Probability theory as a tool for financial mathematics

V.N. Kolokol'tsov
Department of Statistics, University of Warwick,
and Institute of Problems of Informatics (IPI) RAN
email: v.kolokoltsov@warwick.ac.uk

Preface

These Notes, revised and extended, form the basis for my book "Mathematical Analysis of Financial Markets. Probability Theory and Financial Mathematics," Moscow 2010, ISBN 978-5-904073-03-9.

The aim is to give a concise introduction to the basic notions of the elementary probability theory oriented to the applications in quantitative and qualitative analysis of financial processes. The main objective is to give the readers a clear idea about the mathematical structures (and their applications in model situations) lying at the foundation of the modern theories of the financial market. This includes the theory of option pricing, hedging by futures, portfolio optimization, credit risk and basic models of financial dynamics on the stock exchanges. The book can form a basis of undergraduate University courses in applied mathematics, economics, finances and statistics. Probability theory is introduced in the minimal scope (say, multidimensional distributions and characteristic functions are not discussed seriously), and all notions are illustrated by concrete examples (which distinguishes the book from the most of the mathematical texts). At the same time, the exposition is given on a mathematically rigorous level (unlike the most texts in finances and economics), because in the author's opinion the precision of the exposition is crucial for the correct understanding and hence correct application of the underlying theory.

To read the text one is expected to have an elementary knowledge of the basic rules of calculus (differentiation - integration, finding maxima and minima of functions, natural logarithms and exponential functions), to understand the basic operations on sets (union, intersection), to have at least a rudimentary idea about limits, infinite sums and continuity. Elementary knowledge of probabilities would be an advantage for reading, but is not a requirement.

Standard abbreviations and notations

r.h.s. right hand side
l.h.s. left hand side
r.v. random variable
i.i.d. independent identically distributed
$\emptyset$ -empty set
$\mathbf{N}$, $\mathbf{Z}$, $\mathbf{R}$ -the sets of natural, integer and real numbers,
$\mathbf{Z}_+ = \{0, 1, 2, ...\}$
$\exp\{x\} = e^x$
$\sum_{i=1}^n$ - sum over all i from 1 to n
$\prod_{i=1}^n$ - product over all i from 1 to n

$\circ$- composition, i.e. $(f \circ g)(x) = f(g(x))$
$f(x) \sim g(x)$, $x \to \infty$, means $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$
E - expectation, **P** - probability
$\Phi(x)$- distribution function for standard normal r.v.

Contents

Chapter 1

Elementary probability models

The readers having some acquaintance with the elementary notions of the probability theory can start by trying to solve the following curious problems, whose solution may look rather counterintuitive. For the solutions of the first three problems no knowledge is required, apart from the understanding of what is the meaning of probability. Problem 4 is an exercise on the conditional probabilities (Bayes rule) and Problem 5 can be most easily resolved via Poisson approximation. We shall supply the solutions at the end of this chapter after properly introducing all necessary tools.

Problem 1 (**a birthday problem**). In a football team of 23 players, what is the probability that at least two players have birthdays on the same day of a year?

Problem 2 (**daughter or sun**). Suppose you are told that a family has two children and that one of them is a daughter. (i) What is the probability of another child to be a daughter? (ii) Would this probability change, if you ring the bell and a daughter opens the door?

Problem 3 (**three doors quiz show**). A show host leads you to three closed doors. Behind one of them is a new car, and behind the other two are chocolate bars. The game proceeds as follows. You are choosing a door without opening it. Then the host will open one of the remaining doors that hides a chocolate bar (the host knows where is the car). When this is done, you are given the opportunity to switch door you have chosen to another closed door left. You will win whatever is behind the door you choose on this stage. Should you switch? In other words, would it increase your chances of winning? This problem is sometimes called the Monty Hall dilemma after the USA 1970s television game show, where this chance of winning was in fact offered. The solution was a subject of hot discussion in US press of that time. To make a solution more transparent a modification can be suggested. Suppose there are 100 doors (and still one car). You choose one door, and then 98 empty doors (with a chocolate) are revealed. Would you switch your first choice now?

Problem 4 (**test paradox**). Assume an inexpensive diagnostic test is suggested that is however not 100% reliable. Among infected persons the test shows positive in approximately 99% of the cases. For a non infected person there is a 2% chance to get a false positive result. (i) Suppose the test is used in a clinic with a half of all patients being infected. Show that for a given person the probability of being infected after testing positive is approximately 0.98. This probability being closed to one clearly supports the suggestion to use the test. (ii) Based on the above success, it is now suggested to test the entire population (say, yearly). It is known that one out of 1000 persons is infected. Let us put forward the same question. For a given person, what is the probability of being infected after testing positive? The remarkable answer is 0.047. This shows the necessity to monitor the basic proportions. Otherwise, one can arrive at the famous conclusion: 'Statistics show that 10% of traffic accidents are caused by drunken drivers, which means that other 90% are caused by sober drivers. Is it not sensible to allow only drunken drivers onto the road?

Problem 5 (**scratch-and-win lottery**). In a popular European lotteries N tickets are issued carrying two numbers, an open one and a hidden one. Each of these numbers is chosen randomly from the first N integers ($N = 10 \times 10^3$ in Andorra and $N = 10 \times 10^6$ in Spain) but in such a way that no two tickets can have identical open or identical hidden numbers. A person buying a ticket with a hidden number coinciding with an open number wins. What is the probability of at least one win? Can one find a reasonable approximation for the distribution of the number of winning tickets?

The main objective of this chapter is to introduce the basic notion of probability theory, namely a probability space, moving slowly from its most elementary version to a more advanced one.

A finite probability space in its simplest form is specified by a finite set $\Omega = \{1, ..., n\}$, whose elements are called *elementary events*, and a collection of non-negative numbers $p_1, ..., p_n$, called *probability distribution* or *probability law* on Ω, assigning probabilities to these elementary events. The probabilities should sum up to one, i.e. $p_1 + ... + p_n = 1$. For those used to think in percentages, 100% corresponds to probability one, the probability p corresponds to $100p\%$. The subsets of Ω are called the *events* and the probability of any event $I \subset \Omega$ is given by the sum of the probabilities of the elementary events contained in I:

$$\mathbf{P}(I) = \sum_{i \in I} p_i.$$

In examples, the set Ω often describes the outcomes of a certain experiment, and probabilities are assigned according to our perception of the likelihood of these outcomes. In the situation of maximum uncertainty all outcomes are equally probable, i.e. all p_i are equal and hence all $p_i = 1/n$, where n is the number of elements in Ω. This probability distribution is called *uniform*. In this case the probability of any event $A \subset \Omega$ is of course just the ratio $|A|/n$, where A is the number of elements in A.

For instance, tossing a coin, can be modeled by the set Ω consisting of two elements $\{H, T\}$ (head and tail). If a coin is fair, then $\mathbf{P}(H) = \mathbf{P}(T) = 1/2$, i.e. tossing a coin is modeled by a uniform distribution on a two point set.

Two events A, B are called *independent* if the probability of their joint occurrence factorizes:

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

For instance, when tossing a coin two times, the results of these two experiments are of course independent, hence the probabilities of any of the 4 events HH, HT, TH, TT equal

$$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

Thus the experiment consisting of two tossing of a fair coin is described by the uniform distribution on a four point set.

Counting the number of outcomes belong to the domain of mathematics called *combinatorics*. We shall use from it only the so called *binomial coefficients*

$$C_n^k = \frac{n!}{k!(n-k)!} \tag{1.1}$$

which yield the number of ways to choose k elements out of a given set of n (say, the number of ways to choose k cards out of a pack of n).

An experiment with two outcomes (success or failure, head or tail, etc), which one encodes by numbers 1 and 0, with given probabilities $p = \mathbf{P}(1)$ and $1 - p = \mathbf{P}(0)$ is called a *Bernoulli trial*. It is clear that the number of ways to have (or to choose) precisely k successes (or one's) in n independent trials is given precisely by the binomial coefficient (1.1). As by independence, the probability of any outcome with k successes out of n experiments equal $p^k(1-p)^{n-k}$, the probability of k successes in n trials equals

$$p_k(n) = C_n^k p^k (1-p)^{n-k}. \tag{1.2}$$

For given n, p this distribution on the set of $(n+1)$ elements $k = 0, 1, ..., n$ is called the (n, p)-*Binomial*. It plays a distinguished role in probability theory.

Remark 1 *As the probabilities sum up to one, we can deduce from our discussion the following remarkable formula of the binomial expansion:*

$$1 = \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k}$$

for any integer n and any $p \in [0, 1]$.

In the financial context one can illustrate the importance of studying independence by the following elementary example. Suppose you keep a portfolio of bonds belonging to 100 different corporations, and the probability of a default during the next three years for each of these corporations equals to 0.02 (i.e. 2%). So called *credit default swaps* (CDS) allows one to buy a protection against the first or, say, the tenth default inside your portfolio (default means a crucial devaluation of the bond and/or of the corresponding payments). If the firms are independent, the number of defaults are distributed like a Binomial (100.0.02) r.v. implying that the probability of at least one (respectively at least 10) defaults equals 0.867 (respectively 0.003). Thus a first-to-default CDS should be valuable, but the ten-to-default CDS is worth almost nothing. On the other hand, if the defaults of your corporations are perfectly correlated, i.e. they occur simultaneously, then the probabilities of at least one or at least 10 (or 50) defaults coincide and equal 0.02. We shall return to this example in Section 3.6.

Now it is time to extend a bit our concept of a probability space. By our initial definition, it was just a finite set Ω with given probabilities of each element. All subsets were considered as possible events. In the extension we have in mind not all subsets will constitute an event, to which a probability can be assigned. This extension is crucial for further development and is very natural from the practical point of view. In fact, the events of the model, also called *measurable subsets* (as their probabilities can be measured) should be linked with particular purposes for which a model is created. For example, consider the population of a city from a point of view of a financial institution. The relation with a client (say, level of credit offered) will be based on certain financial parameters characterizing this client, primarily on an average income per annum. Thus the set of all people with a given income should be an event. Its probability can be estimated as a ratio of the number of people with this level of income to the size of the population. In other words, this would yield a probability of a randomly chosen individual to have an income of this level. On the other hand, a subset of individuals having, say, green eyes, should not be considered as an event in this model. Further on, often certain decisions on offering a financial product to a customer (say, a mortgage) depends on the income exceeding some threshold, say K thousands dollars per annum. For this purpose then, the model will be reduced to four events (subsets of a population) only: the whole population, its complement (the empty set), and the subsets having income below and above the threshold K.

This discussion was meant to justify the introduction of a general concept of a *finite probability space* as a triple $(\Omega, \mathcal{F}, \mathbf{P})$, where Ω is an arbitrary finite set, $\mathcal{F}$ is a collection of its subsets, called *events* or *measurable subsets*, and $\mathbf{P}$ is a positive function on $\mathcal{F}$ assigning probabilities to the events. Such a triple is called a *probability space* if it enjoys the following natural (convince yourself that they are really natural!) properties.

(P1) If A and B are events (i.e. belong to $\mathcal{F}$), then so are their union and intersection $A \cup B$, $A \cap B$.

(P2) The whole set Ω and the empty set $\emptyset$ belong to $\mathcal{F}$ and have probabilities one and zero respectively.

(P3) If A is an event, then its compliment $\bar{A} = \Omega \setminus A$ is also an event and $\mathbf{P}(\bar{A}) = 1 - \mathbf{P}(A)$.

(P4) If $A_1, A_2, ..., A_k$ is a finite collection of nonintersecting events, then the *additivity of probability* holds

$$\mathbf{P}(A_1 \cup A_2 \cup ... \cup A_k) = \mathbf{P}(A_1) + ... + \mathbf{P}(A_k).$$

Remark 2 *To see why (P1) is reasonable let us return to our above example with customers of a financial institution: if they are assessed by two criteria corresponding to the events A and B (say, income per year and the value of property owned), the situations when both criteria are satisfied (the event $A \cap B$) or at least one holds (the event $A \cup B$) should be considered as relevant events.*

Remark 3 *For a given finite probability model $(\Omega, \mathcal{F}, \mathbf{P})$ one can reduce it, by grouping, to an equivalent one in such a way that all subsets of the* *basic underlying set of outcomes become events (or measurable). Say, in the above example with population characterized by the income per annum, one can alternatively take as Ω not the set of all people, but instead the set of all incomes. However, such a reduction is not always natural practically. In particular, this happens when analyzing the dynamics of random variables (so called random processes). In this situation it becomes instructive to use families $\mathcal{F}_t$ of the sets of events parameterized by time (called a filtration) that store the information available to time t.*

The next step in extending the notion of a probability space is in rejecting the assumption of finiteness of Ω. Namely, analyzing repeated experiments, even the simplest Bernoulli ones, naturally leads to the necessity to introduce at least countable sets of outcomes. Consider, for instance, the well known strategy of betting on a certain outcome till it wins. In other words, we are betting on a success in a Bernoulli trial with the probability of success (encoded as number one) until it occurs. Possible outcomes in this experiments are the strings of events

$$1, 01, 001, 0001, ...$$

There are countably many of them and their probabilities equal

$$p, (1-p)p, (1-p)^2 p, ... \tag{1.3}$$

Thus we have a distribution on a set of all natural numbers. As follows from the summation of a geometric sequence, the (infinite) sum of these probabilities equals one, as it should be. The distribution (1.3) is called *p-Geometric*.

As one can now guess the only correction to the above definition of a probability space should be the possibility to take countable sums, which leads to the following definition. A *discrete* (finite or countable) *probability space* or *probability model* is a triple $(\Omega, \mathcal{F}, \mathbf{P})$, where Ω is an arbitrary finite or countable set, $\mathcal{F}$ is a collection of its subsets, called *events* or *measurable subsets*, and $\mathbf{P}$ is a positive function on $\mathcal{F}$ assigning probabilities to the events. Moreover, this triple enjoys the properties (P1)-(P3) above as well as an improved version of (P4):

(P4') If $A_1, A_2, ...$ is a finite or countable collection of nonintersecting events, then its union is also an event and the following σ- *additivity* holds:

$$\mathbf{P}(A_1 \cup A_2 \cup ...) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + ...$$

(this infinite sum denotes the limit as $n \to \infty$ of the finite sums $\mathbf{P}(A_1) + ... + \mathbf{P}(A_n)$).

In simplest situations, like in the above example of geometric distribution, the collection $\mathcal{F}$ coincides with all subsets of Ω and the probability law $\mathbf{P}$ is then specified just by assigning probabilities, say p_k, to each element of Ω.

By far the most important example of a countable probability law is the so called *Poisson distribution* with parameter c, $c > 0$, where the probabilities p_k on the set of all non-negative integers $k \in \Omega = \{0, 1, ...\}$ are defined as

$$p_k = \frac{c^k}{k!} e^{-ck}. \tag{1.4}$$

The Taylor formula for the exponential function ensures that these probabilities sum up to one. One of the reason for the importance of this distribution is the fact that it approximates the binomial distribution when the probability of success is small. More precisely, the following result holds, which will be proved later in Section 5.1.

Theorem 1.0.1 *If $p_k(n)$ is the (n,p)-Binomial distribution* (1.2)*, then*

$$\lim_{n\to\infty} p_k(n) = \frac{c^k}{k!} e^{-c},$$

where p depends on n and tends to zero in such a way that $np \to c$ as $n \to \infty$.

Crucial importance in both theory and applications of probability belongs to *conditional probabilities*, which allow to update our perception of a certain probability law (or its dynamics) once new information becomes available. If A and B represent events in a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, the *conditional probability* of B given A is defined as

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(B \cap A)}{\mathbf{P}(A)}.$$

Notice that as a function of B conditional probability $\mathbf{P}(B|A)$ specifies a probability law on the measurable subsets of A, as it clearly sums up to one:

$$\mathbf{P}(A|A) = \frac{\mathbf{P}(A \cap A)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A}{\mathbf{P}(A)} = 1.$$

From this definition it follows that A and B are independent if and only if $\mathbf{P}(B|A) = \mathbf{P}(B)$, as one should expect (conditioning on something irrelevant should not be relevant).

If $A_1, ..., A_k$ is a *partition* of Ω, i.e. it is a collection of pairwise disjoint events with the union coinciding with Ω, then by the additivity of probability

$$\mathbf{P}(B) = \sum_{j=1}^{n} \mathbf{P}(B \cap A_j),$$

implying the following fundamental *multiple decomposition* law

$$\mathbf{P}(B) = \sum_{j=1}^{n} \mathbf{P}(B|A_j)\mathbf{P}(A_j).$$

In particular, if A is an event and $\bar{A} = \Omega \setminus A$ is its compliment, it implies the following *decomposition* law

$$\mathbf{P}(B) = \mathbf{P}(B|A)\mathbf{P}(A) + \mathbf{P}(B\bar{A})\mathbf{P}(\bar{A}).$$

In practice one often needs to revet the event and its condition. Namely, since by the definition of conditioning

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B|A) = \mathbf{P}(B)\mathbf{P}(A|B),$$

one has

$$\mathbf{P}(A|B) = \mathbf{P}(B|A)\frac{\mathbf{P}(A)}{\mathbf{P}(B)}. \tag{1.5}$$

Plugging in the decomposition rule, yields

$$\mathbf{P}(A|B) = \mathbf{P}(B|A)\frac{\mathbf{P}(A)}{\mathbf{P}(B|A)\mathbf{P}(A) + \mathbf{P}(B\bar{A})\mathbf{P}(\bar{A})}. \tag{1.6}$$

Equations (1.5) and (1.6) are often refereed to as the *Bayes* rules.

The solution to problem 4 below demonstrates how this rule works. Generally speaking, it often occurs that conditional probabilities and not absolute probabilities are naturally given by the model. As an example let us show how the default probabilities of firms for various time periods are calculated. Suppose it is known (as given, say, by default rating services) that for a company a probability of default during a year conditioned on no earlier default equals 0.02 (or 2%). What are the survival probabilities $\mathbf{P}(S_k)$ up to the year $k = 1, 2, ...$ (S_k denotes the event that the company will not default at least till the end of year k) and the default probabilities $\mathbf{P}(D_k)$ during year k? Well, of course $\mathbf{P}(D_1) = 0.02$ and $\mathbf{P}(S_1) = 1 - 0.02 = 0.98$. To calculate these probabilities for the next years, one proceeds as follows.

$$\mathbf{P}(D_2|S_1) = 0.02 = \mathbf{P}(D_2)/\mathbf{P}(S_1).$$

Hence

$$\mathbf{P}(D_2) = 0.02 \times 0.98 = 0.0196,$$

and consequently

$$\mathbf{P}(S_2) = 1 - \mathbf{P}(D_1) - \mathbf{P}(D_2) = 0.9604,$$

and so on with $\mathbf{P}(S_3)$, etc.

We have now all the tools at hand needed for the solutions of the problems formulated in the beginning.

Solution to Problem 1. Required probability equals $1 - p$, where p is the probability that all players have birthdays at different dates. Assume the number of days in a year is 365. The number of all birthday allocations (that are equiprobable) is thus 365^{23}. The number of ways to choose 23 different days equals $365 \times 364 \times ... \times (365 - 22)$. Hence

$$p = \frac{365 \times 364 \times ... \times (365 - 22)}{365^{23}}.$$

Calculation yields for the required probability of having at least two people with the same birthday a remarkably high value $1 - p = 0.5073$.

Solution to Problem 2. (i) Of course it is assumed here that any new born child turns out to be a son (S) or a daughter (D) with equal probabilities. Hence there are 4 elementary events DD, DS, SD, SS, describing possible appearance of a son or a daughter in a family with two children, and all these outcomes are equiprobable and hence have probability 1/4. Hence the event $\exists D$ meaning that the family has a daughter has probability 3/4. Consequently

$$\mathbf{P}(DD|\exists D) = \mathbf{P}(DD)/\mathbf{P}(\exists D) = 1/3.$$

(ii) Now each of the elementary events DS and SD is decompose in a pair of equiprobable events $DS : D$, $DS : S$ and respectively $SD : D$, $SD; S$ (the last letter indicating who has opened the door), each of which thus having probability 1/8. The probability of the event ($Dseen$) that a daughter opens the door equals 1/2. Consequently

$$\mathbf{P}(DD|Dseen) = \mathbf{P}(DD)/\mathbf{P}(Dseen) = (1/4)/(1/2) = 1/2.$$

Solution to Problem 3. When you choose a door, you decompose the set of three doors into two events (you are right), (you are wrong) with probabilities 1/3 and 2/3. Opening the empty door by the host does not change these probabilities. Hence switching allows you to change the probability of winning from 1/3 to 2/3. In the modified version with 100 doors you would be able to improve your chances from 1/100 to 99/100.

Solution to Problem 4. Let us denote by D (respectively $\bar{D}$) the event of having a disease (respectively not having a disease) and by P the event of being tested positive. Then

$$\mathbf{P}(P) = \mathbf{P}(P|D)\mathbf{P}(D) + \mathbf{P}(P|\bar{D})\mathbf{P}(\bar{D}).$$

Consequently,

$$\mathbf{P}(D|P) = \mathbf{P}(P|D)\frac{\mathbf{P}(D)}{\mathbf{P}(P)} = \frac{\mathbf{P}(P|D)\mathbf{P}(D)}{\mathbf{P}(P|D)\mathbf{P}(D) + \mathbf{P}(P|\bar{D})\mathbf{P}(\bar{D})}. \tag{1.7}$$

Now

$$\mathbf{P}(P|D) = 0.99, \quad \mathbf{P}(P|\bar{D}) = 0.02,$$

yielding

$$\mathbf{P}(D|P) = \frac{0.99 \times \mathbf{P}(D)}{0.99 \times \mathbf{P}(D) + 0.02(1 - \mathbf{P}(D))}. \tag{1.8}$$

In case (i) $\mathbf{P}(D) = 1/2$ leading to $\mathbf{P}(D|P) = 0.98$. In case (ii) $\mathbf{P}(D) = 0.001$ leading to $\mathbf{P}(D|P) = 0.047$.

Solution to Problem 5. The number of winning tickets can be considered as the number of successes in N trials, each trial, being a comparison of a printed number with hidden one, has the success probability $1/N$. For large N the dependence between the trials is weak suggesting to use the Poisson approximation for the number of successes in N Bernoulli trials. By the Poisson limit theorem, the distribution of the number of winnings is approximately Poisson with parameter $N \times 1/N = 1$, which remarkably enough does not depend on N. Thus the probability of no win is approximately $1/e = 0.37$. One can also calculate this probability precisely. It turns out that for $N > 10$ the Poisson approximation matches the exact probabilities in at least eight digits.

Chapter 2

Random variables (r.v.)

2.1 Discrete r.v. and random vectors

Random variables (r.v.) represent the core notion of probability theory. One can say that probability theory is a domain of sciences studying r.v. This chapter is devoted to this concept in its various performances starting with the simple discrete r.v.

In its first description a r.v. is understood as an uncertain number, which takes particular values with prescribed probabilities. In particular, a *discrete random variable* X is described by a finite (or more generally countable) family of its possible values $x_1, ..., x_n$ (the *range* of X) together with the probabilities $p_1, ..., p_n$, with which these values can occur (specifying a probability law on the range of X). A simple example is supplied by the r.v. X denoting the face of a die randomly chosen (or randomly thrown). The range of this X is the collection of numbers $\{1, ..., 6\}$ and the probabilities are all equal: $p_1 = ... = p_6 = 1/6$ (if a die is assumed to be fair). One of the basic characteristics of a r.v. is its expectation or mean value. These averages appear constantly in every day live. In probability theory they get a precise meaning. Namely, for a r.v. X with range $x_1, ..., x_n$ and probabilities $p_1, ..., p_n$, the *expectation* (or *mean*, or *average value*) is defined as

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

If all values of X are equiprobable, i.e. all p_i coincide, then this turns to the usual arithmetic mean of the range:

$$\mathbf{E}(X) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Working with r.v. one often needs to perform various operations on them, like adding, multiplying, calculating functional transformations (thus playing with uncertain numbers as with usual ones). Clearly, for a r.v. X with range $x_1, ..., x_n$ and probabilities $p_1, ..., p_n$, and a function $\phi(x)$, the composition $\phi(X)$ can be naturally defined as a r.v. with the range $\phi(x_1), ..., \phi(x_n)$ and the same probabilities for $p_1, ..., p_n$. As may be surprising from the first sight, the situation is not so obvious when one tries to give sense to the sum of r.v. Namely, suppose X is a r.v. with range $x_1, ..., x_n$ and probabilities $p_1, ..., p_n$, and Y is a r.v. with range $y_1, ..., y_m$ and probabilities $q_1, ..., q_m$. What is the meaning of $X + Y$? Clearly the range should consists of all sums $x_i + y_j$. But which probabilities are to be assigned to these values? Of course, no sane person would try to add temperature (or average temperature) with an income per annum. But to add, say, incomes of a firm in consecutive years is a reasonable operation.

In order to get an understanding of what is going on with the sums of r.v. it is instructive to invoke an alternative description of r.v. In this second description a *discrete r.v.* is defined as a function on a discrete probability space. From a first sight it may look like a tautological reformulation. In fact, if X is a r.v. with range $x_1, ..., x_n$ and probabilities $p_1, ..., p_n$, one can equivalently define X as the function on the probability space $\Omega_X = \{1, ..., n\}$ with the law $\{p_i = \mathbf{P}(i)\}_{i=1}^n$ given by $X(i) = x_i$. The definition of the expectation can be equivalently rewritten now as

$$\mathbf{E}(X) = \sum_{\omega \in \Omega_X} X(\omega)\mathbf{P}(\omega). \tag{2.1}$$

However, the problem of giving sense to the sums seems to be solved now almost automatically. Namely, if X and Y are two functions on a discrete probability space Ω, then $X + Y$ is naturally defined also as a function on Ω as $(X + Y)(\omega) = X(\omega) + Y(\omega)$. It is also obvious from (2.1) that the expectation is a linear operation, i.e.

$$\mathbf{E}(aX + bY) = a\mathbf{E}(X) + b\mathbf{E}(Y) \tag{2.2}$$

for any numbers a, b.

Let us return to our original question. Given a r.v. X with range $x_1, ..., x_n$ and probabilities $p_1, ..., p_n$, and a r.v. Y with range $y_1, ..., y_m$ and probabilities $q_1, ..., q_m$, how to give meaning to the sum $X + Y$? The point is that in order to be able to define it as a point-wise sum of two functions, one has to have X and Y to be specified as functions on the same probability space. But if we would apply to Y the same construction as for X above, we would define it as a function on the probability space $\Omega_Y = \{1, ..., m\}$ with the law $\{q_j = \mathbf{P}(j)\}_{j=1}^m$, which is different from Ω_X. In order to circumvent this problem, the idea is to take as common probability space the product

$$\Omega_{X,Y} = \Omega_X \times \Omega_Y = \{(i,j) : i = 1, ..., n; j = 1, ..., m\}.$$

However, the probability law on $\Omega_{X,Y}$ can not be reconstructed from the laws of X and Y. It should be given as additional input to the model, which leads naturally to the notion of a random vector. Namely, a discrete two-dimensional *random vector* is a pair of random variables (X, Y) with ranges $x_1, ..., x_n$ and $y_1, ..., y_m$ respectively, for which the *joint probabilities* (or *joint probability distribution*) are specified:

$$p_{ij} = \mathbf{P}(X = x_i, Y = y_j), \quad i = 1, ..., n, \quad j = 1, ..., m.$$

Notice that the distributions of the components X, Y (called in this setting the *marginal distributions* of the random vector) are easily reconstructed from the joint probabilities, as by the additivity of probability

$$p_i = \mathbf{P}(X = x_i) = \sum_{j=1}^m \mathbf{P}(X = x_i, Y = y_j) = \sum_{j=1}^m p_{ij},$$

$$q_j = \mathbf{P}(Y = y_j) = \sum_{i=1}^n \mathbf{P}(X = x_i, Y = y_j) = \sum_{i=1}^n p_{ij}.$$

Specifying the joint probability law p_{ij} on the basis of the given marginal distributions, is called the *coupling* of the r.v. X and Y. It can be done in many ways (see for instance the independent coupling, described below).

Once the joint probabilities are specified, both X and Y become defined on the same probability space $(\Omega_{X,Y}, p_{ij})$ as the functions

$$X(i,j) = x_i, \quad Y(i,j) = y_j,$$

and one can apply the previously stated linearity of the expectation for r.v. defined on a single probability space to conclude that (2.2) holds. Notice that though in order to define the sum $X + Y$ the joint probabilities have to be specified, the expectation of the sum actually does not depend on these joint distributions, but only on the marginal ones. This is specific for the operation of taking sums. For the expectation of the product, say, one obtains

$$\mathbf{E}(XY) = \sum_{i=1}^n \sum_{j=1}^m x_i y_j p_{ij},$$

which can not be calculated without the knowledge of the joint probabilities. A particular case of coupling is of importance. One says that the r.v. X, Y are *independent* if the joint probabilities factorize:

$$p_{ij} = \mathbf{P}(X = x_i, Y = y_j) = \mathbf{P}(X = x_i)\mathbf{P}(Y = y_j) = p_i q_j,$$

i.e. the events $(X = x_i)$ and $(Y = y_j)$ are independent for all x_i, y_j. Evidently, in this case the expression for $\mathbf{E}(XY)$ also factorizes:

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

Let us summarize our conclusions in the following

Theorem 2.1.1 *Expectation is a linear function on discrete r.v. , i.e.* (2.2) *holds for any r.v.* X*,*Y*. Moreover, if* X *and* Y *are independent, then the expectation of the product of* X *and* Y *equals to the product of their expectations.*

Similarly one defines d-dimensional random vectors as the collection $(X, Y, ..., Z)$ of d uncertain numbers, for which the *joint probabilities* are specified:

$$\mathbf{P}(X = x_i, Y = y_j, ..., Z = z_k)$$

for all possible values x_i of X, y_j of Y, ... and z_k of Z. The r.v. $X, Y, ..., Z$ are called *independent* if all the joint probabilities factorize:

$$\mathbf{P}(X = x_i, Y = y_j, ..., Z = z_k) = \mathbf{P}(X = x_i)\mathbf{P}(Y = y_j)...\mathbf{P}(Z = z_k).$$

The above theorem extends straightforwardly to the d-dimensional situation:

Theorem 2.1.2 *For any collection of r.v.* $X_1, ..., X_d$ *one has*

$$\mathbf{E}\sum_{i=1}^{d} X_i = \sum_{i=1}^{d} \mathbf{E}(X_i).$$

Moreover, if $X_1, ..., X_d$ *are independent, then*

$$\mathbf{E}\prod_{i=1}^{d} X_i = \prod_{i=1}^{d} \mathbf{E}(X_i).$$

An important example of a r.v. on a probability space Ω is given by the so called *indicators* $\mathbf{1}_A$ for an event A in Ω. By definition $\mathbf{1}_A(x)$ equals one or zero for $x \in A$ or otherwise respectively. One easily sees that for any event A

$$\mathbf{E}(\mathbf{1}_A) = \mathbf{P}(A).$$

The following two classical examples nicely demonstrate the use of indicators in conjunction with the linearity of expectation for the calculation of averages.

Example: Records. Let X_k, $k = 1, ..., n$, denote the independent identically distributed (i.i.d.) r.v.'s of winners's times in n marathons. Assume for simplicity that $\mathbf{P}(Y_i = Y_j) = 0$ for any $i \neq j$. We call Y_m a record, if

$$Y_m < \min(Y_1, ..., Y_{m-1}).$$

We are interested in the average number of records K_n. As for a given m the events $(Y_i < \min\{Y_j : j \neq i\})$, $i = 1, ..., m$ are equiprobable and their union is the full set, one concludes that the probability of Y_m to be a record equals $1/m$. Hence, denoting by R_m the indicator r.v. of the event that the m th time is a record, one concludes that $\mathbf{E}(R_m) = 1/m$. Consequently,

$$\mathbf{E}(K_n) = \mathbf{E}\left(\sum_{m=1}^{n} R_m\right) = \sum_{m=1}^{n} \frac{1}{m}.$$

(In analysis one proves that this sum grows like $\ln n$ for large n.)

Example: Ladies and gentlemen on an a tea party. Suppose 10 ladies and 10 gentlemen sit in 20 chairs around an oval table. All seating arrangements are equiprobable. How many gentlemen G do you expect to have a lady sitting immediately to their left? Let S_i (respectively $\bar{S}_i$) denote the event that a man (respectively a lady) is in i th chair, $i = 1, ..., 20$. Let X_i, $i = 1, ..., 20$, denote the indicator of the event $S_i \cap \bar{S}_{i+1}$ (where we put $S_{21} = S_1$, as the chair 1 is the left neighbor of the chair 20). One has

$$\mathbf{E}(X_i) = \mathbf{P}(S_i \cap \bar{S}_{i+1}) = \mathbf{P}(S_i)\mathbf{P}(\bar{S}_{i+1}|S_i) = \frac{10}{20}\frac{10}{19} = \frac{5}{19}.$$

Consequently

$$\mathbf{E}(G) = \mathbf{E}\left(\sum_{m=1}^{20} X_m\right) = 20 \times \frac{5}{19} = 5\frac{5}{19}.$$

2.2 Continuous r.v. and their expectations

Here we shall introduce another useful class of r.v., continuous r.v., as well as their main examples.

As already discussed above in discrete case, r.v. can be described in two ways: by their distributions and by their functional realizations. We shall introduce here continuous r.v. via the more elementary first description leaving the second one for the next section. One says that X is a *continuous (or more precisely absolutely continuous)* r.v., if there exists a integrable (say, continuous or piecewise continuous) non-negative function f_X, called the *probability density function* of X, such that for any interval $[a, b]$ the probability of X taking values in $[a, b]$ (closed interval) or (a, b) (open interval) is given by the integral:

$$\mathbf{P}(X \in [a,b]) = \mathbf{P}(X \in (a,b)) = \int_a^b f_X(y)\,dy.$$

This implies in particular that for a continuous r.v., unlike a discrete one, any particular value is taken with zero probability.

As the total probability always sums up to one, any probability density function satisfies the normalizing condition

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

If the range of X belongs to $[a, b]$, then clearly f_X vanishes outside this interval and the normalizing condition rewrites as $\int_a^b f_X(x)\,dx = 1$.

A connection with a general theory of r.v. (see next section) is given by the so called *distribution function* F_X of a r.v. X defined as

$$F_X(x) = \mathbf{P}(X \leq x).$$

For a continuous r.v. X with a probability density function f_X the distribution function F_X is clearly represented as the integral

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,dy. \tag{2.3}$$

If $f_X(x)$ is continuous at x, then by the main law of calculus

$$f_X(x) = \frac{d}{dx} F_X(x), \tag{2.4}$$

implying in particular that $F_X(x)$ is differentiable at x.

For example, a r.v. X is called *uniformly distributed* on an interval $[a,b]$ if its probability density function $f_X(x) = f_X$ is a constant (does not depend on x) on $[a,b]$ and vanishes outside it. Since $\int_a^b f_X \, dx = 1$ it follows that this constant f_X equals $1/(b-a)$. Clearly the uniform distribution describes the situation of the maximum uncertainty about the position of X on $[a,b]$.

The second most important class of examples is given by the so called *normal or Gaussian* r.v. ranging in the whole $\mathbf{R}$ and specified by the probability density functions of the form

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \tag{2.5}$$

where $\mu \in \mathbf{R}$ and $\sigma > 0$ are two parameters. These r.v. are usually denoted $N(\mu, \sigma^2)$. The r.v. $N(0,1)$ is called *standard normal* and has the density

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}. \tag{2.6}$$

Its distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{y^2}{2}\right\} dy \tag{2.7}$$

can not be expressed in closed form, but is tabulated in standard statistical tables with great precision due to its importance. The properties of general $N(\mu,\sigma^2)$ are easily deduced from the standard normal r.v. due to their simple linear connections. Namely, if X is $N(0,1)$ r.v., then

$$Y = \mu + \sigma X \tag{2.8}$$

is $N(\mu,\sigma^2)$. In fact,

$$F_Y(x) = \mathbf{P}(\mu + \sigma X \leq x) = \mathbf{P}\left(X \leq \frac{x-\mu}{\sigma}\right),$$

so that the distribution function of Y is

$$F_Y(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Differentiating with respect to x yields for the probability density of Y the expression (2.5) as was claimed.

The meaning of the parameters μ, σ, as well as the importance of the normal r.v. will be revealed later. We notice here that

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}\,dx=1$$

for any $\mu\in\mathbf{R},\sigma>0$ as it should be for any probability density function. This fact is not obvious but can be checked by the methods of calculus.

In the analysis of the stock market the central role belongs to the so called log-normal r.v. A r.v. X is called *log-normal* if its logarithm $Y=\ln X$ is normal $N(\mu,\sigma^2)$.

Another key example constitute the exponential distributions that model memoryless waiting times in a variety of situations from the cash counter queues and default times to the processes of nuclear radiation. One says that a positive r.v. τ is *memoryless* if for any $s,t>0$

$$\mathbf{P}(\tau>t+s|\tau>s)=\mathbf{P}(\tau>t),$$

i.e. the waiting time after a moment s does not depend on the waiting time before this moment.

Lemma 2.2.1 *Let f be a continuous function $(0,\infty)\to(0,\infty)$ such that $f(s+t)=f(s)f(t)$ for all s,t. Then there exists a number a s.t. $f(t)=e^{at}$.*

Proof. Introducing $g=\ln f$ implies $g(s+t)=g(s)+g(t)$. Consequently $g(1/n)=g(1)/n$ for all natural n. Hence for all natural m,n one has $g(m/n)=(m/n)g(1)$ implying by continuity that $g(t)=tg(1)$ for all t. Hence $f(t)=e^{tg(1)}$.

Applying this lemma to the function $f(t)=\mathbf{P}(\tau>t)$ and taking into account that $f(t)\leq 1$ for all t so that the corresponding constant a should be negative, implies that a memoryless r.v. τ is *θ-exponential* for a certain positive constant θ, i.e.

$$\mathbf{P}(\tau>t)=\exp\{-\frac{t}{\theta}\}.$$

The distribution function of this r.v. is then given by

$$F_\tau(t)=\mathbf{P}(\tau\leq t)=1-\exp\{-\frac{t}{\theta}\}.$$

It is differentiable so that the probability density function f_τ is defined as

$$f_\tau(t)=\frac{d}{dt}F_\tau(t)=\frac{1}{\theta}\exp\{-\frac{t}{\theta}\}.$$

A slight extension of this distribution forms a standard model for the default probabilities in the theory of credit risk, see Section 3.6.

When adapting the notions related to discrete r.v. to the continuous r.v. the rule of thumb is to use integrals instead of sums. For instance, for a r.v. X with a probability density function $f_X(x)$ the *expectation* of X is defined as

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx. \tag{2.9}$$

As in discrete case one often needs to work not only with r.v., but also with random vectors. For a pair of r.v. X and Y one says that their *joint distribution* is defined, or the *random vector* (X, Y) is defined, if for any intervals A, B of the real line the joint probabilities $\mathbf{P}(X \in A, Y \in B)$ are specified. One says that X and Y are independent, if these probabilities factorize, i.e. if for any A, B

$$\mathbf{P}(X \in A, Y \in B) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B). \tag{2.10}$$

As in discrete case, specifying joint distributions for two r.v. X, Y with given distribution functions F_X and F_Y is called the *coupling* of the r.v. X and Y. The simplest coupling is of course the *independent coupling* given by (2.10).

The following result represents an extension of the corresponding facts obtained above for discrete r.v. We present it without a proof, as the technical issues required (multiple integrals) are beyond the scope of our exposition (see however the next section).

Theorem 2.2.1 *(i) Suppose X is a r.v. with the probability density function f_X and let ϕ be any continuous (or piecewise continuous) function. Then*

$$\mathbf{E}(\phi(X)) = \int_{-\infty}^{\infty} \phi(x) f_X(x)\, dx. \tag{2.11}$$

(ii) Suppose X and Y are r.v. forming a random vector. Then

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

(iii) If these X and Y are independent, then

$$\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y). \tag{2.12}$$

Similarly one can define a *n-dimensional random vector* $(X_1, ..., X_n)$ by specifying its *joint probabilities*:

$$\mathbf{P}(X_1 \in A_1, ..., X_n \in A_n) \tag{2.13}$$

for all intervals $A_1, ..., A_n$. One says that the r.v. $X_1, ..., X_n$ are independent, if all these probabilities factorize:

$$\mathbf{P}(X_1 \in A_1, ..., X_n \in A_n) = \mathbf{P}(X_1 \in A_1) ... \mathbf{P}(X_n \in A_n).$$

As easily follows by induction from the above stated fact for $n = 2$, in this case the expectation again factorizes:

$$\mathbf{E}(X_1 ... X_n) = \mathbf{E}(X_1) ... \mathbf{E}(X_n). \tag{2.14}$$

The r.v. $X_1, ..., X_n$ are called *mutually Gaussian* (equivalently the random vector $(X_1, ..., X_n)$ is called *Gaussian*), if any linear combinations of X_i is a Gaussian r.v. As for Gaussian r.v., one can specify the joint distributions for Gaussian random vectors explicitly.

Let us calculate the expectation for the uniform, normal and log-normal r.v. If X is uniformly distributed on an interval $[a, b]$, its probability density functions $f_X(x)$ is a constant $1/(b-a)$ so that

$$\mathbf{E}(X) = \int_a^b \frac{x}{b-a}\, dx = \frac{1}{2(b-a)} x^2 \mid_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2},$$

so that, as one expects from the common sense, the mean value of the points taking all values of the interval $[a, b]$ with equal probabilities is just the middle of this interval.

If X is a normal r.v. $N(\mu, \sigma^2)$, then

$$\mathbf{E}(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} dx,$$

which by changing the variable of integration x to $y = (x - \mu)/\sigma$ rewrites as

$$\mathbf{E}(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y\sigma \exp\left\{-\frac{1}{2}x^2\right\} dx + \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}x^2\right\} dx.$$

The first integral here vanishes (which one can see either from the general evident fact that the integral of any odd function vanishes, or by the integration by parts followed by the explicit evaluation), and the second integral equals μ, because the integral of any a probability density function equals one. Thus the parameter μ in the notation $N(\mu, \sigma^2)$ denotes the mean value of this normal r.v.

If X is log-normal, so that $X = e^Y$ with a normal $N(\mu, \sigma^2)$ r.v. Y, then by (2.11)

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} e^x \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} dx.$$

Changing variable of integration to $y = (x - \mu)/\sigma$ yields

$$\mathbf{E}(X) = e^{\mu} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{\sigma y - \frac{y^2}{2}\right\} dy,$$

which rewrites as

$$e^{\mu} e^{\sigma^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y - \sigma)^2\right\} dy.$$

As the integral here equals one it implies finally

$$\mathbf{E}(X) = \exp\{\mu + \sigma^2/2\}. \tag{2.15}$$

The following calculations form the basis for the deduction of the Black-Sholes option pricing formula in Section 3.4.

Proposition 2.2.1 *Suppose X is log-normal, so that $X = e^Y$ with a normal $N(\mu, \sigma^2)$ r.v. Y, and let K be a positive constant. Then*

$$\mathbf{E}\max(X - K, 0) = \tilde{\mu}\Phi\left(\frac{\ln(\tilde{\mu}/K) + \sigma^2/2}{\sigma}\right) - K\Phi\left(\frac{\ln(\tilde{\mu}/K) - \sigma^2/2}{\sigma}\right), \tag{2.16}$$

where

$$\tilde{\mu} = \mathbf{E}(X) = \exp\{\mu + \sigma^2/2\}.$$

Proof. By (2.8),

$$\mathbf{E}\max(X - K, 0) = \mathbf{E}\max(\exp\{\mu + \sigma Z\} - K, 0)$$

with a standard $N(0, 1)$ r.v. Z. Consequently

$$\mathbf{E}\max(X - K, 0) = \int_{(\ln K - \mu)/\sigma}^{\infty} (e^{\mu + \sigma x} - K) \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx$$

$$= \tilde{\mu} \int_{(\ln K - \mu)/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x - \sigma)^2}{2}\right\} dx - K \int_{(\ln K - \mu)/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx,$$

which rewrites as (2.16).

2.3 General r.v.: transformations and simulation

Conceptually, the following discussion of the general notion of r.v. and their transformations would be possibly the most difficult part of the Lectures, as we shall move along a carefully chosen logical path at the border of a deep mathematical discipline (measure theory) trying not to plunge into its abysses, but at the same time to extract the most useful fruits from its immense richness. By * a more advanced material is marked.

The main results obtained will concern the functional transformations of arbitrary r.v. implying in particular 1) the universal method of their simulations (used basically in all applications of probability), 2) the representation of their means via the integral over the unit interval (giving the basic tool for all calculations with r.v.) and finally 3) the method of the transformations of arbitrary r.v. into a given one (in practice usually Gaussian) that lies in the heart of the method of the Gaussian copulas, which is crucial in the theory of credit derivatives and risk measurement. At the end we shall summarize again the main points to be learned from this discussion.

In its first, more elementary description, a r.v. is understood as an uncertain number which takes particular values with prescribed probabilities. More precisely, any r.v. X is described by its *distribution function* $F_X(x)$ that for any $x \in \mathbf{R}$ specifies the probability of the event that the value of X does not exceed x:

$$F_X(x) = \mathbf{P}(X \leq x).$$

Clearly $F_X(x)$ is a non-decreasing function of x taking values 0 and 1 at $-\infty$ and $+\infty$ respectively (more precisely, the latter means that $\lim_{x\to\infty} F_X(x) = 1$, $\lim_{x\to-\infty} F_X(x) = 0$). By the additivity of probability the knowledge of F_X allows to specify the probabilities of the values of X to belong to any given interval, as

$$\mathbf{P}_X((a,b]) = \mathbf{P}(X \in (a,b]) = \mathbf{P}(X \leq b) - \mathbf{P}(X \leq a) = F_X(b) - F_X(a). \tag{2.17}$$

One says that the *range of* X belongs to an interval $[a,b]$ if the probability of the values of X being below a or above b vanishes. This clearly means that $F_X(x) = 0$ for all $x < a$ and $F_X(x) = 1$ for all $x \geq b$.

The discrete r.v., studied in the previous section can be characterized by the fact that their distribution functions are piecewise constant. For instance, a Bernoulli r.v. X taking values 1 and 0 with probability p and

$1 - p$ has the distribution function

$$F_X(x) = \begin{cases} 0, \quad x < 0 \\ 1 - p, \quad 0 \leq x < 1 \\ 1, \quad x > 1 \end{cases} \tag{2.18}$$

On the other hand, continuous r.v. have continuous distribution functions (expressed via their probability density functions by formula (2.3)), which are even differentiable whenever a probability density function is continuous.

A r.v. does not have to be either discrete or continuous (it could contain rather weird mixtures of these two types), though in practical examples r.v. usually belong to one of these two types. On the other hand, as one easily observed in both cases the distribution function is right continuous (i.e. when $x_n \to x_0$ as $n \to \infty$ so that all $x_n > x_0$, then $F(x_n) \to F(x)$). This property turns out to be general, i.e. it holds for any distribution function.

Remark 4 *The right continuity of a distribution function follows from the continuity of the probability measure (see Remark 5 below), as the set* $(X \leq x)$ *is the intersection of the sets* $(X \leq x + 1/n)$ *over all natural* n*. Some authors prefer to use distribution functions defined as* $\tilde{F}_X(x) = \mathbf{P}(X < x)$*. Thus defined* $\tilde{F}_X(x)$ *is always left continuous.*

The above given definition of a r.v. via its distribution function is intuitively appealing, but left unclear many important issues, both theoretical and practical, say how to take the compositions of various r.v. or how to add them. These issues are resolved in another description of r.v. as a function on a probability space.

A discrete probability space was introduced in Chapter 1. For a general definition one only has to denounce the assumption of countability of Ω. Namely, a general *probability space* or *probability model* is a triple $(\Omega, \mathcal{F}, \mathbf{P})$, where Ω is an arbitrary set, $\mathcal{F}$ is a collection of its subsets, called *events* or *measurable subsets*, $\mathbf{P}$ is a positive function on $\mathcal{F}$ assigning probabilities to the events and the conditions (P1)-(P3), (P4') from Chapter 1 hold. The basic examples of an uncountable probability space (sufficient for the most of practical applications) represent the *geometric probability models*, where Ω is a real line, or a plane, or more generally a subset of the d-dimensional Euclidean space. The family $\mathcal{F}$ is then usually chosen as the set of the so called *Borel* sets. The simplest *Borel* subsets of the real line $\mathbf{R}$ are the intervals $[a, b]$, and the whole family $\mathcal{B}$ of the *Borel subsets* of the line $\mathbf{R}$ is defined as the class of subsets of $\mathbf{R}$ obtained by applying at most countably many times the operations of intersection and union to open and closed intervals. Similarly, the family $\mathcal{B}$ of the *Borel subsets* of the plane $\mathbf{R}^2$ is defined as the class of subsets of $\mathbf{R}^2$ obtained by applying at most countably many times the operations of intersection and union to the rectangles $[a, b] \times [c, d]$. Similarly Borel subsets of an arbitrary Euclidean space are constructed from d dimensional parallelepipeds.

By far the most important example of a Borel probability space is the *standard (or Lebesgue) probability space* that models the uniform distribution on the unit interval $[0, 1]$, where all points are equally probable (in other words, the points of $[0, 1]$ are considered as the values of a uniformly distributed r.v., introduced in the previous section). In this model $\Omega = [0, 1]$ and the class $\mathcal{F}$ is given by the Borel class $\mathcal{B}$ of the subsets of Ω. The simplest events (or Borel subsets) of this space are the subintervals $[a, b] \subset [0, 1]$ and their probability equals their length: $\mathbf{P}([a, b]) = b - a$. Also any finite or countable union of nonintersecting subintervals of $[0, 1]$ specifies an event (is a measurable Borel set), the probability of such an event being of course the total length, i.e. the sum of the lengths of all intervals entering this subset. The so called *measure theory* is a special branch of mathematics designed to give a proper understanding to what is measurable or not. For practical applications one just has to have in mind that though not every subset of $[0, 1]$ is measurable, i.e. represents an event to which a probability of occurrence can be assigned, the usual subsets A met in practice are measurable and their probabilities are given by the integral $\mathbf{P}(A) = \int_0^1 \mathbf{1}_A(x)dx$, where 1_A is the *indicator function* of the set A (i.e. $1_A(x)$ equals one or zero when respectively x belongs to A or not).

Remark 5 * *When we extend the notion of a probability space from a finite to a countable setting, we have also extended the additivity axiom (P4) to the countable (or σ-) additivity (P4'). However, denouncing the countability of the basic set Ω, we still have to keep the additivity countable (an attempt to use arbitrary, i.e. uncountable, unions would lead to inconsistency). It is easy to deduce from the σ- additivity of $\mathbf{P}$ that it also enjoys the following continuity property: if $\{A_n\}_{n=1}^\infty$ is a collection of measurable subsets such that $A_{n+1} \subset A_n$ for all n, then $A = \cap_n A_n$ is measurable and*

$$\mathbf{P}(A) = \lim_{n\to\infty} \mathbf{P}(A_n).$$

In particular, this property implies that distribution functions of any r.v. should be right continuous. Finally let us notice that unlike the case of finite *spaces, one usually can not reduce a given probability space to get a space with all sets being measurable. The existence of nonmeasurable sets is a crucial feature of uncountable probability spaces.*

In its second description a general *random variable* (r.v.) is defined as a *measurable* function $X(\omega)$ on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, i.e. a function, for which the subsets $(X \leq x) = \{\omega \in \Omega : X(\omega) \leq x\}$ belong to $\mathcal{F}$ for any x, i.e. they represent events, for which certain probabilities can be assigned. The notion of measurability is one of the central in probability theory and is not so easy to grasp in its full richness. For practical purposes it is usually enough to know that any continuous, or piecewise continuous, or any monotone function is measurable.

A measurable function $X(\omega)$ on the standard probability space $[0,1]$ (or more generally on geometric or Borel probability spaces) introduced above is often called *Borel measurable*, or simply a *Borel function*).

The link to our previous definition of a r.v. via its distribution is as follows. If $X(\omega)$ is a measurable function on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, its distribution function is defined of course as

$$F_X(x) = \mathbf{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbf{P}(X \leq x). \tag{2.19}$$

For a given distribution function F_X, the probability law $\mathbf{P}_X$ is defined by (2.17) on the intervals and then can be extended by (countable) additivity to all Borel subsets $\mathcal{B}$ of the real line and hence specifies a probability space $(\mathbf{R}, \mathcal{B}, \mathbf{P}_X)$. The simplest functional representation for a r.v. given by its distribution function F_X is then the identical map $X(x) = x$ on this Borel probability space. As was already clear from the discussion of the discrete r.v., functional representation of a r.v. is not unique (it is in some sense similar to choosing coordinates on a plane). In particular, the next result (that could possibly seem surprising) states that any r.v. given by an arbitrary distribution function can be represented as a measurable function on the standard probability space, in other words, as a function of a uniform r.v.

Theorem 2.3.1 *Let $F(x)$ be an arbitrary non-decreasing right continuous function on $\mathbf{R}$ such that $\lim_{x\to\infty} F_X(x) = 1$, $\lim_{x\to-\infty} F_X(x) = 0$. Then there exists a nondecreasing right continuous function $X(\omega)$ on $[0,1]$ such that $F = F_X$ with F_X given by (2.19). Explicitly, X can be chosen as the generalized inverse of F introduced below.*

Proof (simple case). Suppose that $F(x)$ is a continuous and strictly increasing function, the latter meaning that $x < y$ always implies $F(x) <$

$F(y)$. In this case the inverse function F^{-1} is clearly well defined and is a continuous strictly increasing function on $[0,1]$ (recall that $x = F^{-1}(y)$ denotes the solution to the equation $F(x) = y$). We claim that one can take $X(\omega) = F^{-1}(\omega)$. In fact, for this X,

$$F_X(x) = \mathbf{P}(\{\omega \in [0,1] : X(\omega) \leq x\}) = \mathbf{P}(\{\omega \in [0,1] : \omega \leq F(x)\}) = F(x),$$

as required, the latter equation is due to the fact that $\mathbf{P}$ is the uniform measure and the length of the interval $[0, F(x)]$ equals $F(x)$.

*Proof (general case)**. In general situation one clearly has to look for an appropriate extension of the notion of the inverse function. Namely, for any non-decreasing function $F(x)$ on $\mathbf{R}$ let us define its *generalized inverse* function F^{-1} as follows.

$$F^{-1}(y) = \sup\{u : F(u) \leq y\}. \tag{2.20}$$

It is easy to see that when $F(x)$ is continuous and strictly increasing then $x = F^{-1}(y)$ yields the unique solution to the equation $F(x) = y$. In order to proceed in the general case, we shall need the following properties of this generalized inverse:

(I1) for any nondecreasing function F on $\mathbf{R}$ one has

$$F^{-1}(y) = \inf\{u : F(u) > y\}; \tag{2.21}$$

(I2) if additionally F is right continuous, then

$$\{F^{-1}(y) < x\} \subset \{y < F(x)\} \subset \{F^{-1}(y) \leq x\} \subset \{y \leq F(x)\} \tag{2.22}$$

(right continuity is needed only for the last inclusion);

(I3) if additionally F is continuous, then F^{-1} is the *right inverse* of F in the sense that for all y

$$(F \circ F^{-1})(y) = F(F^{-1}(y)) = y \tag{2.23}$$

whenever y is inside the range of F (which partially justify the notation F^{-1}).

Remark 6 * *These properties are easily obtained. Namely, to get (I1) assume that the r.h.s. of* (2.21) *and* (2.20) *do not coincide. Then there exists a number a such that*

$$\sup\{u : F(u) \leq y\} < a < \inf\{u : F(u) > y\}.$$

But then the left inequality implies $F(a) > y$ and the right inequality implies $F(a) \leq y$ leading to a contradiction and hence proving (2.21)*. Turning to (I2) observe that the first two inclusions in* (2.22) *follow directly from* (2.21)*. To prove the last one it remains to see that if $F^{-1}(y) = x$, then $F(u) > y$ for all $u > x$ again by* (2.21)*, and consequently $F(x) \geq y$ by the right continuity of F. Finally the property (I3) follows directly from the coincidence of the r.h.s. of* (2.21) *and* (2.20)*. Let us notice also for completeness that (as one easily sees) the function* (2.21) *is always right continuous (even if F is not).*

Let us show now that $X(\omega) = F^{-1}(\omega)$ satisfies the requirement of the theorem under its general assumptions. Since ω is uniformly distributed on $[0, 1]$,

$$\mathbf{P}(\omega < F(x)) = \mathbf{P}(\omega \leq F(x)) = F(x).$$

Hence from the last two inclusions in (2.22) with $\omega = y$ it follows that $\mathbf{P}(F^{-1}(\omega) \leq x\} = F(x)$, as required.

The above theorem yields a universal tool for simulating arbitrary r.v., as it reduces the problem to the simulation of uniform variables. A generator of uniformly distributed numbers on $[0, 1]$ is attached to all basic program packages, and taking the (generalized) inverse function F^{-1} specifies a r.v. with the distribution function F. It is useful even for discrete r.v., where $F(x)$ is clearly not strictly increasing (hence the practical necessity to use a more complicated expression (2.20). For instance, assume we are interested in simulating the discrete r.v. X that takes $(n + 1)$ integer values $0, 1, ..., n$ with equal probabilities. Thus the distribution function is

$$F_X(x) = \begin{cases} 0, \quad x < 0 \\ \dfrac{1}{n+1}, \quad 0 \leq x < 1 \\ \dfrac{2}{n+1}, \quad 1 \leq x < 2 \\ \ldots \\ \dfrac{n}{n+1}, \quad n-1 \leq x < n \\ 1, \quad x \geq n \end{cases} \tag{2.24}$$

Using (2.20) one finds the generalized inverse

$$F_X^{-1}(y) = \begin{cases} 0, & 0 \le y < \dfrac{1}{n+1} \\ 1, & \dfrac{1}{n+1} \le y < \dfrac{2}{n+1} \\ \ldots & \\ n, & \dfrac{n}{n+1} \le y < 1 \\ n+1, & y = 1 \end{cases} \tag{2.25}$$

that can be expressed as a single formula

$$F^{-1}(y) = [(n+1)y],$$

using the integer part function $[u]$ that equals the maximum integer not exceeding u. Consequently, if θ is uniformly distributed on $[0,1]$, the r.v. $X(\theta) = [(n+1)\theta]$ is a discrete r.v. taking values $0, 1, ..., n$ with equal probabilities.

The representation of a r.v. X as a function $X(\omega)$ on the standard probability space allows to define its *expectation* in a unified way (independently on whether X is discrete, or continuous or from neither of these classes) as

$$\mathbf{E}(X) = \int_0^1 X(\omega)\, d\omega. \tag{2.26}$$

Moreover, for any piece-wise continuous function ϕ, one can now naturally define the r.v. $\phi(X)$ as the composition $\phi(X(\omega))$. Using definition (2.26) to the r.v. $\phi(X)$ yields the formula for its expectation:

$$\mathbf{E}(\phi(X)) = \int_0^1 \phi(X(\omega))\, d\omega. \tag{2.27}$$

The next result shows that in case of continuous r.v. our new general definition of the expectation coincides with the previously given one.

Theorem 2.3.2 *Let a r.v. X, given as a function $X(\omega)$ on the standard probability space, has a probability density function f_X. And let ϕ be a piecewise continuous function. Then*

$$\mathbf{E}(\phi(X)) = \int_0^1 \phi(X(\omega))\, d\omega = \int_{-\infty}^{\infty} \phi(x) f_X(x)\, dx. \tag{2.28}$$

Proof*. It is based on the ideas of linearity and approximation. Namely, as any continuous (or piece-wise continuous) function can be uniformly approximated on any interval by piece-wise constant functions, it is enough to show (2.28) for piece-wise constant ϕ. But any such ϕ is a linear combination of the indicator functions of the form $\mathbf{1}_A$ with an interval A. Hence, by the linearity of the integral, it is sufficient to show (2.28) for these indicators. But in case $\phi = \mathbf{1}_A$ the equation (2.28) reduces to the equation

$$\mathbf{P}(X \in A) = \int_A f_X(x)\, dx,$$

which follows from the definition of the probability density function f_X.

In practice one often meets with the problem of finding a distribution of a certain function of a random parameter, whose distribution is known. In elementary examples this function may be linear, as say, in the problem of calculating the distribution of your income in dollars, when you know it in pounds, or the problem of recalculating the distribution of your income after tax. The next result solves this problem for a general monotone function.

Theorem 2.3.3 *Let X be a continuously distributed r.v. with a distribution function F_X being given by* (2.3) *with a certain f. And let G be a nondecreasing function. Then the distribution function of the r.v. $G(X)$ is the composition of F_X and G^{-1}, i.e. it equals $F_X \circ G^{-1}$, where G^{-1} is the generalized inverse of G.*

Proof (simple case). If G is continuous and strictly increasing, then

$$\mathbf{P}(G(X) \le y) = 1 - \mathbf{P}(G(X) > y) = 1 - \mathbf{P}(X > G^{-1}(y))$$
$$= \mathbf{P}(X \le G^{-1}(y)) = F_X(G^{-1}(y)),$$

as required.

Proof (general case)*. For general G the generalized inverse G^{-1} should be used. By (2.22)

$$\{G^{-1}(y) < X\} \subset \{y < G(X)\} \subset \{G^{-1}(y) \le X\}$$

for any number y. Since X is continuous, the probabilities of the left and the right sets in this inclusion coincide. Consequently all three sets have the same probability. Therefore, the same calculations as above do the job.

Of importance are the following corollaries.

Theorem 2.3.4 *Under the assumptions of the previous theorem assume additionally that G is itself a distribution function (i.e. it is right continuous and takes values 0 and 1 at $-\infty$ and ∞ respectively). Then (i) the r.v. $F_X(X)$ is uniformly distributed on $[0,1]$ and (ii) the r.v. $G^{-1}(F_X(X)) = (G^{-1} \circ F_X)(X)$ has the distribution function G.*

Proof. By Theorem 2.3.3 the r.v. $F_X(X)$ has the distribution function $F_X \circ F_X^{-1}$. But since X is continuous its distribution function F_X is continuous, and hence $F_X \circ F_X^{-1}(x) = x$ for all $x \in [0,1]$ by (2.23), implying (i). Statement (ii) then follows from (i) and Theorem 2.3.1.

Remark 7 *It is worth noting that the assumption that X is continuous is essential for the validity of Theorem 2.3.4, because, say, if X takes only finite number of values, then so does the composition $F_X(X)$, which therefore can not be uniformly distributed.*

Finally, let us discuss random vectors and sums. Representing two r.v. as functions on the same standard probability space allows naturally to define their sums $(X+Y)(\omega) = X(\omega) + Y(\omega)$ (if X and Y are defined only via their distributions, the sense of these combinations is rather obscure) and to observe the basic property of the linearity of the expectation. Namely, if they both are defined on the standard probability space, it follows directly from (2.26), that for any real numbers a, b one has

$$\mathbf{E}(aX + bY) = a\mathbf{E}(X) + b\mathbf{E}(Y).$$

However, as in the discrete case, the possibility to define two r.v. on a single probability space is linked with their coupling, i.e. with specifying certain joint probabilities

$$\mathbf{P}_{X,Y}(A \times B) = \mathbf{P}(X \in A, Y \in B)$$

for intervals A, B. Once such a distribution is defined, it specifies a structure of a geometric probability space $(\mathbf{R}^2, \mathcal{B}, \mathbf{P}_{X,Y})$ on the plane, where both r.v. X and Y can be nicely defined as the coordinate functions $X(x,y) = x, Y(x,y) = y$ implying the linearity of the expectation.

Remark 8 * *Suppose X and Y are specified as functions on a standard probability space (say, by Theorem 2.3.1). Then one can define the r.v. X, Y simultaneously on the product probability space $[0,1] \times [0,1]$, being the square on a plane (with the probabilities of the Borel subsets being given by* *their geometric areas), as* $X(\omega_1, \omega_2) = X(\omega_1)$ *and* $Y(\omega_1, \omega_2) = Y(\omega_2)$*. This definition specifies the independent coupling. Moreover, it makes it clear that the expectation*

$$\mathbf{E}(XY) = \int_0^1 \int_0^1 X(\omega_1)Y(\omega_2)\, d\omega_1\, d\omega_2$$

decomposes into the product $\mathbf{E}(X)\mathbf{E}(Y)$ *by the Fubbini's theorem.*

For conclusion, let us summarize the main issues of this section. Random variables can be defined either by their distribution function F_X or as (usual) functions on a probability space. As such a space one can always choose the standard probability space $[0, 1]$ with the probability being defined as the length. The connection between the tho representations is given by equation (2.19). In case of continuous r.v. X, its distribution function has form (2.3), and the expectation of any function ϕ of X is given by the two equivalent expressions in (2.28). Furthermore, using generalized inverse functions, one can express any r.v. as a function of the standard uniform one (giving a universal tool for the simulation of r.v.) and define explicit functional transformations between any given continuous random variables.

Chapter 3

Volatility - spread - risk

3.1 Variance and correlation

Trying to describe a r.v. X by some simple characteristics, one is naturally led to look for its location statistics and its spread, i.e. one looks for a point d, where the spread of the deviation $\mathbf{E}[(X-d)^2]$ is minimal and then assesses this spread.

As by the linearity of the mean

$$\mathbf{E}[(X-d)^2] = \mathbf{E}(X^2) - 2d\mathbf{E}(X) + d^2,$$

one easily finds that the minimum is attained when $d = \mathbf{E}(X)$ (check it!), yielding a characterization of the expectation as the location statistics. The minimum itself

$$Var(X) = \mathbf{E}[(X - \mathbf{E}(X))^2] = \mathbf{E}(X^2) - [\mathbf{E}(X)]^2 \tag{3.1}$$

(check that the two expressions coincide!) is called the *variance* of X and constitutes the second basic characteristics of a r.v. describing its spread statistics. In application, the variance often represents a natural measure of risk, as it specifies the average deviation from the expected level of gain, loss, win, etc. A variance is measured in square units, and the equivalent spread statistics measured in the same units as X is the *standard deviation* of X:

$$\sigma = \sigma_X = \sqrt{Var(X)}.$$

For instance, if $\mathbf{1}_A$ is the indicator function on a probability space, then $\mathbf{E}(\mathbf{1}_A) = \mathbf{P}(A)$ and $\mathbf{1}_A = (\mathbf{1}_A)^2$ so that

$$Var(\mathbf{1}_A) = \mathbf{E}(\mathbf{1}_A) - [\mathbf{E}(\mathbf{1}_A)]^2 = \mathbf{P}(A)[1 - \mathbf{P}(A)].$$

In particular, for a Bernoulli r.v. X taking values 1 and 0 with probabilities p and $1-p$ (in other words X is Binomial $(1,p)$) it implies

$$Var(X) = p(1-p). \tag{3.2}$$

The simplest measure of the dependence of two r.v. X, Y (forming a random vector, i.e with the joint probabilities specified) is supplied by their *covariance*:

$$Cov(X,Y) = \mathbf{E}[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))]$$

or equivalently by their *correlation coefficient*

$$\rho = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

Exercise 3.1.1 *Convince yourself that $Cov(X,X) = Var(X)$ and that*

$$Cov(X,Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y). \tag{3.3}$$

If $Cov(X,Y) = 0$, one says that the r.v. X and Y are *uncorrelated.* As it follows from (2.12), if X and Y are independent, then they are uncorrelated. The converse statement does not hold true.

However, it is possible to show (we will not go into detail here) that it does hold for Gaussian r.v., i.e. two (joint) Gaussian r.v. are uncorrelated if and only if they are independent. This implies, in particular, that the dependence structure of, say, two standard Gaussian r.v. can be practically described by a singe parameter, their correlation. This property makes Gaussian r.v. particularly handy when assessing dependence, as is revealed, say, by the method of Gaussian copulas, see Section 3.6.

The linearity of the expectation implies that for a collection of r.v. $X_1, ..., X_k$ one has

$$Var(\sum_{i=1}^k X_i) = \sum_{i=1}^k Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j).$$

In particular, if r.v. $X_1, ..., X_k$ are pairwise uncorrelated, for example if they are independent, the variance of their sum equals the sum of their variances:

$$Var(X_1 + ... + X_k) = Var(X_1) + ... + Var(X_k). \tag{3.4}$$

This linearity is very useful in practical calculations. For example, if X is Binomial (n,p) r.v., then

$$X = \sum_{i=1}^n X_i,$$

where X_i are independent Bernoulli trials, or independent Binomial $(1, p)$ r.v., so that by (3.2) and by linearity one gets

$$Var(X) = np(1-p). \tag{3.5}$$

By (2.28), for a continuous r.v. X with the probability density function f, the variance can be calculated as

$$Var(X) = \int [x - \mathbf{E}(X)]^2 f_X(x)\, dx. \tag{3.6}$$

As an example, let us calculate the variance for a normal r.v. To begin with, let X be $N(0,1)$. Then by (3.6) and (2.5) (taking into account that the expectation of X vanishes and using integration by parts)

$$Var(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \exp\left\{-\frac{1}{2}x^2\right\} dx$$

$$= -\frac{1}{\sqrt{2\pi}}\, x \exp\left\{-\frac{1}{2}x^2\right\}\Bigg|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}x^2\right\} dx = 1,$$

as the second term is the integral of a probability density function. It now follows from (2.8) that the variance of a $N(\mu, \sigma^2)$ normal r.v. equals σ^2. Thus two parameters in the notation $N(\mu, \sigma^2)$ for a normal r.v. stand for its mean and variance.

3.2 Waiting time paradox

As an illustration of an application of variance, let us discuss the so called *waiting time paradox*. One could be very annoyed by regularly waiting an hour for a bus at a bus stop, where buses are scheduled to run at 20-minutes intervals. However, only if the buses run precisely at 20 minutes intervals, your average waiting time (when you arrive at the stop at a random moment) will be 10 minutes. Of course this effect is relevant not only for buses, but for a variety of situations, when waiting times and queues are to be handled (i.e. supermarket cash points, internet sites, mobile phone networks, etc). The full development belongs to a highly applied domain of probability, called the queueing theory. Here we shall only sketch a (not quite rigorous) argument leading to the calculation of the average waiting times.

To simplify the matter, let us assume that possible intervals between busses take only finite number of values, say $T_1, ..., T_k$, with certain probabilities $p_1, ..., p_k$. The average interval (usually posted on a timetable) is therefore

$$\mathbf{E}(T) = \sum_{j=1}^{k} p_j T_j.$$

We are interested in the average waiting time. Suppose first that all possible intervals follow each other periodically in a strictly prescribed order, say $T_1, T_2, ..., T_k, T_1, T_2, ...$ What will be your average waiting time, if you arrive at the stop at a random time t uniformly distributed on the interval $[0, T_1 + ... + T_k]$. If $T_1 + ... + T_{j-1} < t \le T_1 + ... + T_j$ (i.e.your arrive at the j th period between busses), then you will wait the time $W = T_1 + ... + T_j - t$. Taking expectation with respect to the uniform distribution of t yields

$$\mathbf{E}_{T_1,...,T_k}(W) = \frac{1}{T_1 + ... + T_k} \left(\int_0^{T_1} (T_1 - t)\, dt \right.$$

$$\left. + \int_{T_1}^{T_1+T_2} (T_1 + T_2 - t)\, dt + ... + \int_{T_1+...+T_{k-1}}^{T_1+...+T_k} (T_1 + ... + T_k - t)\, dt \right)$$

$$= \frac{1}{T_1 + ... + T_k} \left(\int_0^{T_1} t\, dt + \int_0^{T_2} t\, dt + ... + \int_0^{T_k} t\, dt \right) = \frac{T_1^2 + ... + T_k^2}{2(T_1 + ... + T_k)}.$$

Similarly, if during a period of time there were m_1 intervals of the length T_1, m_2 intervals of the length T_2,..., m_k intervals of the length T_k (in any order), and you arrive uniformly randomly on this period, then your average waiting time will be

$$\mathbf{E}_{T_1,...,T_k;m_1,...,m_k}(W) = \frac{m_1 T_1^2 + ... + m_k T_k^2}{2(m_1 T_1 + ... + m_k T_k)} = \frac{q_1 T_1^2 + ... + q_k T_k^2}{2(q_1 T_1 + ... + q_k T_k)},$$

where $q_j = k_j/(k_1 + ... + k_m)$ denote the frequencies of the appearance of the intervals T_j. But frequencies are equal approximately to probabilities (and approach them as the number of trials go to infinity), so that approximately, if the intervals T_j occur with probabilities p_j, the above formula turns to

$$\mathbf{E}(W) = \frac{q_1 T_1^2 + ... + q_k T_k^2}{2(q_1 T_1 + ... + q_k T_k)} = \frac{\mathbf{E}(T^2)}{2\mathbf{E}(T)}, \tag{3.7}$$

which can be equivalently rewritten as

$$\mathbf{E}(W) = \frac{1}{2} \left[\mathbf{E}(T) + \frac{Var(T)}{\mathbf{E}(T)} \right]. \tag{3.8}$$

Hence, as we have expected, for a given average interval time $\mathbf{E}(T)$, the average waiting time can be be arbitrary large depending on the variance of the interval lengths.

Similar arguments can be used in assessing traffic flows. To wit, suppose n cars are driven along the race track, formed as a circumference of radius 1 km, with the speeds v_1,...,v_n kilometers per hour, and a speed camera is placed at some point that registers the speeds of passing cars and then calculates the average (arithmetic mean) speed from all the observed ones. Would the average be $(v_1 + ... + v_n)/n$? Of course not! In fact, during a time T, the cars will cover $v_1T, ..., v_nT$ circumferences respectively, so that the camera will register $(v_1 + ... + v_n)T$ cars with the average speed being

$$\frac{v_1^2 + ... + v_n^2}{v_1 + ... + v_n}.$$

If the speed of a car is a r.v. with a given distribution, then this would turn to the expression $\mathbf{E}(V^2)/\mathbf{E}(V)$, which is similar to (3.7).

3.3 Hedging via futures

Futures and forwards represent contracts to buy or sell a commodity or an asset by a fixed price on a prescribed date in future.

Futures markets can be used to hedge the risk, i.e. to neutralize it as far as possible. The plan of this section is the following: 1) we shall show how this kind of hedging works on a simple numerical example; 2) deduce the main formula for the optimal hedge ratio; 3) introduce the main idea (arbitrage) underlying the pricing of the futures.

Assume that on the 1st of April an oil producer negotiated a contract to sell 1 million barrels of crude oil on the 1st of August by the market price that will form on this latter date. The point is that this market price is unknown on the 1st of April. Suppose that on the 1st of April the crude oil futures price (on a certain Exchange) for August delivery is \$19 per barrel and each futures contract is for the delivery of 1000 barrels. The oil producer can hedge its risk by *shorting* 1000 futures contracts, i.e. by agreeing to sell $1000 \times 1000 = 10^6$ barrels on the 1st of August by the price \$19 per barrel. Let us see what can then happen. Suppose the price for crude oil will go down and on the 1st of August become, say, \$18 per barrel. Then the company will realize 18 millions from its initial contract. On the other hand, the right to sell something by \$19 per barrel when the actual price is \$18 per barrel means effectively the gain of \$1 per barrel, i.e. the company will realize from its futures contract the sum of 1 million. The total gain then will equal $18 + 1 = 19$ millions. Suppose now the opposite scenario takes place, namely, the price for crude oil will go up and on the 1st of August become, say, \$21 per barrel. Then the company will realize 21 millions from its initial contract. On the other hand, the obligation to sell something by \$19 per barrel when the actual price is \$21 per barrel means effectively the loss of \$2 per barrel, i.e. the company will loose from its futures contract the sum of 2 million. The total gain then will equal $21 - 2 = 19$ millions. In both cases the total gain of the company is the one obtained by selling its oil according to the futures price for the August delivery. Thus the risk is totally eliminated.

Such a perfect hedging was possible, because the commodity underlying a futures contract has been the same as the commodity whose price is being hedged. However, this is not always possible, so that one can only hedge the risk using futures on a related commodity or an asset (say to hedge a contract on a certain product of crude oil by the futures on the oil itself). To assess (and to minimize) risk the correlations between these commodity prices should be then taken into account. Suppose N_A units of assets are to be sold at time t_2 and a company is hedging its risk at time $t_1 < t_2$ by shorting futures contract on N_F units of a similar (but not identical asset). The *hedge ratio* is defined as

$$h = N_F/N_A.$$

Let S_1 and F_1 be the (known at time t_1) asset and futures prices at the initial time t_1 (let us stress that F_1 is the price, at which one can agree at time t_1 to deliver a unit of the related asset on the time t_2), and let S_2 and F_2 be the (unknown at time t_1) asset and futures prices at the time t_2 (F_2 is the price, at which one can agree at time t_2 to deliver a unit of the underlying asset at this time t_2, so it is basically equal to the asset price at time t_2). The gain of the company at time t_2 becomes

$$Y = S_2 N_A - (F_2 - F_1) N_F = [S_2 - h(F_2 - F_1)] N_A = [S_1 + (\Delta S - h\Delta F)] N_A,$$

where

$$\Delta S = S_2 - S_1, \quad \Delta F = F_2 - F_1.$$

We aim at minimizing the risk, i.e. the spread of this r.v. around its average. In other words we are aiming at minimizing the variance $Var(Y)$. As S_1 and N_A are given, it is equivalent to minimizing

$$Var(\Delta S - h\Delta F).$$

Let σ_S, σ_F and ρ denote respectively the variance of ΔS, the variance of ΔF and the correlation between these r.v. Then

$$Var(\Delta S - h\Delta F) = (\sigma_S^2 + h^2\sigma_F^2 - 2h\rho\sigma_S\sigma_F).$$

Minimizing this quadratic expression with respect to h one obtains (check!) that the minimum is attained on the value

$$h^\star = \rho\frac{\sigma_S}{\sigma_F}, \tag{3.9}$$

which is the basic formula for the it optimal hedge ratio.

In practice, the coefficients σ_S, σ_F and ρ are calculated from the historical behavior of the prices using the standard statistical estimators (see Section 4.2).

Finally, for completeness, let us give a rough idea on how futures contracts are priced. Let S_0 denote the price of this asset at time t_0 and let r denotes the risk free rates, with which one can borrow and/or lend money (in practice rates r can be variable, but we only consider a simple situation with fixed rates). The futures price F to deliver this asset after a time T (called *time to maturity*) from the initial time t_0 should be equal to

$$F = S_0 e^{rT}. \tag{3.10}$$

Otherwise, arbitrage opportunities arise that would drive the price back to this level. In fact, suppose $F > S_0 e^{rT}$ (opposite situation is considered similarly). Then one can buy the asset by the price S_0 and short a future contract on it (i.e. enter into the agreement to sell an asset at time $t_0 + T$). Realizing this contract at time $t_0 + T$ would yield the sum F, which, when discounted to the present time equals Fe^{-rT}. Since this is greater than S_0, one gets a free income.

3.4 Black-Sholes option pricing

We shall sketch here the celebrated Nobel price winning Black-Sholes option pricing theory taking for granted the so called *risk-neutral evaluation* principle.

Thinking about stock prices as a result of a large number of influences of various sizes and sources, one naturally comes to the analogy with a small particle put in a liquid and moving under a constant bombardment of the immense number of the molecules of the liquid. The erratic behavior of such a particle was observed and protocolled by R. Brown and is called the

Brownian motion. The analogy of share prices with the Brownian motion was exploited by Bachelier at the dawn of the 20th century. Referring to the Central Limit Theorem that states that the sum of independent r.v. can be approximated by normal laws under rather general assumptions (see Section 5.2), Bachelier assumed the prices S_T of a stock at time T to be normally distributed. However, the assumption of normality (distribution on all numbers, positive and negative)) clearly contradicts the crucial positivity property of prices. Later on, the Bachelier model was improved by assuming that the rates of changes are normal, i.e. the logarithm of prices $\ln S_T$ is normal, in other words S_T themselves are log-normal. This was the starting point for the Black-Sholes theory.

Suppose now that a r.v. X_t can be considered as the sum of independent inputs acting at shorter times, i.e. one can write

$$X_1 = X_t^1 + ... + X_t^n, \quad t = 1/n,$$

for any integer n with i.i.d r.v. X_t^i, distributed like X_t. By the linearity of the expectation and variance for independent r.v. it follows that

$$\mathbf{E}X_1 = n\mathbf{E}(X_t) = \frac{1}{t}\mathbf{E}(X_t), \quad Var(X_1) = n\,Var(X_t) = \frac{1}{t}\,Var(X_t),$$

so that

$$\mathbf{E}X_t = t\mathbf{E}(X_1), \quad Var(X_t) = t\,Var(X_1),$$

leading to the conclusion that the expectation and variance of such X_t should depend linearly on time. Applying this to $\ln S_T$ allows to make our model for stock prices more precise by assuming that $\ln S_T$ is normal $N(\mu t, \sigma^2 t)$ and thus leaving only two parameters μ, σ to be specified (say, by observed data, see Section 4.2). The quadratic variation σ in this model is usually referred to as the *volatility* of the stock.

A standard *European call option* gives to the holder the right (but unlike a futures contract, not the obligation) to by a stock by certain time T in the future, called the *expiration date* or *maturity*, for a fixed price K, called the *strike price* or the *exercise price.* If by the time T the stock price would not rise above K, there would be no reason for an option holder to exercise this right turning his/her gain to zero. On the other hand, if by the time T the stock price would rise above K, an option holder would exercise his/her right yielding the payoff of $S_T - K$. Hence the price of an option at the expiry date is $\max(S_T - K, 0)$. We are interested in a fair price of an option at the initial time 0.

Suppose that r is the risk free interest rate in our model, i.e. one can borrow money at this rate (borrowing now the amount V, you have to return $e^{rt}V$ after time t, or equivalently, the discounted present cost of the amount V payed at the time t is $e^{-rt}V$). The Black-Sholes *risk-neutral evaluation principle* states that the fair price of an option does not depend on the parameter μ in the normal $N(\mu t, \sigma^2 t)$ model for $\ln S_T$, so that it can be calculated as the discounted expectation of the price at time T assuming the *risk neutral* distribution for S_T, i.e. that it is log-normal with $\mathbf{E}(S_T) = e^{rT}S_0$. The deduction of this principle is rather deep and is beyond the scope of the present exposition. Taking this principle for granted yields for the fare price of an option the value

$$c = e^{-rT}\hat{\mathbf{E}}[\max(S_T - K, 0)],$$

where $\hat{\mathbf{E}}$ means the expectation with respect to the risk-neutral distribution. Applying Proposition 2.2.1 yields

$$c = e^{-rT}[S_0 e^{rT}\Phi(d_1) - K\Phi(d_2)] = S_0\Phi(d_1) - Ke^{-rT}\Phi(d_2), \tag{3.11}$$

where

$$d_1 = \frac{\ln \hat{\mathbf{E}}(S_T)/K) + \sigma^2 T/2}{\sigma\sqrt{T}} = \frac{\ln(S_0/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}},$$

$$d_2 = \frac{\ln \hat{\mathbf{E}}(S_T)/K) - \sigma^2 T/2}{\sigma\sqrt{T}} = \frac{\ln(S_0/K) + (r - \sigma^2/2)T}{\sigma\sqrt{T}}.$$

This is the celebrated Black-Sholes formula.

3.5 Markowitz portfolio optimization

We shall sketch here another Nobel prize wining idea. Let a market contain n securities with prices $S_0^1, ..., S_0^n$ at the initial moment of time. Their prices S_T^i at the future time T are considered as r.v. The returns of the securities in time T are defined as the ratios $R_i = R_i(T) = S_T^i/S_0^i$. Assume their means and covariances are known (or have been estimated):

$$\mathbf{E}(R_i) = \mu_i, \quad Cov(R_i, R_j) = \sigma_{ij}.$$

An investor with initial wealth $x > 0$ is supposed to choose ϕ_i securities of type i, $i = 1, ..., n$, complying with the budget constraint

$$\sum_{i=1}^n \phi_i S_0^i = x.$$

In terms of the *portfolio vector* $\pi = (\pi_1, ..., \pi_n)$ defined by $\pi_i = \phi_i S_0^i / x$, the budget constraint rewrites as the normalizing condition $\sum_{i=1}^n \pi_i = 1$. Once the choice of ϕ_i (or equivalently π_i) is made, the final wealth of the investor in time T will constitute

$$X(T) = \sum_{i=1}^n \phi_i S_T^i,$$

and the total portfolio return will be equal to

$$R = \frac{X(T)}{x} = \sum_{i=1}^n \frac{\phi_i S_0^i}{x} R_i = \sum_{i=1}^n \pi_i R_i.$$

The mean and variance of the portfolio return are then given by

$$\mathbf{E}(R) = \sum_{i=1}^n \pi_i \mu_i, \quad Var(R) = \sum_{i=1}^n \sum_{j=1}^n \pi_i \sigma_{ij} \pi_j.$$

Choosing a portfolio an investor aims at maximizing its return. However, possible risk should also be taken into account. The basic Markowitz idea was to look for a balance between the portfolio mean return and the risk measured by the portfolio variance. This leads to the two closely connected problems: finding a portfolio with a minimal variance for a given level of portfolio return or finding a portfolio with the maximum return for a given level of portfolio variance. These are the problems of quadratic programming. A variety of methods can be used for their effective numeric solutions.

3.6 Credit risk; Gaussian copulas

Default time τ for a company are usually assessed (or stored) by the *cumulative probabilities of survival*

$$V(t) = \mathbf{P}(\tau > t)$$

(no default till time t). It is often assumed that the change rate of V is proportional to V, i.e.

$$\frac{dV(t)}{dt} = -\lambda(t) V(t), \tag{3.12}$$

with a positive function λ, called *default intensities* or *hazard rates*. Approximately, for small s

$$\lambda(t) \sim \frac{V(t) - V(t+s)}{V(t)} = \mathbf{P}(\tau \in [t, t+s] | \tau > t),$$

so that the intensity specifies the conditional default probabilities. From (3.12) one deduces that

$$V(t) = \exp\{-\int_0^t \lambda(s)\, ds\},$$

so that the distribution function of the r.v. τ equals

$$F_\tau(t) = \mathbf{P}(\tau \leq t) = 1 - \exp\{-\int_0^t \lambda(s)\, ds\}. \tag{3.13}$$

Thus the distribution of τ is a time non-homogeneous extension of the exponential distribution.

If a firm goes bankrupt, its creditors file claims against its assets to recover some of its dept. The *recovery rate* R for a bond is defined as the bond's market price immediately after default as a percent of its face value. It is reasonable to assume (at least for rough estimates) that the possibility of a default is the only reason for higher return rates of the bonds issued by a firm as compared to the corresponding risk free bonds (say, issued by the government). Consequently, if s denotes the additional interest on a firm's bond as compared with an equivalent risk free bond, and if p denotes the default probability per year so that $p(1-R)$ is the average rate of loss due to default, one should have

$$s = p(1-R),$$

yielding an important estimate for the default probability

$$p = s/(1-R) \tag{3.14}$$

via the prices of bonds the company has issued.

As we observed already in Chapter 1 for assessing possible defaults and hence pricing the corresponding credit derivatives (like CDS) the structure of the dependence between various firms defaults becomes crucial. This requires specifying joint default probabilities, which practically are difficult to assess. Only for Gaussian r.v. the dependence structure can be incorporate in a few key parameters (say, for two standard Gaussian r.v., only one parameter, their correlation coefficient, specifies the joint distribution uniquely). As the default times are principally different from Gaussian r.v. (in particular, because they are positive), the idea is to transform them into Gaussian and then to specify the dependence structure on these Gaussian images. To do this, Theorem 2.3.4 is applied with $G = \Phi$ being the distribution function (2.7) of a standard normal r.v. implying that if τ is a random time to default for a firm and the distribution function of τ is F_τ (say, given by equation (3.13)), then

$$X = \Phi^{-1}(F_\tau(\tau))$$

is $N(0,1)$ Gaussian r.v. For n firms with default times τ_i and their distribution functions F_i, the *Gaussian copula* assumption states that the r.v. $X_i = \Phi^{-1}(F_i(\tau_i))$ are mutually Gaussian with the correlation coefficients between X_i and X_j being given numbers ρ_{ij}.

Avoiding here a discussion of multivariate distributions (Gaussian vectors), let us introduce only the simplest model of the Gaussian copula approach, namely the so called *one factor model*, where one assumes that

$$X_i = a_i M + \sqrt{1-a_i^2} Z_i, \tag{3.15}$$

where $M, Z_1, ..., Z_n$ are independent normal $N(0,1)$ r.v. and a_i are constants from $[-1,1]$. Here M stands for the common factor affecting all firms and Z_i stand for their individual characteristics.

Exercise 3.6.1 *Check that each X_i is also $N(0,1)$ and $Cov(X_i, X_j) = \rho_{X_i,X_j} = a_i a_j$ for $i \neq j$.*

Consequently the event $(\tau_i \leq T)$ (i th firm defaults before time T) can be described as

$$X_i = a_i M + \sqrt{1-a_i^2} Z_i \leq \Phi^{-1}(F_i(T))$$

or equivalently as

$$Z_i \leq \frac{\Phi^{-1}(F_i(T)) - a_i M}{\sqrt{1-a_i^2}}.$$

Hence, as Z_i is $N(0,1)$ r.v., for the probability $D_i(T|M)$ of the ith firm default conditional on the value of the factor M one has

$$D_i(T|M) = \Phi\left(\frac{\Phi^{-1}(F_i(T)) - a_i M}{\sqrt{1-a_i^2}}\right).$$

In particular, if all $F_i = F$ are the same (firms are chosen from a certain common class) and all correlations are the same, say $a_i = \sqrt{\rho}$, then all $D_i(T|M)$ coincide and equal

$$D(T|M) = \Phi\left(\frac{\Phi^{-1}(F(T)) - \sqrt{\rho} M}{\sqrt{1-\rho}}\right). \tag{3.16}$$

Practically, for a large portfolio of similar assets, this probability estimates the frequency of defaults, i.e. the ratio of the number of assets belonging to firms that defaulted to time T to the total number of assets of the portfolio. For risk assessment one is often interested in the worst scenario. Choosing a confidence level C (say, 0.999, or 99.9%), one is interested in a bound for losses that can occur C almost certainly, i.e. with probability C (99.9% say). Since M is $N(0,1)$ normal, $M \geq m$ with probability $1-\Phi(m)=\Phi(-m)$, so that $M \geq -\Phi^{-1}(C)$ with probability C implying that with this probability the frequency of defaults over T years will be less than

$$V(C,T)=\Phi\left(\frac{\Phi^{-1}(F(T))+\sqrt{\rho}\Phi^{-1}(C)}{\sqrt{1-\rho}}\right). \tag{3.17}$$

This is so called *Vasicek* formula.

Let us denote by $P(k,T|M)$ the probability of k defaults by time T conditioned on M. When M is fixed the default probabilities are independent. Hence, the binomial distribution can be applied to yield

$$P(k,T|M)=\frac{n!}{(n-k)!k!}[D(T|M)]^k[1-D(T|M)]^{n-k} \tag{3.18}$$

with $D(T|M)$ given by (3.16). To get unconditional probabilities it remains to integrate over the normally distributed factor M. These formulas become the standard market tool for valuing default ratios and hence pricing the corresponding n th-to-default CDS (credit default swaps).

Chapter 4

The law of large numbers

4.1 Markov's inequality and the law of large numbers

The law of large numbers (in its various forms) constitutes one of the oldest result in probability theory. We shall prove only its weak version, which is a consequence of the classical Markov and Chebyshev inequalities. As applications, we shall discuss the basic statistic estimator for a volatility and the optimal betting system.

The importance of the following elementary inequalities is difficult to overestimate.

Theorem 4.1.1 Markov's Inequality: *If X is a non-negative random variable, then for any $\epsilon > 0$*

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

Chebyshev's Inequality: *For any $\epsilon > 0$ and a random variable Y*

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

Proof. Evident inequalities

$$\mathbf{E}X \geq \mathbf{E}(X\mathbf{1}_{X\geq\epsilon}) \geq \epsilon\mathbf{E}\mathbf{1}_{X\geq\epsilon} = \epsilon\mathbf{P}(X \geq \epsilon)$$

imply Markov's one. Applying Markov's Inequality to $X = |Y - \mathbf{E}Y|^2$ yields Chebyshev's one.

Theorem 4.1.2 *(***Weak law of large numbers***) If $X_1, X_2, \ldots$ is a collection of i.i.d. random variables with $\mathbf{E}(X_j) = m$ and $Var(X_j) < \infty$, then the means $(X_1 + \ldots + X_n)/n$ approaches m as $n \to \infty$ in the following sense: for any $\epsilon > 0$*

$$\mathbf{P}\left(|\frac{X_1 + \ldots + X_n}{n} - m| > \epsilon \right) \to 0, \tag{4.1}$$

as $n \to \infty$.

Proof. By (3.4)

$$Var\left(\frac{X_1 + \ldots + X_n}{n} - m\right) = Var\frac{(X_1 - m) + \ldots + (X_n - m)}{n} = \frac{VarX_1}{n},$$

which of course tends to zero as $n \to \infty$. By Chebyshev's inequality

$$\mathbf{P}\left(|\frac{X_1 + \ldots + X_n}{n} - m| > \epsilon \right) \le \frac{Var(X_1)}{n\epsilon^2},$$

implying (4.1).

As the simplest example, we can deduce that if Y_n denotes the number of successes in a series of n independent Bernoulli trials with the probability of success in each trial being p, the frequency of successes Y_n/n converges to p in the sense that for any $\epsilon > 0$

$$\mathbf{P}\left(|\frac{Y_n}{n} - p| > \epsilon \right) \to 0,$$

as $n \to \infty$. This fact is sometimes called the *Golden Theorem.*

The convergence in Theorem 4.1.2 can be essentially improved yielding the celebrated Kolmogorov's *strong law of large numbers* that states that the average sum $(X_1 + \ldots + X_n)/n$ converges to m as $n \to \infty$ for almost all realizations of the sequences X_n. 'For almost all' means *with probability one*, i.e the sets of converging sequences has the full probability one (see Remark below).

Remark 9 * *In order to talk about convergence of a sequence of r.v., say $(X_1 + \ldots + X_n)/n$, with probability one, one has to define all X_n on a single probability space. Suppose for simplicity that X_n are independent Bernoulli trials, i.e. X_n take values 1 and 0 with probabilities p and $1 - p$ respectively. As a common probability space for the sequence X_n, $n = 1, 2, \ldots$, one can take the set Ω of all sequences of zero's and one's. Probability law on Ω is uniquely specified by the requirement that the sets Ω_k^0 of sequences with* *zero on the k th place have probability $1/2$ for all k. The important point is that the set of all sequences Ω is not countable (this is in fact a natural example of the appearance of an uncountable state when working with discrete probabilities). To see this one has to note that each sequence of zero's and one's specifies a number on the interval $[0,1]$ by its binary expansion. This correspondence even becomes one-to-one if one excludes the countable subset of sequences (with vanishing probability) that have all one's starting from a certain index. Thus the set of all sequences is (essentially) in a one-to-one correspondence with the points of $[0,1]$. Moreover, under this correspondence the probability law on sequences introduced above corresponds to the standard probability measures on $[0,1]$ (just because the sets Ω_k^0 corresponds to the union of 2^{k-1} intervals of the length 2^{-k}. Thus the sequence X_n becomes naturally defined on the standard probability space, yielding precise meaning to the 'convergence with probability one' mentioned above.*

It is worth stressing the importance of the assumption of independence in both weak and strong laws of large numbers. In fact, suppose, say that all X_j are identical (note the crucial difference between 'identical' and 'identically distributed'!), i.e. $X_k = X_1$ for all k. Then the average Y_n/n equals X_1 and of course does not converge to any constant (as long as X_1 is not a constant itself).

4.2 Volatility and correlation estimators

A standard problem in applications of probability consists in estimating the mean, variance and correlations of random variables on the basis of their observed realizations during a period of time. Such estimates are routinely performed, say, by traders, for assessing the volatility of the stock market or a particular common stock. In its simplest mathematical formulation the problem is to estimate the mean and variance of a r.v. X, when a realization of a sequence $X_1, ..., X_n$ of independent r.v. distributed like X was observed. It is of course natural to estimate the expectation μ of X by its empirical mean

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This estimate is *unbiased* in the sense that $\mathbf{E}(\hat{\mu}_n) = \mu$ and *asymptotically effective* in the sense that $\hat{\mu}_n$ converges to μ by the law of large numbers. As the variance σ^2 is the expectation of $(X-\mu)^2$, the above reasoning suggests also the estimate for the variance in the form

$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2,$$

which is again unbiased and asymptotically effective. The problem with this estimate lies in utilizing the unknown μ. To remedy this shortcoming, it is natural to plug in the above given estimate instead of μ leading to the estimate

$$\hat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \hat{\mu}_n^2.$$

But here a surprise is awaiting. This estimate is no longer unbiased. In fact, as (by the i.i.d. property of X_i)

$$\mathbf{E}(\hat{\mu}_n^2) = \frac{1}{n}(\mathbf{E}(X^2) + (n-1)[\mathbf{E}(X)]^2) = \frac{1}{n}\sigma^2 + \mu^2,$$

one has

$$\mathbf{E}(\hat{\sigma}_2^2) = \sigma^2 + \mu^2 - (\frac{1}{n}\sigma^2 + \mu^2) = \frac{n-1}{n}\sigma^2.$$

So, to have unbiased estimate one has to take instead $\hat{\sigma}_2^2$ the estimate

$$\hat{\sigma}_3^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \hat{\mu}_n)^2.$$

Of course, for large n the difference between $\hat{\sigma}_2^2$ and $\hat{\sigma}_3^2$ disappears, and both these estimates are asymptotically effective.

Similarly, suppose we observe two sequences of i.i.d. r.v. $X_1, X_2, ...$, $Y_1, Y_2, ...$ distributed like X and Y respectively (let us stress that actually we observe the realizations of i.i.d. random vectors $(X_1, Y_1), (X_2, Y_2), ...$). An unbiased and asymptotically effective estimate for the covariance can be constructed as

$$\hat{Cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \hat{\mu}_n)(Y_i - \hat{\nu}_n),$$

where $\hat{\nu}_n = (Y_1 + ... + Y_n)/n$.

4.3 Optimal betting (Kelly's system)

A nice illustration of the law of large numbers presents the analysis of optimal betting.

Suppose you have an edge in betting on a series of similar trials, i.e. the average gain in each bet is positive (say, as you might expect by trading on Forex using some advanced strategy). Then you naturally expect to win in a long run. However, if you are going to put all your bankroll on each bet, you would definitely loose instead. An obvious idea is therefore to bet on a fraction of you current capital. What is the optimal fraction?

To answer this question assume that you are betting on a series of Bernoulli trials with the probability of success p, when you get m times the amount you bet. And otherwise, with probability $1-p$, you loose your bet. Consequently, if you invest one dollar, the expectation of you gain is mp dollars. Thus, assuming that you have an edge, is equivalent to assuming $mp > 1$, as we shall do now.

Let V_0 be your initial capital, and your strategy is to invest in each bet a fixed fraction α, $0 < \alpha < 1$, of your current bankroll. Let X_k denote the r.v. that equals m, if the k th bet is winning, and equals zero otherwise, so that all X_k are independent and identically distributed. Hence after the first bet your capital becomes $V_1 = (1-\alpha+\alpha X_1)V_0$, after the second bet it will be

$$V_2 = (1-\alpha+\alpha X_2)(1-\alpha+\alpha X_1)V_0,$$

and for any n your capital at time n will become

$$V_n = (1-\alpha+\alpha X_n)...(1-\alpha+\alpha X_1)V_0.$$

We are aiming at maximizing the ratio V_n/V_0, or equivalently its logarithm

$$\ln\frac{V_n}{V_0} = \ln(1-\alpha+\alpha X_n) + ... + \ln(1-\alpha+\alpha X_1).$$

Here the law of large numbers comes into play. Namely, according to this law, the average of the winning rates per bet

$$G_n = \frac{1}{n}\ln\frac{V_n}{V_0} = \frac{1}{n}[\ln(1-\alpha+\alpha X_n) + ... + \ln(1-\alpha+\alpha X_1)]$$

converges to the expectation

$$\begin{aligned}\phi(\alpha) &= \mathbf{E}\ln(1-\alpha+\alpha X_1)\\ &= p\ln(1-\alpha+\alpha m) + (1-p)\ln(1-\alpha).\end{aligned}$$

Therefore, to maximize the gain in a long run, one has to look for α that maximizes the function $\phi(\alpha)$. But this is an easy exercise in calculus yielding (check it!) that the maximum is attained at

$$\alpha^\star = \frac{pm-1}{m-1}. \tag{4.2}$$

This is the final formula we aimed at, known as Kelly's betting system formula. This is not however the full story, as we did not take into account the risk (we only used averages). A careful analysis (that we are not going into) allows to conclude that choosing the betting fraction α slightly smaller than $\alpha^\star$ one can essentially reduce the risk with only a slight decrease of the expected gain.

4.4 Playing on a stock market

We shall touch here only one aspect of financial trading, namely the *money management*, which by many active and successful traders is considered by the most crucial one.

As a first (possibly rather artificial) example, suppose you are trading on a stock that each week goes up 80% or down 60% with equal probabilities 1/2. You clearly have an edge in this game, as the expectation of your gain is

$$\frac{1}{2}80\% - \frac{1}{2}60\% = 10\%.$$

On the other hand, suppose you invest (and reinvest) your capital for many weeks, say n times. Each time your capital is multiplied either by 1.8 (success) or by 0.4 (failure) with equal probabilities. During a large period of time, you can expect the number of successes and failures to be approximately equal. Thus your initial capital V_0 would turn to

$$1.8^{n/2} 0.4^{n/2} V_0 = (0.72)^{n/2} V_0,$$

which quickly tends to zero as $n \to \infty$. So, do you really have an edge in this game?

The explanation to this paradoxical situation should be seen from the previous section. The point is that you should not invest all your capital at once (money management!).

To properly sort out the situation it is convenient to work in a more general setting. Namely, assume that the distribution of the rate return is given by a positive random variable R (i.e. in a one time investing your capital is multiplied by R) with a given distribution. Following the line of reasoning of the previous section, we can conclude that if you always invest the fraction α of your capital, then the average rate of return per investment would converge to

$$g(\alpha) = \mathbf{E} \ln(1 - \alpha + \alpha R).$$

We are interested in the maximum of this expression over all $\alpha \in [0, 1]$.

Theorem 4.4.1 *Assume* $\mathbf{E}(R) > 1$ *(i.e. you have an edge in the game). Then (i) if* $\mathbf{E}(1/R) \leq 1$ *(your edge is overwhelming), then* $\alpha = 1$ *is optimal and your maximum average return per investment equals* $g(1) = \mathbf{E} \ln(R)$*;*

(ii) if $\mathbf{E}(1/R) > 1$*, then there exists a unique* $\hat{\alpha} \in (0, 1)$ *maximizing* $g(\alpha)$*, which is found from the equation*

$$g'(\hat{\alpha}) = \mathbf{E} \frac{R - 1}{1 + (R - 1)\hat{\alpha}} = 0. \tag{4.3}$$

Proof. Since the second derivative

$$g''(\alpha) = -\mathbf{E} \left(\frac{R - 1}{1 + (R - 1)\hat{\alpha}} \right)^2$$

is negative, the function g is concave (i.e. its derivative is decreasing), and there can be at most one solution of equation (4.3) that necessarily would specify a maximum value of g. The assumption $\mathbf{E}(R) > 1$ is equivalent to saying that $g'(0) > 0$. In case (i) one has $g'(1) \geq 0$, so that the solution $\hat{\alpha}$ can lie only to the right of $\alpha = 1$ and hence the maximum of $g(\alpha)$ is attained at $\alpha = 1$. In case (ii), $g'(1) < 0$ implying the existence of the unique solution to (4.3) in $(0, 1)$.

Exercise 4.4.1 *Show that if* R *takes only two values* m_1 *and* m_2 *with equal probabilities (in the example at the beginning* $m_1 = 1.8$*,* $m_2 = 0.4$*), the condition* $\mathbf{E}(1/R) > 1$ *rewrites as* $m_1 + m_2 > 2m_1 m_2$*, and if it holds, the optimal* $\hat{\alpha}$ *equals*

$$\hat{\alpha} = \frac{1}{2(1 - m_1)} + \frac{1}{2(1 - m_2)}.$$

Exercise 4.4.2 *Specify the results of Theorem 4.4.1 in case when* R *is a log-normal r.v. (which is often assumed in stock pricing models).*

Chapter 5

Basic Limiting distributions

5.1 Generating functions and Poisson's limit; an epidemics model

For a r.v. X its *generating function* (called also the *probability generating function*) $G_X(s)$ is defined for non-negative s as

$$G_X(s) = \mathbf{E}(s^X).$$

Generating functions represent a useful tool for working with r.v. taking values in $\mathbf{Z}_+$ - the set of non-negative integers. In this case the above formula makes sense for all $s \in [-1,1]$ and by the definition of expectation rewrites as

$$G_X(s) = \sum_{n=0}^{\infty} p_n s^n, \tag{5.1}$$

where p_n stand for the probabilities $p_n = \mathbf{P}(X = n)$. From this expansion it follows that

$$p_n = \frac{1}{n!} G_X^{(n)}(0),$$

so that the probabilities p_n are uniquely specified by G_X. More generally, differentiating the defining formula for G_X yields

$$G_X^{(k)}(s) = \mathbf{E}[X(X-1)...(X-(k-1))s^{X-k}] = \sum_{n=k}^{\infty} n(n-1)...(n-k+1)s^{n-k}p_n,$$

implying

$$G_X^{(k)}(1) = \mathbf{E}[X(X-1)...(X-(k-1))].$$

This formula allows to express the moments $\mathbf{E}(X^k)$ of X in terms of the derivatives of the generating function. In particular, as

$$G'_X(1) = \mathbf{E}(X), \quad G''_X(1) = \mathbf{E}(X^2) - \mathbf{E}(X),$$

one obtains

$$\mathbf{E}(X) = G'_X(1), \quad Var(X) = G''_X(1) + \mathbf{E}(X) - [\mathbf{E}(X)]^2. \tag{5.2}$$

The following result explains, why generating functions are especially useful for studying the sums of i.i.d. r.v.

Theorem 5.1.1 *Let $X_1, X_2, ...$ be a sequence of independent $\mathbf{Z}_+$-valued r.v. with the generating functions $G_{X_i}(s)$. Then the generating function of the sum $Y_n = X_1 + ... + X_n$ equals the product*

$$G_{Y_n} = G_{X_1}(s)...G_{X_n}(s).$$

Proof.

$$G_{Y_n} = \mathbf{E}s^{X_1+...+X_n} = \mathbf{E}[s^{X_1}...s^{X_n}],$$

which by the independence of X_i rewrites as

$$\mathbf{E}(s^{X_1})...\mathbf{E}(s^{X_n}) = \prod_{i=1}^n G_{X_i}(s),$$

as required.

Examples. 1. If X is Bernoulli with the success probability p, then

$$G_X(s) = (1-p) + sp.$$

2. If X is p-Geometric, then

$$G_X(s) = \sum_{n=0}^\infty p(1-p)^n s^n = \frac{p}{1-s(1-p)}.$$

3. If X is c-Poisson, then

$$G_X(s) = \sum_{n=0}^\infty \frac{c^n s^n}{n!} e^{-c} = e^{c(s-1)}.$$

4. If X is Binomial (n,p), then X is the sum of n independent Bernoulli r.v. so that by Theorem 5.1.1

$$G_X(s) = [(1-p) + sp]^n.$$

Exercise 5.1.1 *Calculate the expectation and variance for the r.v. from the above examples using formula* (5.2).

The method of generating function is a natural tool for justifying the statement of Theorem 1.0.1. Namely let us show that the generating function of a Binomial (n,p) r.v. X converges to the generating function of a c-Poisson r.v. as $n \to \infty$ whenever p tends to zero in such a way that $np \to c$. In fact, since $\ln(1+x) = x(1+\alpha(x))$ with a function $\alpha(x)$ tending to zero as $x \to 0$ (Taylor's formula), one has

$$G_X(s) = [(1-p)+sp]^n = e^{n\ln[1+p(s-1)]} = e^{n(s-1)p(1+\alpha(p))}$$

with $\alpha(p) \to 0$ as $p \to 0$, and this tends to $e^{c(s-1)}$, which is the generating function of the Poisson r.v. with parameter c. Of course, this does not constitute a complete proof, as one still have to deduce from the point-wise convergence of the generating functions the convergence of all its derivatives at zero that specify the corresponding probability laws (which follows in fact from a general result of the complex analysis), but at least makes the result of Theorem 1.0.1 plausible.

We have observed in many examples the appearance of the sums of i.i.d. r.v. Hence the importance of Theorem 5.1.1 above. In many situations one encounters as well the sums of i.i.d. r.v. with a random number of terms. The next result shows how this case can be handled.

Theorem 5.1.2 *Let $X_1, X_2, ...$ be a sequence of i.i.d. $\mathbf{Z}_+$-valued r.v. each having the generating function G_X. And let N be another $\mathbf{Z}_+$-valued r.v., independent of all X_i, and having the generating function G_N. Then the generating function G_Y of the sum of a random number of term $Y = X_1 + ... + X_N$ equals the composition of G_N and G_X, i.e. $G_Y(s) = G_N(G_X(s))$.*

Proof. By the probability decomposition law

$$G_Y = \mathbf{E}s^{X_1+...+X_N} = \sum_{n=0}^{\infty} \mathbf{E}[s^{X_1+...+X_N}\mathbf{1}_{N=n}] = \sum_{n=0}^{\infty} \mathbf{E}[s^{X_1+...+X_n}\mathbf{1}_{N=n}].$$

Hence by the independence of X_i and N

$$G_Y(s) = \sum_{n=0}^{\infty} \mathbf{E}(s^{X_1})...\mathbf{E}(s^{X_n})\mathbf{E}(\mathbf{1}_{N=n}) = \sum_{n=0}^{\infty} [G_X(s)]^n \mathbf{P}(N=n),$$

which equals $G_N(G_X(s))$, as required.

Example. *An epidemics model.* Suppose a carrier can expose N contacts to infection and p denotes the probability that a contact actually leads to infection. One can speak, say, about AIDS expansion for people, or foot-and-mouth disease for the cattle. One can naturally model N as a Poisson r.v. with a certain parameter λ. Hence the number of infected individuals will be given by

$$\sum_{i=1}^{N} X_i,$$

where X_i is the Bernoulli r.v. with probability of success p. The generating function of the r.v. Y (and hence its distribution) can be found from Theorem 5.1.2.

5.2 Asymptotic normality (central limit theorem).

We shall explain here shortly why the normal r.v. are called 'normal', i.e. why they are so universal.

For continuous r.v. X, more handy than generating functions become other equivalent characteristics, namely the *moment generating function* M_X and the *Laplace transform* L_X defined as

$$M_X(t) = \mathbf{E}e^{Xt}, \quad L_X(t) = \mathbf{E}e^{-Xt}. \tag{5.3}$$

Clearly

$$M_X(t) = L_X(-t) = G_X(e^t).$$

From the definition it follows straightforwardly that the moments $\mathbf{E}(X^k)$ are found from the derivatives of M_X as

$$M_X^{(k)}(0) = \mathbf{E}(X^k),$$

and expanding the exponential function in the Taylor series yields

$$M_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{E}(X^k) \tag{5.4}$$

(hence the name 'moment generating function').

If Z is normal $N(\mu, \sigma^2)$, then by (2.15)

$$\mathbf{E}(e^Z) = \exp\{\mu + \sigma^2/2\}.$$

But then tZ is normal $N(t\mu, t^2\sigma^2)$ implying that

$$M_Z(t) = \mathbf{E}(e^{Zt}) = \exp\{t\mu + t^2\sigma^2/2\}. \tag{5.5}$$

Exercise 5.2.1 *Calculate the moment generating function of a θ-exponential r.v. and then deduce its expectation and variance. Answer: moments generating function is* $(1-\theta t)^{-1}$.

Similarly to Theorem 5.1.1 one shows that for a sequence $X_1, X_2, ...$ of independent r.v. with the moment generating functions $M_{X_i}(s)$, the moment generating function of the sum $Y_n = X_1 + ... + X_n$ equals the product

$$M_{Y_n} = M_{X_1}(s)...M_{X_n}(s). \tag{5.6}$$

For example, of interest in application is the r.v. X representing the sum of, say n, independent θ-exponential r.v. (total waiting time for n sequential independent events). The distribution of such X is called (θ, n)- *Gamma distribution.* By Exercise 5.2.1 and equation (5.6) its moment generating function equals $(1-\theta t)^{-n}$.

In the same way as we deduced above the Poisson limit theorem from the the limiting behavior of the generating function of the binomial law, one can deduce the so called *central limit theorem* (representing one of the major strongholds of the probability theory) that states that the sum of i.i.d. r.v. becomes *asymptotically normal* as the number of terms increases. If they are properly normalized, then the limiting distribution is the standard normal one. Namely, assume $X_1, X_2, ...$ are i.i.d. r.v. with $\mu = \mathbf{E}(X_i)$, $\sigma^2 = Var(X_i)$. Let

$$Y_n = \frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 - \mu}{\sigma\sqrt{n}} + ... + \frac{X_n - \mu}{\sigma\sqrt{n}}. \tag{5.7}$$

By linearity $\mathbf{E}(Y_n) = 0$ and $Var(Y_n) = 1$. The *central limit theorem* states that Y_n converges to the standard normal r.v. in the sense that

$$\lim_{n\to\infty} \mathbf{P}(a \le Y_n \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b \exp\{-\frac{x^2}{2}\}\, dx \tag{5.8}$$

for any $a < b$. To see why this result is plausible let us show that the moment generating function M_{Y_n} of Y_n converges to the moment generating function (5.5) (to transform this argument into a rigorous proof one would have of course to show that from this convergence one can deduce the convergence (5.8), which is beyond our scope). From (5.7) and using that the expectation of a product of i.i.d. r.v. equals the product of their expectations it follows that

$$M_{Y_n}(t) = \left[M_{(X_1-\mu)/\sigma\sqrt{n}}(t)\right]^n.$$

By (5.4)

$$M_{(X_1-\mu)/\sigma\sqrt{n}}(t) = 1 + \frac{t^2}{2n} + \frac{\alpha(n)}{n}$$

with a certain function α that tends to zero as $n \to \infty$. Hence

$$M_{Y_n}(t) = \left[1 + \frac{t^2 + \alpha(n)}{2n}\right]^n \to \exp(t^2/2),$$

as required.

Since Y_n is asymptotically $N(0,1)$ normal, the sum $Z_n = X_1 + ... + X_n$ is asymptotically $N(n\mu, n\sigma^2)$ normal and the mean Z_n/n is asymptotically $N(\mu, \sigma^2/n)$ normal. Consequently, approximately, as $n \to \infty$, one gets the following equation allowing to calculate the distributions of arbitrary sums of i.i.d. r.v.:

$$\mathbf{P}(Z_n \leq x) = \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right), \tag{5.9}$$

where Φ is the distribution function of a standard normal r.v.

Example. *Biased betting.* Suppose you are betting on independent Bernoulli trials with success probability slightly less than 1/2, say $p = 0.497$. Let Y_n be the total number of winning in n experiments (similar model would work in many situations, say when estimating the number of males or females born in a population in a given period). As p is close to 1/2 one can expect that $P_n = \mathbf{P}(Y_n \geq n/2)$ should be slightly less than 1/2. Can you guess an approximation, say for $n = 500000$? By the asymptotic normality, Y_n is approximately $N(np, np(1-p)) = N(248500, 354^2)$. Hence approximately

$$P_n = 1 - \mathbf{P}(Y_{500000} \leq 250000) = 1 - \Phi(\frac{250000 - 248500}{354}) = 1 - \Phi(4.24) = 0.00001.$$

'Slightly' less than 1/2 indeed!

5.3 Fat (or heavy) tails.

The results of the previous section can lead to a misleading impression that essentially all distributions met in real life are normal. In fact, the practical applications of normal laws often go far beyond the situations, where this application can be justified by the above given central limit theorem. This leads to superficial conclusions, based on what one can call the prejudice of normality.

Experiments with concrete statistical data give clear evidence that many real life distributions (including in particular stock prices) have quite different qualitative features than Gaussian. In particular, they are often *heavy-tailed*, i.e. unlike the *light tails* of Gaussian laws, where $\mathbf{P}(X > x)$ decreases exponentially as $t \to \infty$, the r.v. X with *heavy (or fat) tails* has a property that $\mathbf{P}(X > x)$ decreases as a power $x^{-\omega}$ as $x \to \infty$ for some $\omega > 0$. We are going to use the method of generating functions in order to show how heavy tailed distributions can appear as the limits of the sums of i.i.d. r.v. when their second moment is not finite. We shall reduce our attention to positive r.v. X (modeling, say, waiting times) such that

$$\mathbf{P}(X > x) \sim \frac{c}{x^\alpha}, \quad x \to \infty, \tag{5.10}$$

in other words $\lim_{x\to\infty} \mathbf{P}(X > x)\, x^\alpha = c$ with some $c > 0$, $\alpha \in (0, 1)$. As one easily sees, even the expectation $\mathbf{E}(X)$ is not defined (it equals infinity) in this cases, so that already the law of large numbers can not be applied. For positive r.v. it is usually more convenient to work with the Laplace transform defined by the second equation in (5.3), rather than with the moment generating function.

Proposition 5.3.1 *Let $X_1, X_2, ...$ be a sequence of positive i.i.d. r.v. with the probability density function $p(x)$, $x \in (0, \infty)$, such that*

$$\mathbf{P}(X_1 > x) = \int_x^\infty p(y)\, dy \sim \frac{c}{x^\alpha}, \quad x \to \infty, \tag{5.11}$$

with some $c > 0$, $\alpha \in (0, 1)$. Then the Laplace transforms $L_{Y_k}(t)$ of the normalized sums $Y_k = (X_1 + ... + X_k)/k^{1/\alpha}$ converge, as $k \to \infty$, to $\exp\{-c\beta_\alpha t^\alpha\}$ with a certain constant $\beta_\alpha > 0$.

Proof. First one notes that, for a bounded differentiable function $g : [0, \infty) \to [0, \infty)$ such that $g'(0) = 0$ and $\int_1^\infty g(y) y^{-(1+\alpha)}\, dy < \infty$, it follows from (5.11) that

$$\lim_{k\to\infty} k \int_0^\infty g\left(\frac{x}{k^{1/\alpha}}\right) p(x)\, dx = \alpha c \int_0^\infty g(y) y^{-(1+\alpha)}\, dy.$$

In fact, by linearity and approximation, it is enough to show this for the indicators $\mathbf{1}_{[a,\infty)}$ with any $a > 0$. But in this case

$$k \int_0^\infty \mathbf{1}_{[a,\infty)} \left(\frac{x}{k^{1/\alpha}}\right) p(x)\, dx = k \int_{ak^{1/\alpha}}^\infty p(x)\, dx \to \frac{c}{a^\alpha} = \alpha c \int_a^\infty y^{-(1+\alpha)}\, dy,$$

as required. Now one can follow the same arguments as for the central limit theorem above. Namely,

$$L_{Y_k}(t) = \left[L_{X_1}\left(\frac{t}{k^{1/\alpha}}\right)\right]^k = \left[\int_0^\infty \exp\{-\frac{xt}{k^{1/\alpha}}\} p(x)\, dx\right]^k$$

$$= \left[1 + \int_0^\infty (\exp\{-\frac{xt}{k^{1/\alpha}}\} - 1) p(x)\, dx\right]^k = \left[1 + \alpha c \int_0^\infty (e^{-tx} - 1)\frac{dx}{x^{1+\alpha}} \frac{1+\omega(k)}{k}\right]^k,$$

where $\omega(k) \to 0$, as $k \to \infty$. Consequently, since by the change of variables

$$\int_0^\infty (e^{-tx} - 1)\frac{dx}{x^{1+\alpha}} = t^\alpha \int_0^\infty (e^{-x} - 1)\frac{dx}{x^{1+\alpha}},$$

one concludes that

$$L_{Y_k}(t) \to \exp\{-c\beta_\alpha t^\alpha\}$$

as required with $\beta_\alpha = \alpha \int_0^\infty (1 - e^{-x}) x^{-(1+\alpha)}\, dx$.

Remark 10 * *By the integration by parts*

$$\beta_\alpha = \alpha \int_0^\infty (1 - e^{-x})\frac{dx}{x^{1+\alpha}} = \int_0^\infty x^{-\alpha} e^{-x}\, dx,$$

so that $\beta_\alpha = \Gamma(1-\alpha)$, where Γ denotes the Euler Gamma function.

Positive r.v. (and their distributions), whose Laplace transform equals $\exp\{-c' t^\alpha\}$ with some constants $c' > 0$ and $\alpha \in (0,1)$ are called *α-stable* with the *index of stability α*. One can show that such random variable X satisfies condition (5.10). In particular, it has fat tails. The term 'stable' stems from the observation that the sum of independent copies of such a random variable belongs to the same class, because the Laplace transform of this sum equals $\exp\{-c' n t^\alpha\}$, where n denotes the number of terms.

The above Proposition shows that not only Gaussian r.v. can describe the limits of the sums of i.i.d. r.v. Namely, one says that a r.v. X belongs to the *domain of attraction* of an α-stable law, if (5.10) holds. By Proposition 5.3.1, the distributions of the normalized sums of i.i.d. copies of such r.v. converge to the α-stable law. Unfortunately, the probability density functions of stable laws with $\alpha \in (0,1)$ can not be expressed in a closed form (in elementary function), which makes their analysis more involved than, say, of Gaussian or exponential distributions.

Index