = 3$. Find the conditional probability that $W_8 > 7$.
3) Find $E(W_3 - W_1)(W_9 - W_5)$.
4) Show that $E(W_5 - W_1)(W_9 - W_2) = 3$. (Present the increments as sums of two increments, second taken over the joint time interval).

10.2 Some properties of Brownian motion

1) Finite-dimensional distributions of SBM
The joint distribution of $(W_{t_1}, W_{t_2}, \ldots, W_{t_n})$ where $0 < t_1 < t_2 < \ldots < t_n$ can easily be calculated.

For each t_i the density of W_{t_i} is

$$f_{W_{t_i}}(x) = \frac{1}{\sqrt{2\pi t_i}}\ e^{-\frac{x^2}{2t_i}},$$

provided that $C = 1$. Since the equalities

$$\begin{cases} W_{t_1} = x_1 \\ W_{t_2} = x_2 \\ \dots \\ W_{t_n} = x_n \end{cases}$$

are equivalent to the equalities

$$\begin{cases} W_{t_1} = x_1 \\ W_{t_2} - W_{t_1} = x_2 - x_1 \\ \dots \\ W_{t_n} - W_{t_{n-1}} = x_n - x_{n-1} \end{cases}$$

and since the increments $W_{t_1}, W_{t_2} - W_{t_1}, \dots, W_{t_n} - W_{t_{n-1}}$ are independent, we have

$$\begin{aligned} & f_{W_{t_1},\dots,W_{t_n}}(x_1, \dots, x_n) = \\ & = f_{W_{t_1},W_{t_2}-W_{t_1},\dots,W_{t_n}-W_{t_{n-1}}}(x_1, x_2 - x_1, \dots, x_n - x_{n-1}) = \\ & = f_{W_{t_1}}(x_1) \cdot f_{W_{t_2}-W_{t_1}}(x_2 - x_1) \cdot \dots \cdot f_{W_{t_n}-W_{t_{n-1}}}(x_n - x_{n-1}) = \\ & = \frac{1}{\sqrt{2\pi t_1}} e^{-\frac{{x_1}^2}{2t_1}} \cdot \frac{1}{\sqrt{2\pi(t_2 - t_1)}} e^{-\frac{(x_2-x_1)^2}{2(t_2-t_1)}} \cdot \dots \cdot \frac{1}{\sqrt{2\pi t_n - t_{n-1}}} e^{-\frac{(x_n-x_{n-1})^2}{2(t_n-t_{n-1})}}. \end{aligned}$$

The formula obtained can be used for many purposes.

2) Conditional distribution

Let's use the formula above to solve one particular problem. Suppose we know that at time t BM has taken value $W_t = B$. Let s be an earlier time, $s < t$. What is the conditional distribution of W_s given the event $W_t = B$? It is known that the conditional density is the ratio of joint density and the density of the condition, we can calculate

$$f_{W_s|W_t}(x|B) = \frac{f_{W_s,W_t}(x,B)}{f_{W_t}(B)} =$$

$$= \frac{\frac{1}{\sqrt{2\pi s}} \cdot e^{-\frac{x^2}{2s}} \cdot \frac{1}{\sqrt{2\pi(t-s)}} \cdot e^{-\frac{(B-x)^2}{2(t-s)}}}{\frac{1}{\sqrt{2\pi t}} \cdot e^{-\frac{B^2}{2t}}} = \ldots = \frac{1}{\sqrt{2\pi \frac{s}{t}(t-s)}} \cdot e^{-\frac{t(x-B\frac{s}{t})^2}{2s(t-s)}}.$$

Hence, the conditional distribution of W_s is normal distribution with mean $B \cdot \frac{s}{t}$ and variance $\frac{s(t-s)}{t}$.

3) First passage time

Let $W_0 = 0$ and $a > 0$. We are interested in the time which elapses before BM attains the level a. We call it the *first passage time* to the point a and denote $T_a = \inf\{t : W_t = a\}$. T_a is a random variable since its value depends on the path of BM (on ω). Let's find its distribution function $F_{T_a}(t) = P\{T_a \leq t\}$. Using the formula of total probability, we have:

$$P\{W_t \geq a\} = P\{W_t \geq a | T_a \leq t\} \cdot P\{T_a \leq t\} + P\{W_t \geq a | T_a > t\} \cdot P\{T_a > t\}.$$

For symmetry of normal distribution $P\{W_t \geq a | T_a \leq t\} = \frac{1}{2}$. At the same time obviously $P\{W_t \geq a | T_a > t\} = 0$. Therefore $P\{W_t \geq a\} = \frac{1}{2} P\{T_a \leq t\}$, and

$$P\{T_a \leq t\} = 2P\{W_t \geq a\}.$$

Since $W_t \sim N(0, \sqrt{t})$, we have

$$F_{T_a}(t) = P\{T_a \leq t\} = 2\frac{1}{\sqrt{2\pi t}} \int_a^\infty e^{-\frac{x^2}{2t}} dx = 2[1 - \Phi\left(\frac{a}{\sqrt{t}}\right)].$$

If $a < 0$, then by symmetry $P\{T_a \leq a\} = 2[1 - \Phi\left(\frac{|a|}{\sqrt{t}}\right)]$.
If $a = 0$, then $T_0 = 0$. Taking all together, we have that, for any a,

$$P\{T_a \leq t\} = 2[1 - \Phi\left(\frac{|a|}{\sqrt{t}}\right)]. \tag{36}$$

By differentiating the distribution function above, one gets the density function of T_a. For $a > 0$ it calculates as

$$f_{T_A}(t) = F'_{T_a}(t) = -2\varphi(\frac{a}{\sqrt{t}}) \cdot (-\frac{a}{2}) \cdot t^{-\frac{3}{2}} = \frac{a}{\sqrt{2\pi t^3}}\, e^{-a^2/2t}.$$

This distribution is called *inverse Gaussian distribution* (also Wald distribution). From (36) it is also seen that if $t \to \infty$, then

$$\mathbf{P}(T_a < \infty) = \lim_{t\to\infty} \mathbf{P}(T_a \le t) = 1. \tag{37}$$

4) Maxima of Brownian motion

If $a > 0$, then $P\{\max_{0\le s\le t} W_s \ge a\} = P\{T_a \le t\} = 2[1 - \Phi\left(\frac{|a|}{\sqrt{t}}\right)]$.

If $a < 0$, then $P\{\max_{0\le s\le t} W_s \ge a\} = 1$.

5) Brownian motion between two boundaries

Let $A > 0, B > 0$. Let us find the probability that, starting from 0, BM reaches level A before $-B$. Recall that in the case of SRW the answer to the same question is $\frac{B}{A+B}$. As the same answer remains true for any time and space steps sizes, we have

$$P\{W_t \text{ reaches } A \text{ before } -B\} = P\{T_A < T_B\} = \frac{B}{A+B}.$$

HW27. It is often believed that if X_t represents the market value of a given stock at time t then the $\log X_t$ is approximately a Brownian motion process. If this is so and if the current price of the stock is 10 euros, then what is the probability that the stock will reach the price 13 eur before price 8 eur ?

HW28. Let W_t be a standard Brownian motion. What is the distribution of $W_s + W_t,\ s \le t$? (Hint: Recall that the sum of two independent normally distributed random variables is again normally distributed.)

HW 29. Let T_α be the time it takes a standard Brownian motion to hit α. Find the probability $P(T_1 < T_{-1} < T_2]$.