Script for the Course Fundamental Probability for Finance*

Jeffrey Näf

August 15, 2022

*This script is based in some parts on the lecture slides of an earlier version of the course by Marc Paolella.

Contents

1 Introduction

The concept of probability or "chance" is abundant in our world, from the throw of a coin to complex stochastic processes, such as financial returns data.[1] In this lecture the goal is to struck a balance between mathematical rigorosity and comprehensibility of the presented material (which admittedly is not always easy in this particular branch of mathematics).

1.1 Motivation

Let us start by considering definitions already encountered in a first statistics course (at least partly). This is just an overview, we will study most of these concepts in more detail during the course. We deal with some sample space Ω (often called S in statistic courses for economists), with certain *outcomes* $\omega \in \Omega$. On the other hand there are *events* $A \subset \Omega$ on which we can define probabilities $P(A)$. These are real numbers which are constrained to lie between 0 and 1 and have to fulfill some other conditions. These conditions will be studied in more detail in the next section. Consider the following example:

Example 1. A fair die is tossed once and the number of dots is observed. The set of outcomes is $\Omega = \{1, 2, 3, 4, 5, 6\}$ with natural ordering $\omega_i = i$, $i = 1, \ldots, 6$. Each outcome is equally likely: $P(\{\omega_i\}) = 1/6$. Possible events include $E =$"rolling an even number" and $O =$"rolling a odd number". In this case $E = \{2, 4, 6\}$, $O = \{1, 3, 5\}$. $\Diamond$

So P is a function defined on sets, which means we need to define a domain of P. We will see that the notion of a "σ-algebra" is most useful here. One then usually defines random variables X, with values in $\mathbb{R}$, and random vectors $\mathbf{X}$ with values in $\mathbb{R}^d$. As will be seen these are actually functions(!) with domain Ω and codomain $\mathbb{R}$ or $\mathbb{R}^d$, e.g. $X : \Omega \to \mathbb{R}$. Then we call for an arbitrary $\omega \in \Omega$, $X(\omega) = x$ a *realization of* X. Usually one then differentiates between X being "discrete" (i.e. X takes at most countably many values) or "continous" (i.e. it takes values in a continuoum). There is actually much more to this distinction and we will attempt to be more precise about this in Sections 3/4.

There are many different ways of describing a random variable X (i.e. many different angles we can look at it). Often we are solely interested in its distribution, which tells us which values occur with which probability/frequency. In Section 3, we'll see that any random variable has a cummulative distribution function (cdf):

$$\mathbb{R} \ni x \mapsto F(x) = P(X \leq x).$$

If X is continuous, this cdf is differentiable everywhere, and the derivative is called the probability density function (pdf), $x \mapsto f(x)$. This is most often the first thing we want and there are a great

[1] Deeply connected to this is the fact that we experience uncertainty in almost every step of our lives. This is exemplified by the field of artificial intelligence (AI); First intelligences, built solely upon logic, were not able to handle any "difficult" task. Only after the recent introduction of probabilistic methods, allowing for a degree of uncertainty (e.g. something might be only true in 80 % of times, or on average), the field exploded with complex tasks it could suddenly solve. See for example Russell and Norvig (2003).

number of important pdf's around. For example, the pdf or density of $X \sim N(0,1)$ is plotted in Figure 1. The analogous object for discrete X is the function

$$\mathbb{N} \cup \{0\} \ni x \mapsto P(X = x).$$

In both cases (continuous and discrete) we may also define "moments" of X. It all starts with the definition of the most important concept of expected value:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x) dx \quad \text{or} \quad \mathbb{E}[X] = \sum_{x=0}^{\infty} x P(X = x).$$

This expression may be seen as a measure of location of a distribution. The variance on the other hand is a measure of dispersion. It is defined as

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

For $k \in \mathbb{N}$, the kth moment of a distribution is defined as

$$\mathbb{E}[X^k] = \int_{-\infty}^{+\infty} x^k f(x) dx \quad \text{or} \quad \mathbb{E}[X^k] = \sum_{x=0}^{\infty} x^k P(X = k).$$

Taking $k = 1$ gives back the expected value. Very importantly, it might be the case that any kth moment of a distribution does not exist! That is to say $\mathbb{E}[X^k]$ might not be well-defined, but $\pm\infty$, or it might not even be defined at all. We will see an example where not even $\mathbb{E}[X]$ is defined in Section 3.

Example 2. Consider again the die role above, but now $\Omega = \{one, two, three, four, five, six\}$. Then Ω is a set with six elements $\omega_1, \ldots, \omega_6$, which are however not easy to handle mathematically. We may still say that $P(\{one\}) = \ldots = P(\{six\}) = 1/6$, but not much more. Define the random variable $X : \Omega \to \mathbb{R}$, $X(one) = 1, X(two) = 2, \ldots, X(six) = 6$. Now we are back on the nicely behaved space $\mathbb{R}$. Furthermore, for $x \in \{1, \ldots, 6\}$, $P(X = x) := P(\{\omega \in \Omega : X(\omega) = x\}) = P(\{\omega_x\}) = 1/6$. Thus we are able to calculate:

$$\mathbb{E}[X] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \cdot + \frac{1}{6} \cdot 6 = 3.5,$$

or

$$\mathbb{V}(X) = \frac{1}{6} \cdot (1 - 3.5)^2 + \frac{1}{6} \cdot (2 - 3.5)^2 + \cdot + \frac{1}{6} \cdot (6 - 3.5)^2 = \frac{35}{12}$$

Note that we did not really need the underlying space Ω above. This is a repeating patter: Using appropriate random variables, we are able to forget about the underlying sample space Ω sooner or later and focus on $\mathbb{R}$ or $\mathbb{R}^d$. $\Diamond$

1.2 Some mathematical notions

In mathematics basically everything is a set, see e.g. Dudley (2002, Chapter 1). In particular sets play an immensely important role in probability theory, since "events" are modeled as sets. Thus we recall a few set operations. First, for any set Ω, 2^Ω is the power set, the set of all subsets of Ω. That means any subset $A \subset \Omega$ has $A \in 2^\Omega$. So 2^Ω is a "set of sets" (sets whose elements are themselves sets). Let $A, B \in 2^\Omega$ (or $A, B \subset \Omega$), then important definitions are

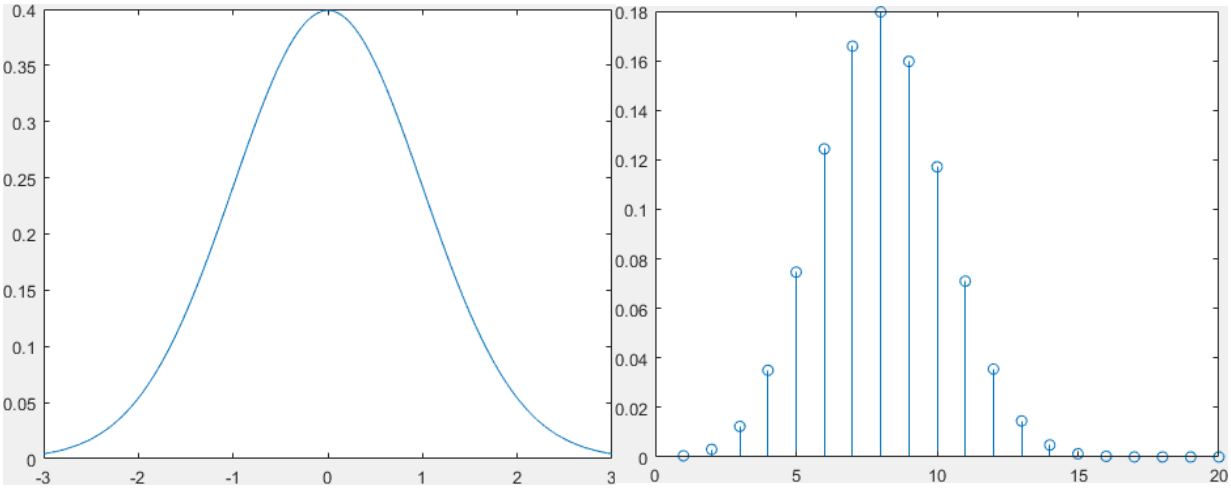

Figure 1: pdf of a standard normal (left) and binomial with number of trials $N = 20$ and probability of success $p = 0.4$ (right)

- intersection: $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$.
- union: $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$
- complement: $\Omega \setminus A = A^c = \{\omega \in \Omega : \omega \notin A\}$.
- difference: $A \setminus B = A \cap B^c$.
- inclusion: $A \subset B$ means $\omega \in A$ implies $\omega \in B$. Note that in this lecture, if $A \subset B$, then $A = B$ is also possible (so "$\subset$" is not a strict inclusion).

Further, A and B are called (mutually) disjoint if $A \cap B = \emptyset$, i.e. there is no point in Ω that lies both in A and B. By the principle of "extensionality" (Dudley, 2002, p. 3) $A = B$ iff (if and only if) $A \subset B$ and $B \subset A$. This is often used when one wants to proof $A = B$; We then first proof $A \subset B$ by looking at arbitrary $\omega \in A$ and showing that $\omega \in B$ also holds. Doing the same thing the other way around gives $B \subset A$ and thus $A = B$. Let us look at an example:

Example 3. Let $\Omega = \{-10, 2, 10\}$, $A = \{-10, 2\}$ and $B = \{2, 10\}$ (notice the brackets). Then

$$2^\Omega = \{\{-10\}, \{2\}, \{10\}, \{-10, 2\}, \{-10, 10\}, \{2, 10\}, \{-10, 2, 10\}, \emptyset\}.$$

In particular $\Omega \in 2^\Omega$, and the same is true for A, B. Furthermore, $A \cap B = \{2\}$, $A \cup B = \{-10, 2, 10\}$, $A^c = \{10\}$, $A \setminus B = \{-10\}$. ◊

One can even define intersection and union for arbitrary sets! Important for us will be the following case; Let $(A_n)_{n\in\mathbb{N}}$ be a countable collection of sets.[2] Then we may define

$$\bigcap_{n=1}^{\infty} A_n = \{\omega \in \Omega : \omega \in A_n \text{ for all } n \in \mathbb{N}\},$$

[2]We will soon be more precise about what countable means, for the moment just take it to be that the indices of the sets A_n all lie in the natural numbers $\mathbb{N}$.

and

$$\bigcup_{n=1}^{\infty} A_n = \{\omega \in \Omega : \omega \in A_n \text{ for some } n \in \mathbb{N}\}.$$

It is also follows immediately from the definitions that

$$\left(\bigcap_{n=1}^{\infty} A_n\right)^c = \bigcup_{n=1}^{\infty} A_n^c \quad \text{and} \quad \left(\bigcup_{n=1}^{\infty} A_n\right)^c = \bigcap_{n=1}^{\infty} A_n^c. \tag{1}$$

Finally note that everything we can proof for an infinite union/intersection, also holds for a finite one. The reason is that we are able to write every finite union $\bigcup_{n=1}^{N} A_n$ as

$$\bigcup_{n=1}^{N} A_n = \bigcup_{n=1}^{\infty} A_n,$$

simply by taking $A_m = \emptyset$ for all $m \geq N+1$. Similarly, we may write

$$\bigcap_{n=1}^{N} A_n = \bigcap_{n=1}^{\infty} A_n,$$

by taking $A_m = \Omega$ for all $m \geq N+1$. Another useful fact is that whenever $A \subset B$, then $B^c \subset A^c$ (see the Venn diagram for some intuition). Indeed, if $\omega \in B^c$, then $\omega \notin B$ by definition, and, since $A \subset B$, $\omega \notin A$, or $\omega \in A^c$.

Important examples of sets are also given by the different spaces one considers in mathematics. Dudley (2002, Chapter 2) presents a very nice treatment of this. A topological space $(T, \mathcal{T})$, or a measurable space $(\Omega, \mathcal{A})$ are highest in generality. A metric space (M, d) has a metric d defined on it (a function into $\mathbb{R}$ to measure distance between points) and is a special case of a topological space (i.e. every metric space is a topological space). A normed space $(\mathcal{V}, \|\cdot\|)$ is a vector space with a norm $\|\cdot\|$ on it. Again, every normed space is also a metric space. Finally $(\mathbb{R}^d, \|\cdot\|_d)$, $(\mathbb{R}, \|\cdot\|_1)$ or $(\mathbb{N}, \|\cdot\|_1)$ are examples of normed spaces, with, for $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$\|\mathbf{x}\|_d = \sqrt{\sum_{i=1}^{d} x_i^2},$$

the "Euclidean norm". They are mathematically speaking the "nicest" spaces we will encounter and also the most important ones for our purposes.

2 Axiomatic Probability

We will now make all of the above more mathematically formal. The goal is to define things as accurately as possible, without causing too much confusion. It thereby helps to remember, that mathematics has actually little to do with reality. It just turned out to be a very powerful tool in modeling real world phenomena. But in principle it is simply a logical coherent fairy tale, based on axioms that might be true or false. One example of such unprovable axioms is given by the Axioms of Probability, in Definition 2.2 below. The goal was to define something that is mathematically useable and at the same time allows to model the concept of probability or randomness that one observes in nature.

2.1 Countable Probability Spaces

Any set Ω is finite, if it has finitely many elements. It is *countably infinite* if there exists a bijective function to $\mathbb{N}$. This is a complicated way of saying that Ω and $\mathbb{N}$ have the same number of elements, as each element of Ω can be uniquely represented as an element of $\mathbb{N}$. An example is $\Omega = \mathbb{Q}$, the set of rational numbers. Then one can show that there exists a bijective (and therefore invertible) function $f : \mathbb{N} \to \mathbb{Q}$. Another example is the Cartesian product

$$\Omega = \mathbb{N} \times \mathbb{N} = \mathbb{N}^2 = \{(n, m) : n \in \mathbb{N} \text{ and } m \in \mathbb{N}\}.$$

In this case we can define the bijective function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ simply as $f(m, n) = 2^m(2n+1)-1$, so $\mathbb{N} \times \mathbb{N}$ is also countable.

Let first Ω be a finite set. That is, denote the sample space Ω with elements, or possible outcomes, ω_i, $i = 1, \ldots, N$. Then Ω and a function P with domain 2^Ω and range $[0, 1]$ such that $\sum_{i=1}^{N} P(\{\omega_i\}) = 1$ is referred to as a *finite probability space*. In these cases we often need "counting", that is to assess with combinatorial arguments how large a subset will be.[3] It then often helps to think of Ω as an urn, from which we draw balls of different kinds. Thus assume we have N unique balls in an urn: we randomly draw n of them. We wish to know how many ways there are of doing this, but we need to specify if the balls are drawn with or without replacement, and also if the ordering of the balls is relevant or not. If we draw two blue balls $b1, b2$ for instance, it does not matter which one we draw first, so we don't have to count both $(b1, b2)$ and $(b2, b1)$. Consider drawing the balls,

- "*Ordered without replacement*": The first draw is one of N possibilities; the second is one of $(N-1)$, ..., the n^{th} is one of $N - n + 1$. In total: $R_{[n]} = N! \,/\, (N-n)!$.
- "*Ordered with replacement*": the first draw is one of N possibilities; the second is one of N, etc., so N^n possibilities.
- "*Unordered without replacement*": similar to ordered without replacement, but we need to divide $R_{[n]}$ by $n!$ to account for the irrelevance of order, giving "N choose n", $\frac{N!}{(N-n)!n!} = \binom{N}{n}$.

We left out the case of "Unordered with replacement", which is quite involved (as combinatorial arguments often are), see e.g. Paolella (2006, Chapter 2.1). Let us look at an example instead:

Example 4. A lottery consists of 100 tickets, labeled $1, 2, \ldots, 100$, three of which are "winning numbers". You buy 4 tickets (that is you choose 4 tickets with equal probability from Ω). Calculate the probability, p, that you have at least one winning ticket.

This is a situation with a drawing *without replacement* and where *order does not matter*. In this case Ω is somewhat complicated. Denote $N = \{1, 2, 3, \ldots, 100\}$ the numbers of the 100

[3] This should not be the focus of this lecture, so we will go through this quite quickly, even though combinatorial arguments are of utmost importance in certain areas of probability. For details see Paolella (2006).

lottery tickets. Then

$$\Omega = \{\{1,2,3,4\},\{1,2,3,5\},\ldots,\{97,98,99,100\}\},$$

that is each element of Ω is of the form $\omega = \{m,n,q,r\}$ with $m,n,q,r \in N$ and $m \neq n \neq q \neq r$. It is a set (i.e. curly brackets) because order does not matter, so that for instance $\{1,2,3,4\} = \{4,3,1,2\}$. The condition $m \neq n \neq q \neq r$ encodes the fact that we draw without replacement, so no number can be drawn twice. Note that Ω has cardinality $\binom{100}{4}$. Now, there are 3 winning numbers i,j,l say, while the rest are "losers". Let $A \in 2^\Omega$ be all ω with at least one winning number in it, i.e. (assuming i,j,l are not equal to $1,2,3,100$ for the sake of illustration):

$$A = \{\{1,2,3,i\},\{1,2,3,j\},\{1,2,3,l\},\ldots,\{i,j,l,100\}\}.$$

In this case A has

$$\binom{3}{1}\binom{97}{3} + \binom{3}{2}\binom{97}{2} + \binom{3}{3}\binom{97}{1}$$

elements. The question is now what $p = P(A)$ is. One way to solve this is the following: We do not win with probability $1-p$. In this case, from the 3 winning tickets, you choose none, from the 97 losing tickets, you draw 4. This means A^c has $\binom{3}{0}\binom{97}{4}$ elements (so that in fact $|A| + |A^c| = |\Omega|$). Thus, since $P(\{\omega\}) = 1/\binom{100}{4}$ for all $\omega \in \Omega$,

$$1 - p = P(A^c) = \frac{\binom{3}{0}\binom{97}{4}}{\binom{100}{4}} = \frac{97\cdot 96\cdot 95\cdot 94}{4!}\frac{4!}{100\cdot 99\cdot 98\cdot 97} = \frac{96\cdot 95\cdot 94}{100\cdot 99\cdot 98} = 0.8836.$$

$\Diamond$

We can also consider a small example of a countably infinite probability space:

Example 5. Consider tossing a fair coin until a tail appears. It is theoretically possible that the first 10,000 trials will result in heads, or that a tail may never occur. Letting ω_i be the total number of required tosses, $i = 1,2,\ldots$, we see that Ω is countably infinite. If associated with Ω is a function P with domain 2^Ω and range $[0,1]$ such that $\sum_{i=1}^{\infty} P(\{\omega_i\}) = 1$, $\left(\Omega, 2^\Omega, P\right)$ is referred to as a *probability space*. For example, taking $P(\{\omega_i\}) = (1/2)^i$ is valid. $\Diamond$

It will be demonstrated in the next section that things are indeed quite simple in this case, in the sense that we can characterize the whole probability measure P with a collection of numbers.

2.2 General Situation

Before we can define probability formally, we need to talk about the concept of sets of sets. In particular these are sets with elements that are themselves sets.

Let us again properly introduce the terms already encountered in the introduction.

- A *realization* is the result of some well-defined *trial* or *experiment* performed under a given *set of conditions*, whose *outcome* is not known in advance, but belongs to a *set of possibilities* or *set of outcomes* which are known in advance.

- The set of possible outcomes could be *countable* (either *finite* or *denumerable*, i.e., *countably infinite*) or *uncountable*.
- Denote the *sample space* as Ω, the set of all possible outcomes, with individual outcomes or *sample points* $\omega_1, \omega_2, \ldots$. A *subset* of the sample space is known as an *event*, usually denoted by a capital letter, possibly with subscripts, i.e., A, B_1, etc., the totality of which under Ω will be denoted $\mathcal{A}$, and forms the *collection* of events—also called the *collection of measurable events*. This is the σ-algebra introduced in Definition 2.1.
- An outcome $\omega \in \Omega$ may belong to many events, always belongs to the *certain event* Ω, and never to $\emptyset$, the *empty set* or *impossible event*.
- The usual operations in set theory can be applied to two events, i.e., *complement*, *intersection*, *union*, *difference*, *symmetric difference*, *inclusion*, etc.
- Two events are mutually exclusive or *disjoint* if $A \cap B = \emptyset$.
- If a particular set of events A_i, $i \in J$, are such that $\bigcup_{i \in J} A_i \supseteq \Omega$, they *cover* (or *exhaust*) the same space Ω.
- If events A_i, $i \in J$, are such that $\bigcup_{i \in J} A_i = \Omega$ are disjoint and exhaust Ω, they *partition* Ω, i.e., one and only one of the A_i will occur on a given trial.

Now the probability P is defined on sets. So what is its domain? When Ω was finite or countably infinite, we could simply define P on 2^Ω. It turns out that does not work any longer if Ω is uncountable: One can construct a counter example for which it is not possible to find $P : 2^\Omega \to [0, 1]$ and also have the properties in Definition 2.2. The problem, as it turns out, is that 2^Ω is too large, see also Jacod and Protter (2004, p. 35). So we need to take a smaller set of sets to define P on. However it should still be large enough to fulfill certain sensible properties. To formalize this, the notion of a "σ-algebra" is introduced:

Definition 2.1 (σ-algebra)**.** Let Ω be an arbitrary set. $\mathcal{A} \subset 2^\Omega$ is called a σ-algebra if

(i) $\Omega \in \mathcal{A}$

(ii) If $A \in \mathcal{A}$ then also $A^c \in \mathcal{A}$

(iii) If $(A_n)_{n \in \mathbb{N}}$ is a family/sequence of sets in $\mathcal{A}$ (i.e. $A_n \in \mathcal{A}$ for all n), then also

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$$

For (iii) we say $\mathcal{A}$ is closed under countable unions. Since any finite union can be written as a countably infinite one (by choosing A_m as the empty set for all m larger than the number of sets,

as in Section 1.2), it is also closed under finite unions. Moreover, if $(A_n)_{n\in\mathbb{N}}$ is a family/sequence of sets in $\mathcal{A}$ then A_n and also $A_n^c \in \mathcal{A}$ and thus

$$\bigcap_{n=1}^{\infty} A_n = \left(\bigcup_{n=1}^{\infty} A_n^c\right)^c \in \mathcal{A}.$$

So $\mathcal{A}$ is also closed under countable intersection (and thus also under finite intersection). A first important example of a σ-algebra is 2^Ω for any set Ω. All subsets A of Ω are elements of 2^Ω, so clearly (i)-(iii) of the above definition hold. Crucially, there are however σ-algebras strictly smaller than 2^Ω on which we can define probability measures. For instance, if we take any set $A \subset \Omega$ and define $\mathcal{A} = \{A, A^c, \Omega, \emptyset\}$, then you may verify that $\mathcal{A}$ is a σ-algebra.

Having the notion of a σ-algebra at hand, we can finally define probability formally:

Definition 2.2 (Probability measure)**.** Let $\mathcal{A}$ be a σ-algebra on Ω. A *probability measure* is a function, which assigns a real number $P(A)$ to each event $A \in \mathcal{A}$ such that

(i) $P(A) \geq 0$,

(ii) $P(\Omega) = 1$, and

(iii) for a countably infinite sequence of mutually exclusive events $(A_n)_{n\in\mathbb{N}}$,

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

The latter requirement (iii) is known as *(countable) additivity*. If $A_i \cap A_j = \emptyset$, $i \neq j$ and $A_{n+1} = A_{n+2} = \cdots = \emptyset$, then $P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$, which is *finite additivity*. The triplet $(\Omega, \mathcal{A}, P)$ refers to the *probability space* with sample space Ω, collection of measurable events $\mathcal{A}$ and probability measure P. Finally note that while (i) and (ii) are clearly sensible assumptions to put on a probability, condition (iii) has more to do with mathematical convenience than anything else. Many important results could not be proven without this axiom.

Let us go back to the case of countably infinite or finite probability spaces (again, this is what we refer to as being "countable"). In this case any set $A \subset \Omega$ has at most a countably infinite number of points. But two single points are always disjoint sets (that is $\{\omega_i\}$ and $\{\omega_j\}$ are always disjoint for $j \neq i$), thus point (iii) means:

$$P(A) = \sum_{\omega\in A} P(\{\omega\}).$$

In words; the probability of A is just the sum of the probability of its elements. This seems to make intuitive sense and is often also mentioned in a first statistics course.[4] For instance, in Example 1, the probability of rolling an even number $P(E) = P(\{2,4,6\}) = P(\{2\}) + P(\{4\}) + P(\{6\})$. So in fact if we know what $P(\{\omega\})$ is for all $\omega \in \Omega$, we can measure the probability of *any* $A \subset \Omega$. In other words P can be defined on 2^Ω, as mentioned before. This has an important consequence:

[4] As we will see this is of utmost importance for discrete random variables.

Theorem 2.1 (Adaptation of Theorem 4.1 in Jacod and Protter (2004)). *If Ω is countable, then:*

(a) Any collection of real numbers $(p_\omega)_{\omega\in\Omega}$ with $\sum_{\omega\in\Omega} p_\omega = 1$ and $p_\omega \geq 0$ uniquely defines a probability on $(\Omega, 2^\Omega)$.

(b) Conversely, any probability P on $(\Omega, 2^\Omega)$ is characterized by its values on the atoms, i.e. by $p_\omega := P(\{\omega\}), \omega \in \Omega$.

Proof. For (a), let us define a function P on the σ-algebra 2^Ω and check that it meets the properties in Definition 2.2. For any $A \in 2^\Omega$, define $P(A)$ as,

$$P(A) = \begin{cases} \sum_{\omega\in A} p_\omega, & \text{if } A \neq \emptyset \\ 0, & \text{if } A = \emptyset \end{cases}.$$

This $P(A)$ is indeed well-defined, meaning that each $A \in 2^\Omega$, $P(A)$ is a unique number. This follows because $p_\omega \geq 0$ and

$$\sum_{\omega\in A} p_\omega \leq \sum_{\omega\in\Omega} p_\omega = 1,$$

meaning the sum, if uncountably infinite, is *absolutely converging*. Also see Remark 1.

Then, since $p_\omega \geq 0$, $P(A) \geq 0$ for all $A \in 2^\Omega$, which is (i). Further

$$P(\Omega) = \sum_{\omega\in\Omega} p_\omega = 1,$$

by assumption giving (ii). Now let $(A_n)_{n\in\mathbb{N}}$ be a countable, mutually disjoint collection of sets in 2^Ω, then it holds that

$$P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{\omega\in\cup_n A_n} p_\omega,$$

by definition. Now, because the A_n are mutually disjoint, it follows that $\omega \in \bigcup_n A_n$ means ω is exactly in *one* A_n: It has to be in at least one A_n, A_i say, but if it were in A_j as well for some $j \neq i$, then A_i and A_j would not be disjoint. But this indeed means that

$$\sum_{\omega\in\cup_n A_n} p_\omega = \sum_{\omega\in A_1} p_\omega + \sum_{\omega\in A_2} p_\omega + \ldots = \sum_{n\in\mathbb{N}} \sum_{\omega\in A_n} p_\omega = \sum_{n\in\mathbb{N}} P(A_n),$$

with Remark 1 below. In other words we have just shown that

$$P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n),$$

and (iii) holds as well! So $P : 2^\Omega \to [0, 1]$ meets the conditions in Definition 2.2 and is thus a probability measure.

For (b), it is enough to remember what we have shown above, namely that for any probability P on $(\Omega, 2^\Omega)$ and any $A \in 2^\Omega$:

$$P(A) = \sum_{\omega\in A} P(\{\omega\}).$$

So clearly we can find the probability of any set $A \in 2^\Omega$, as soon as we know $p_\omega := P(\{\omega\})$ for all ω. This is exactly what we mean by the statement that P is characterized by $(p_\omega)_{\omega\in\Omega}$ ■

Remark 1. Note that the expression $\sum_{\omega\in A} p_\omega$ above does not say anything about the order of how we sum (and might not even be well-defined a priori). All we know is that we sum potentially infinitely many terms. In general, one has to be careful with such things; Consider the sum

$$\sum_{n=1}^{\infty} u_n$$

with $u_n \in \mathbb{R}$ for all n. This sum may not defined in the sense that the sequence $S_N = \sum_{n=1}^{N} u_n$ does not converge (this is what we mean, by saying the sum is not "well-defined"). Consider for example $u_1 = u_3 = \ldots = 1$ and $u_2 = u_4 = \ldots = -1$. In this case S_N alternates between 1 and 0, without ever converging, i.e. the limit is not well-defined. This is not a problem however if all $u_n \geq 0$: In this case $S_N = \sum_{n=1}^{N} u_n$ is a monotone sequence. Its limit might be $+\infty$, but it will always be well-defined. What is more, it turns out that in this case, we may change the order of the series as we like.[5] Consequently, not only is the term $\sum_{\omega\in A} p_\omega$ well-defined (and even smaller or equal one in our case), but the problem that we did not specify the order in which we sum evaporates. Also see (Jacod and Protter, 2004, Chapter 4). ◊

Example 6. An important application of Theorem 2.1 is a finite sample space Ω for which the probability is the same for each outcome, i.e. $P(\{\omega\}) = a \in [0, 1]$. The condition

$$1 = \sum_{\omega\in\Omega} P(\{\omega\}) = |\Omega| \cdot a$$

yields $a = 1/|\Omega|$. In words: The probability of each outcome is simply the inverse of the number of elements in Ω. In Example 1 we had $a = 1/6$ (the typical die), in Example 4 we had $a = 1/\binom{100}{4}$. The formula $P(A) = \sum_{\omega\in A} P(\{\omega\}) = |A|a$ with $a = 1/|\Omega|$, means that

$$P(A) = \frac{|A|}{|\Omega|}.$$

So indeed in these examples (when all single outcomes are equally likely) the naive

$$p = \frac{\text{Number of things we want}}{\text{Number of total things}}$$

is correct. ◊

As said a few times already, things get more complicated for an Ω which is not countable:

Example 7. Consider the space $\Omega = \mathbb{R}$. If we want to define a probability on this space, what kind of σ-algebra should we choose? The most common one is the so called Borel σ-algebra, denoted $\mathcal{B}(\mathbb{R})$. It is the smallest σ-algebra containing all open sets of $\mathbb{R}$ (one also say that it is generated by all open sets in $\mathbb{R}$). So for instance any open interval (a, b), with $-\infty \leq a \leq b \leq +\infty$, is in $\mathcal{B}(\mathbb{R})$. What about other types of intervals? Take $(a, b]$ instead, for $-\infty \leq a \leq b < +\infty$. Then we can actually write

$$(a, b] = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right) \in \mathcal{B}(\mathbb{R}),$$

[5]This could easily be proven using the fact that we can change the order of the summation for any absolutely convergent sequence and then differentiate between the cases $\sum_{n=1}^{\infty} u_n < \infty$ and $\sum_{n=1}^{\infty} u_n = \infty$.

since $(a, b + 1/n) \in \mathcal{B}(\mathbb{R})$ and by definition of the σ-algebra. In fact it turns out that it is really hard to find a subset $B \subset \mathbb{R}$ which is not element of $\mathcal{B}(\mathbb{R})$.[6] We will see in the section on random variables why this is convenient in practice. What kind of probability measure may we find on $\mathbb{R}$? We will encounter numerous examples in Section 3, since any known distribution/density is actually a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. So the Gaussian density plotted in Figure 1 for instance defines a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Here we focus on a different example: Consider the "dirac" measure at $x \in \mathbb{R}$, $\delta_x : \mathcal{B}(\mathbb{R}) \to [0, 1]$. This probability measure is simply defined as

$$\delta_x(A) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{else} \end{cases}.$$

Let us check the conditions of Definition 2.2: Clearly, $\delta_x \geq 0$ and since $x \in \mathbb{R}$, $\delta_x(\mathbb{R}) = 1$. So (i) and (ii) are fine. Now let $(A_n)_{n\in\mathbb{N}}$ be a countable sequence of mutually exclusive events in $\mathcal{B}(\mathbb{R})$ (for example $A_n = (n, n + 1]$, or $A_n = (1/(n + 1), 1/n)$). Then for any $x \in \bigcup_{n=1}^{\infty} A_n$, x is element of *exactly one* A_n. just as in the proof of Theorem 2.1: By definition it is element of at least one A_n, but since the sets are disjoint it can not lie in more than one. On the other hand if $x \in A_n$ for one n, then $x \in \bigcup_{n=1}^{\infty} A_n$, so:

$$x \in \bigcup_{n=1}^{\infty} A_n \iff x \in A_n \text{ for exactly one } n$$

With this, it is easy to see that whenever $\delta_x\left(\bigcup_{n=1}^{\infty} A_n\right) = 1$, so is $\sum_{n=1}^{\infty} \delta_x(A_n)$, and the same for $\delta_x\left(\bigcup_{n=1}^{\infty} A_n\right) = 0$. So indeed

$$\delta_x\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \delta_x(A_n).$$

This simple measure can in fact be used to construct any "discrete" probability measure as we shall see in Section 3.3. It is also bears strong resemblance to a probability measure on a countable space, which we will also shortly explore in said section. ◊

In fact the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is most important for us. The reason is that through random variables, as introduced in Section 3, we can actually in most cases get to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, no matter with what probability space we start. Now, going back to an arbitrary probability space $(\Omega, \mathcal{A}, P)$, we study properties that Definition 2.2 implies. Many of these properties of $(\Omega, \mathcal{A}, P)$ are intuitive and easy to see, notably so with the help of a *Venn diagram*, such as in Figure 2. We will nonetheless show them all formally:

(i) If $A \subset B$, then $P(A) \leq P(B)$: Indeed in this case we can use additivity of P, to obtain:

$$P(B) = P((B \setminus A) \cup A) = \underbrace{P(B \setminus A)}_{\geq 0} + P(A) \geq P(A).$$

[6] An example of such a set is the cleverly constructed "Cantor set", see e.g. Dudley (2002, Chapter 3.4). So indeed $\mathcal{B}(\mathbb{R})$ is smaller than $2^{\mathbb{R}}$.

(ii) $P(A) \leq 1$, since $P(A) \leq P(\Omega)$ by (i).

(iii) $P(A^c) = 1 - P(A)$. Again using additivity together with the fact that A and A^c are disjoint by definition:

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c),$$

or $P(A^c) = 1 - P(A)$.

(iv) $P(\emptyset) = 0$, directly implied by (iii) since $\Omega^c = \emptyset$.

(v) $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$. This follows because $A_1 \cup A_2 = (A_1 \setminus A_2) \cup (A_1 \cap A_2) \cup (A_2 \setminus A_1)$, so that, since these are all disjoint sets,

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1 \setminus A_2) + P(A_1 \cap A_2) + P(A_2 \setminus A_1) \\ &= P(A_1) + P(A_2 \setminus A_1) \\ &= P(A_1) + P(A_2 \setminus A_1) + P(A_1 \cap A_2) - P(A_1 \cap A_2) \\ &= P(A_1) + P(A_2) - P(A_1 \cap A_2). \end{aligned}$$

A few comments for the above: First, (i) is referred to as the monotonicity of the probability measure, and is of great importance, despite its simplicity. Second, essentially the same argument as in (iii) can be used to show a slightly more general statement: It is clear from the Venn diagram that event A can be partitioned into two disjoints sets $A \cap B$ and $A \cap B^c$, i.e., $P(A) = P(A \cap B) + P(A \cap B^c)$ or $P(A \setminus B) = P(A \cap B^c) = P(A) - P(A \cap B)$. For $A = \Omega$, we get (iii) back. Finally, if one defines a general measure ν (instead of probability measures), then the condition $\nu(\Omega) = 1$ in Definition 2.2 is replaced by (iv), i.e., $\nu(\emptyset) = 0$.

A bit harder to proof is the important result in Theorem 2.2. To make sense of it, we first need to mention that one can actually define a limit for a *monotone increasing sequence* $(A_n)_{n \in \mathbb{N}}$:

$$\lim_{n \to \infty} A_n := \bigcup_{n=1}^{\infty} A_n.$$

Similarly, for $(B_n)_{n \in \mathbb{N}}$ a *monotone decreasing sequence*:

$$\lim_{n \to \infty} B_n := \bigcap_{n=1}^{\infty} B_n.$$

We are not completely rigorous here, but only want to mention that these limits of sets are "well-defined" in the following sense: Take $A = \bigcup_{n=1}^{\infty} A_n$ and $B = \bigcap_{n=1}^{\infty} B_n$. Then A, B are always well-defined (even if A_n and B_n are not monotone). However with monotonicity, A_n or B_n indeed get "closer and closer" to A or B as n gets larger, since for each n, $\bigcup_{i=1}^{n} A_i = A_n$ and $\bigcap_{i=1}^{n} B_i = B_n$. Thus $\lim_{n \to \infty} A_n$ and $\lim_{n \to \infty} B_n$ make sense for monotone (increasing or decreasing) sequences. Finally, we also note that if $(C_n)_{n \in \mathbb{N}}$ is an arbitrary sequence of sets (not necessarily monotone), then the sequence $(A_n)_{n \in \mathbb{N}}$ with

$$A_n := \bigcup_{i=1}^{n} C_i,$$

is always monotone increasing, while $(B_n)_{n\in\mathbb{N}}$ with

$$B_n := \bigcap_{i=1}^{n} C_i,$$

is always monotone decreasing. So even for arbitrary sequences $(C_n)_{n\in\mathbb{N}}$,

$$\lim_{n\to\infty} \bigcup_{i=1}^{n} C_i = \bigcup_{i=1}^{\infty} C_i, \quad \lim_{n\to\infty} \bigcap_{i=1}^{n} C_i = \bigcap_{i=1}^{\infty} C_i.$$

Theorem 2.2. *For* any *sequence of sets* $(A_n)_{n\in\mathbb{N}}$ *in* $\mathcal{A}$*,*

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} P(A_n). \tag{2}$$

If further $(A_n)_{n\in\mathbb{N}}$ *is a monotone increasing sequence in* $\mathcal{A}$ *(i.e.,* $A_n \subset A_{n+1}$ *for all* n*), then*

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcup_{n=1}^{\infty} A_n\right) = P\left(\lim_{n\to\infty} A_n\right). \tag{3}$$

Similarly if $(B_n)_{n\in\mathbb{N}}$ *is a monotone decreasing sequence in* $\mathcal{A}$ *(i.e.,* $B_{n+1} \subset B_n$ *for all* n*), then*

$$\lim_{n\to\infty} P(B_n) = P\left(\bigcap_{n=1}^{\infty} B_n\right) = P\left(\lim_{n\to\infty} B_n\right). \tag{4}$$

Proof. The proof of this theorem requires essentially just one (maybe not immediately obvious) idea: Define for an arbitrary sequence of sets $(A_n)_n$, $C_1 = A_1$ and

$$C_n = A_n \setminus \bigcup_{i=1}^{n-1} A_i$$

for all $n = 2, 3, \ldots$. Thus, C_2 is the part of A_2 which is "new", i.e., not already in A_1 and,

$$\begin{aligned} \bigcup_{i=1}^{n} A_i &= A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus (A_1 \cup A_2)) \cup \cdots \cup (A_n \setminus A_1 \cup \ldots \cup A_{n-1}) \\ &= C_1 \cup C_2 \cup \cdots \cup C_n = \bigcup_{i=1}^{n} C_i. \end{aligned}$$

We furthermore have: (a) $C_n \subset A_n$, (b) $(C_n)_n$ are disjoint (!), and, as we have just shown,

$$(c)\ \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} C_i.$$

Crucially, with (b) we can use the countable additivity of P. Furthermore since the sequences of sets $\left(\bigcup_{i=1}^{n} A_i\right)_{n\in\mathbb{N}}$ and $\left(\bigcup_{i=1}^{n} C_i\right)_{n\in\mathbb{N}}$ are automatically increasing (we add more and more sets), their limits are well defined. In particular, we can take the limit in (c) on both sides to obtain

$$(c')\ \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} C_i.$$

Let us first use this trick for an arbitrary sequence $(A_n)_{n\in\mathbb{N}}$ in $\mathcal{A}$, i.e. construct $(C_n)_n$ as above. Then by countable additivity and the fact that $P(C_i) \leq P(A_i)$ by monotonicity of P,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \overset{(c')}{=} P\left(\bigcup_{i=1}^{\infty} C_i\right) \overset{(b)}{=} \sum_{i=1}^{\infty} P(C_i) \overset{(a)}{\leq} \sum_{i=1}^{\infty} P(A_i).$$

This proves Equation (2). Now let us use the trick above on a monotone increasing sequence $(A_n)_n$ in $\mathcal{A}$. In this case it holds that for all n,

$$\begin{aligned}\bigcup_{i=1}^{n} A_i &= \{\omega : \omega \in A_i \text{ for some } i\} \\ &= A_n.\end{aligned}$$

So in fact we have from (c) that $A_n = \bigcup_{i=1}^{n} C_i$ for all n. Since for any n also $P(A_n) = P\left(\bigcup_{i=1}^{n} C_i\right) = \sum_{i=1}^{n} P(C_i)$,

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

and, from countable additivity,

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

So indeed also (3) holds. It remains to proof (4): Let $(B_n)_n$ be any decreasing sequence in $\mathcal{A}$. Recall that,

$$\left(\bigcup_{n=1}^{\infty} B_n\right)^c = \bigcap_{n=1}^{\infty} B_n^c \quad \text{and} \quad \left(\bigcap_{n=1}^{\infty} B_n\right)^c = \bigcup_{n=1}^{\infty} B_n^c,$$

and that $A \subset B$ implies $B^c \subset A^c$. For monotone decreasing $(B_n)_n$, $B_1 \supset B_2 \supset \cdots$, and thus $B_1^c \subset B_2^c \subset \cdots$, so that

$$\lim_{n\to\infty} P(B_n^c) = P\left(\lim_{n\to\infty} B_n^c\right)$$

from the previous result. Then

$$\lim_{n\to\infty} P(B_n^c) = \lim_{n\to\infty} (1 - P(B_n)) = 1 - \lim_{n\to\infty} P(B_n)$$

and, from the above results,

$$P\left(\lim_{n\to\infty} B_n^c\right) = P\left(\bigcup_{n=1}^{\infty} B_n^c\right) = 1 - P\left(\bigcap_{n=1}^{\infty} B_n\right) = 1 - P\left(\lim_{n\to\infty} B_n\right),$$

so that

$$1 - \lim_{n\to\infty} P(B_n) = 1 - P\left(\lim_{n\to\infty} B_n\right).$$

■

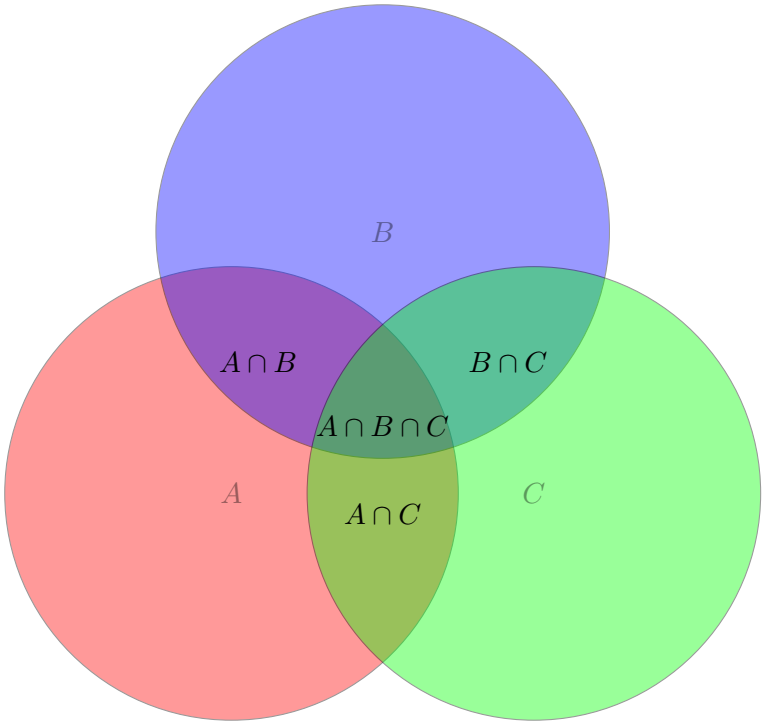

Figure 2: Illustration: Venn diagramm of 3 sets.

Equation (2) is equivalent to saying that $P(\cdot)$ is *subadditive* and is also referred to as Boole's inequality; if the A_i are disjoint, then equality holds. By taking complements of both sides (i.e., 1−), Booles's inequality can also be written as $P\left(\bigcap_{i=1}^{n} A_i^c\right) \geq 1 - \sum_{i=1}^{n} P(A_i)$. Equations (3) and (4) actually say that for an increasing sequence $(A_n)_n$, $\lim_n P(A_n) = P(\lim_n A_n)$ and for a decreasing sequence $(B_n)_n$, $\lim_n P(B_n) = P(\lim_n B_n)$. In other words the map $P : \mathcal{A} \to [0,1]$ is *continuous*!

Example 8. Let $A_n = \left[0, 1 + n^{-1}\right]$, $n = 1, 2, \ldots$. Show that $(A_n)_n$ is monotone and compute $L := \lim_{n\to\infty} A_n$. Let $B_n := A_n \setminus A_{n+1}$, $n = 1, 2, \ldots$. Express B_n as an interval and express A_n in terms of the B_i and L.

As $1/(n+1) < 1/n$, $A_n = [0, 1 + 1/n] \supset [0, 1 + 1/(n+1)] = A_{n+1}$ and so $(A_n)_n$ is monotone decreasing. So, $L = \lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i$ which, as $\lim_{n\to\infty} n^{-1} = 0$, is $[0,1]$. We have $B_1 = [0,2] \setminus [0,1.5] = (1.5, 2]$; $B_2 = (1 + 1/3, 1 + 1/2]$; and

$$B_n = [0, 1 + 1/n] \setminus [0, 1 + 1/(n+1)] = (1 + 1/(n+1), 1 + 1/n].$$

Also,

$$\begin{aligned} A_n &= \left[0, 1 + \frac{1}{n}\right] \\ &= [0,1] \cup \left(1 + \frac{1}{n+1}, 1 + \frac{1}{n}\right] \cup \left(1 + \frac{1}{n+2}, 1 + \frac{1}{n+1}\right] \cup \cdots \\ &= L \cup \bigcup_{i=n}^{\infty} B_i. \end{aligned}$$

It should be clear that L and the B_i are mutually exclusive.

$\Diamond$

One final note concerns sets with probability measure zero. Again take an arbitrary probability space $(\Omega, \mathcal{A}, P)$. As we have seen

$$A = \emptyset \implies P(A) = 0$$

However it is actually the case that other sets $A \in \mathcal{A}$, besides the empty set, may also have measure zero! That is $P(A) = 0$ does not necessarily imply $A = \emptyset$. As a very simple example consider the dirac measure from Example 7 at $x \in \mathbb{R}$, i.e. $(\Omega, \mathcal{A}, P) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \delta_x)$. Then *any* $A \in \mathcal{B}(\mathbb{R})$ with $x \notin A$ has measure zero. More important examples are given by the Gaussian measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, for which all single points (and in extension all countable sets) have measure zero.

2.3 Conditioning and Independence

In most applications, there will exist information which, when taken into account, alters the assignment of probability to events of interest. As a simple example, the number of customer transactions requested per hour from an on-line bank might be associated with an *unconditional* probability which was ascertained by taking the average of a large collection of hourly data. However, the *conditional* probability of receiving a certain number of transactions might well depend on the time of day, the arrival of relevant economic or business news, etc. If these events are taken into account, then more accurate probability statements can be made. Other examples include the number of years a manufacturing product will continue to work, conditional on various factors associated with its operation, and the batting average of a baseball player conditional on the opposing pitcher, etc. If $P(B) > 0$, then the *conditional probability of event A given the occurrence of event B*, or just the *probability of A given B*, is

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

This definition is motivated by observing that the occurrence of event B essentially reduces the relevant sample space, as indicated in the Venn diagram. The probability of A given B is the intersection of A and B, scaled by $P(B)$. If $B = \Omega$, then the scaling factor is just $P(\Omega) = 1$, which coincides with the unconditional case. If the occurrence or "non-occurrence" of event B does not influence that of A, and visa-versa, then the two events are said to be *independent*, i.e., $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$. From the definition of conditional probability, if events A and B are independent, then $P(A \cap B) = P(A) P(B)$. This is also referred to as *pairwise* independence. In general, events A_i, $i = 1, \ldots, n$ are (*completely*) independent if, and only if, for every collection $A_{i_1}, A_{i_2}, \ldots, A_{i_m}$, $1 \leq m \leq n$,

$$P(A_{i_1} A_{i_2} \cdots A_{i_m}) = \prod_{j=1}^{m} P\left(A_{i_j}\right).$$

For $n = 3$, this means that

$$P(A_1 A_2) = P(A_1) P(A_2),$$

$$P(A_1A_3) = P(A_1)\,P(A_3),$$
$$P(A_2A_3) = P(A_2)\,P(A_3),$$

and $P(A_1A_2A_3) = P(A_1)\,P(A_2)\,P(A_3)$. Clearly, pairwise independence and complete independence are the same thing if we have $n = 2$ sets. For $n > 2$ complete independence clearly implies pairwise independence. That pairwise independence does not imply mutual independence is referred to as *Bernstein's Paradox*. It is in fact easy to find a counterexample, such that three events A, B, C are mutually independent, but not fully independent: Take $\Omega = \{a, b, c, d\}$, $A = \{a, b\}$, $B = \{b, c\}$, $C = \{c, a\}$ and $P(\{\omega_i\}) = 1/4$ for $i = 1, \ldots, 4$. We have seen earlier that then

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = P(\{a\}) + P(\{b\}) = \frac{1}{2},$$

so that $P(A) = P(B) = P(C) = 1/2$. Then we have that $1/4 = P(A)\cdot P(B) = P(\{b\}) = P(A \cap B)$ and similarly for every other combination as above. Thus A, B, C are indeed pairwise independent. However

$$P(A \cap B \cap C) = P(\emptyset) = 0 \neq (1/4)^3 = P(A) \cdot P(B) \cdot P(C),$$

so A, B, C are not fully independent.

From a Venn diagram with (overlapping) events A and B, event A may be partitioned into mutually exclusive events $A \cap B$ and $A \cap B^c$, so that

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^c) \\ &= P(A \mid B)\,P(B) + P(A \mid B^c)\,(1 - P(B)). \end{aligned}$$

This is best understood as expressing $P(A)$ as a *weighted sum of conditional probabilities in which the weights reflect the occurrence probability of the conditional events*. In general, if events B_i, $i = 1, \ldots, n$ are exclusive and exhaustive, then the *law of total probability* states that

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A \mid B_i)\,P(B_i).$$

Example 9. Interest centers on the probability of getting at least three girls in a row among seven children. Assume that each child's sex is independent of the all the others and let $p = P(\text{girl on any trial})$. Denote the event that three girls in a row occur as R and the total number of girls as T. Then, from the law of total probability,

$$P(R) = \sum_{t=0}^{7} P(R \mid T = t)\,P(T = t).$$

Clearly, $P(R \mid T = t) = 0$ for $t = 0, 1, 2$ and $P(R \mid T = 6) = P(R \mid T = 7) = 1$. For $T = 3$, there are only 5 possible configurations, i.e.,

$$gggbbbb,\ bgggbbb, \ldots, bbbbggg,$$

so that

$$P(R \mid T=3) = \frac{5p^3(1-p)^4}{\binom{7}{3}p^3(1-p)^4} = \frac{5}{35}.$$

Some work shows that $P(R \mid T=4) = 16/\binom{7}{4} = 16/35$ and

$$P(R \mid T=5) = 18/\binom{7}{5} = 18/21,$$

so that

$$\begin{aligned} P(R) &= 0+0+0+\frac{5}{35}\binom{7}{3}p^3(1-p)^4 \\ &\quad +\frac{16}{35}\binom{7}{4}p^4(1-p)^3+\frac{18}{21}\binom{7}{5}p^5(1-p)^2 \\ &\quad +\binom{7}{6}p^6(1-p)+p^7 = 5p^3-4p^4-p^6+p^7. \end{aligned}$$

For $p=1/2$, $P(R) \approx 0.367$. ◊

From the law of total probability, *Bayes' rule* is given by

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B)+P(A \mid B^c)P(B^c)}.$$

For mutually exclusive and exhaustive events B_i, $i=1,\ldots,n$, the *general Bayes' rule* is given by

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{\sum_{i=1}^{n} P(A \mid B_i)\ \ P(B_i)}.$$

Example 10. A very important example of Bayes' rule is the following: A test for a disease possesses the following accuracy. If a person has the disease (event D), the test detects it 95% of the time; if a person does *not* have the disease, the test will falsely detect it 2% of the time. Let d_0 denote the *prior probability* of having the disease before the test is conducted. (This could be taken as an estimate of the proportion of the relevant population believed to have the disease). Assume that, using this test, a person is detected as having the disease (event $+$). To find the probability that the person actually has the disease, given the positive test result, we use Bayes' rule,

$$P(D \mid +) = \frac{0.95d_0}{0.95d_0+0.02(1-d_0)}.$$

For a rare disease such that $d_0 = 0.001$, $P(D \mid +)$ is only 0.045! There is evidence to suggest that many medical doctors are not capable of this calculation and vastly overestimate the probability of having a disease given a positive test result; see Gerd Gigerenzer's "Reckoning with Risk" (2002) for numerous examples and some of the social and economic implications of this. To vastly aid understanding of Bayes' rule in this context, Gigerenzer recommends expressing things not in terms of probabilities, but rather in "natural frequencies". (Gigerenzer, 2002, p. 41)

Consider posing the following question to a physician (let alone a layperson): Referring to asymptomatic (when the patient does not experience any noticeable symptoms) women aged 40

to 50 undergoing a routine mammography screening: The probability that one of these women has breast cancer is 0.8 percent. If a woman has breast cancer, the probability is 90 percent that she will have a positive mammogram. If a woman does not have breast cancer, the probability is 7 percent that she will still have a positive mammogram. Imagine a woman who has a positive mammogram. What is the probability that she actually has breast cancer?

Bayes' rule of course gives the answer: With $d_0 = 0.008$,

$$P(D \mid +) = \frac{0.90 d_0}{0.90 d_0 + 0.07 (1 - d_0)} = 0.094,$$

i.e., less than a 1 in 10 chance! When the question is posed as above, it is not obvious for most people to apply Bayes' rule.

$\Diamond$

3 Random variables

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(S, \mathcal{S})$ an arbitrary measurable space (think of $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ or $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Recall the definition of the inverse image of a function $f : \Omega \to S$: For any $B \subset S$

$$f^{-1}(B) = \{\omega \in \Omega : f(\omega) \in B\}.$$

In words this a set of points $\omega \in \Omega$, such that $f(\omega) \in B$ holds true. Very useful properties of the inverse image follow almost immediately from the definition above. For any sets $(B_n)_{n\in\mathbb{N}}$,

$$f^{-1}(\bigcup_{n=1}^{\infty} B_n) = \bigcup_{n=1}^{\infty} f^{-1}(B_n) \tag{5}$$

$$f^{-1}(\bigcap_{n=1}^{\infty} B_n) = \bigcap_{n=1}^{\infty} f^{-1}(B_n) \tag{6}$$

$$f^{-1}(B^c) = f^{-1}(B)^c \tag{7}$$

In fact (5) and (6) hold for *arbitrary* unions and intersections (i.e. even over an uncountable number of sets). But we are satisfied with union/intersections of finite or countably infinite families of sets. Now let us look at random maps from the most general perspective:

3.1 General Situation

In mathematics "random" elements (random variables, vectors, or even random functions) are simply so-called measurable functions:[7]

Definition 3.1. A function $f : \Omega \to S$ is called measurable (or more precisely $\mathcal{A}/\mathcal{S}$-measurable) if

$$f^{-1}(B) \in \mathcal{A} \text{ for all } B \in \mathcal{S}.$$

For example:

Definition 3.2. If $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $X : \Omega \to \mathbb{R}$ is $\mathcal{A}/\mathcal{B}(\mathbb{R})$ measurable, then X is called *random variable*. If for some natural number $d > 1$, $(S, \mathcal{S}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $\mathbf{X} : \Omega \to \mathbb{R}^d$ is $\mathcal{A}/\mathcal{B}(\mathbb{R}^d)$-measurable, then $\mathbf{X}$ is called *random vector*. If $(S, \mathcal{S})$ is a space of functions with some σ-algebra $\mathcal{S}$ (for example $\ell^\infty(T)$ the metric space of all bounded real-valued functions on some space T, together with $\mathcal{B}(\ell^\infty(T))$) and $\psi : \Omega \to S$ is $\mathcal{A}/\mathcal{S}$-measurable, then ψ is called a *random function*.

The condition in Definition 3.1 might seem utterly strange. We will now explore why it is important. Notice that we have a probability measure on $(\Omega, \mathcal{A})$, namely P, but none on $(S, \mathcal{S})$. Given a measurable function $X : \Omega \to S$, how could we find a probability measure on $(S, \mathcal{S})$? It turns out that for any $B \in \mathcal{S}$,

$$\mu_X(B) := P(X^{-1}(B)) \tag{8}$$

[7] Obviously nothing in mathematics is "random" in the usual meaning of the word.

is a measure on $(S, \mathcal{S})$. First of all $P(A)$ is only defined for $A \in \mathcal{A}$. But that is no problem, since by measurability, $X^{-1}(B) \in \mathcal{A}$. So μ_X is defined on $(S, \mathcal{S})$. Let us now formally show that it is indeed a probability measure:

- For any $B \in \mathcal{S}$, $\mu_X(B) = P\left(X^{-1}(B)\right) \geq 0$.
- $\mu_X(S) = P\left(X^{-1}(S)\right) = P(\{\omega \in \Omega : f(\omega) \in S\}) = P(\Omega) = 1$
- Finally if $(B_n)_{n\in\mathbb{N}}$ is a sequence of disjoint sets in $\mathcal{S}$, then $\bigcup_n B_n \in \mathcal{S}$ as well (σ-algebra) and it also

$$\mu_X\left(\bigcup_{n=1}^{\infty} B_n\right) = P\left(X^{-1}\left(\bigcup_{n=1}^{\infty} B_n\right)\right) \overset{(*)}{=} P\left(\bigcup_{n=1}^{\infty} X^{-1}\left(B_n\right)\right) = \sum_{n=1}^{\infty} \mu_X(B_n),$$

  since also the sequence $(X^{-1}(B_n))_{n\in\mathbb{N}}$ is disjoint: Take any two sets $X^{-1}(B_j)$, $X^{-1}(B_i)$, $i \neq j$. If $\omega \in X^{-1}(B_j)$, then by definition $X(\omega) \in B_j$ but since $B_j \cap B_i = \emptyset$, $X(\omega) \notin B_i$ or $\omega \notin X^{-1}(B_i)$. Since $\omega \in X^{-1}(B_j)$ was arbitrary, this means also $X^{-1}(B_j) \cap X^{-1}(B_i) = \emptyset$ and since in turn $X^{-1}(B_i)$ and $X^{-1}(B_j)$ were arbitrarily chosen from $(X^{-1}(B_n))_{n\in\mathbb{N}}$, this means the sequence is disjoint as well.

To proof the last point we made the bold claim that $P\left(X^{-1}\left(\bigcup_n B_n\right)\right) = P\left(\bigcup_n X^{-1}\left(B_n\right)\right)$ in (*). This of course simply follows from (6) above, but we will quickly verify this: If $\omega \in X^{-1}\left(\bigcup_n B_n\right)$, then $X(\omega) \in \bigcup_n B_n$. By definition that means there exists at least one B_n such that $X(\omega) \in B_n$. But then $\omega \in X^{-1}(B_n)$ and thus $\omega \in \bigcup_n X^{-1}\left(B_n\right)$. Thus

$$X^{-1}\left(\bigcup_n B_n\right) \subset \bigcup_n X^{-1}\left(B_n\right).$$

We can make very similar arguments in the other direction, to see that

$$X^{-1}\left(\bigcup_n B_n\right) \supset \bigcup_n X^{-1}\left(B_n\right).$$

So in fact

$$X^{-1}\left(\bigcup_n B_n\right) = \bigcup_n X^{-1}\left(B_n\right),$$

and thus their probabilities are also the same. So X induces a new probability measure μ_X on $(S, \mathcal{S})$. We call this measure the *distribution* of X. We will see many examples of distributions for the case $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ later. In fact for the important special case $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (the random variable we usually look at), it can be shown that the above condition of measurability is equivalent to

$$X^{-1}((-\infty, a]) \in \mathcal{A} \text{ for all } a \in \mathbb{R}. \tag{9}$$

So it is enough to show the above condition for X to be measurable. Why does that help? The σ-algebra $\mathcal{B}(\mathbb{R})$ is already an unwieldy object, we don't really know what is in there, it is simply

too big. So for a given map $X : \Omega \to \mathbb{R}$ checking directly whether $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}(\mathbb{R})$ is next to impossible. In contrast, condition (9) is often relatively easy to verify.

Let us have a look at examples for which we will check measurability:

Example 11. Let $(\Omega, \mathcal{A})$ be an arbitrary probability space and $X : \Omega \to \mathbb{R}$ take only two values,

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

for some $A \in \mathcal{A}$ (you could think of $\mathcal{A} = \mathcal{B}(\mathbb{R})$ and $A = (-\infty, a]$, for instance).

Then it holds for any $B \in \mathcal{B}(\mathbb{R})$ that

$$X^{-1}(B) = \begin{cases} A, & \text{if } 1 \in B, 0 \notin B \\ A^c, & \text{if } 1 \notin B, 0 \in B \\ \Omega, & \text{if } 1 \in B, 0 \in B \\ \emptyset & \text{if } 1 \notin B, 0 \notin B \end{cases}$$

Every set above is in $\mathcal{A}$: A by assumption, and $\Omega, \emptyset$ and A^c because $\mathcal{A}$ is a σ-algebra. So X is indeed $\mathcal{A}/\mathcal{B}(\mathbb{R})$ measurable.

$\Diamond$

We also consider a more advanced (and important) example:

Example 12. Let $(\Omega, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $f : \mathbb{R} \to \mathbb{R}$ be monotone increasing (i.e. $x \leq y$ implies $f(x) \leq f(y)$), then f is $\mathcal{B}(\mathbb{R})/\mathcal{B}(\mathbb{R})$ measurable.

This is one of the cases where checking measurability directly is doomed to fail. However, if we look at the inverse image of an interval, we see that, for any $a \in \mathbb{R}$, either

$$f^{-1}((-\infty, a]) = (-\infty, u)$$

or

$$f^{-1}((-\infty, a]) = (-\infty, u],$$

for some $u \in \mathbb{R}$. That is we know that the inverse image of an interval is an interval again, though we don't know whether it is open or half open! (This depends on the further properties of f). We will prove this in a minute, first assume that this is indeed true. It was said before that any interval (closed, open, or half-closed, half-open) is element of $\mathcal{B}(\mathbb{R})$. Consequently,

$$f^{-1}((-\infty, a]) \in \mathcal{B}(\mathbb{R})$$

for *any* $a \in \mathbb{R}$. In other words f is indeed measurable.

To proof the claim above, remember that f is monotone increasing, and that $f^{-1}((-\infty, a]) = \{x \in \mathbb{R} : f(x) \leq a\}$. So, whenever $x_1 \leq x_2 \in \mathbb{R}$ and $x_2 \in f^{-1}((-\infty, a])$, then

$$f(x_1) \underbrace{\leq}_{\text{monotonicity}} f(x_2) \underbrace{\leq}_{x_2 \in f^{-1}((-\infty, a])} a.$$

In other words also $x_1 \in f^{-1}((-\infty, a])$. Since x_1, x_2 above were arbitrary, we have just shown that whenever $x \in f^{-1}((-\infty, a])$, then also $y \in f^{-1}((-\infty, a])$ for any $y \leq x$. It is not hard to see that this means we deal with an interval stretching to $-\infty$. However this interval might have the form $(-\infty, u)$ or $(-\infty, u]$, there is no way of telling without knowing more about f. By the way, $u = \sup\{x : f(x) \leq a\}$, but again, we cannot say more without assuming more about f. $\Diamond$

The above example highlights the reason we need not think about measurability much more: It is really hard to construct a map that is not measurable! For instance, together with increasing functions, any continuous function $f : \mathbb{R} \to \mathbb{R}$ is $\mathcal{B}(\mathbb{R})/\mathcal{B}(\mathbb{R})$ measurable. (In fact any continuous function from a topological space $\mathcal{X}$ to another topological space $\mathcal{Y}$ is $\mathcal{B}(\mathcal{X})/\mathcal{B}(\mathcal{Y})$ measurable!) Moreover, many operations involving measurable functions result in measurable functions again ($f = \max(f_1, f_2)$ for instance). This is why in practice we often don't need to check measurability. Though one should remember that this is always meant when we talk about "random" in a mathematical sense. Lastly, there are important cases when measurability indeed breaks down, in the field of stochastic processes for instance. However this goes beyond the scope of this course. Those interested may for instance consult the monumental work of van der Waart and Wellner (1996). Let us give a final example in which we check measurability in a quite general case:

Example 13. Let $(\Omega, \mathcal{A})$, $(S_1, \mathcal{S}_1)$, $(S_2, \mathcal{S}_2)$ be arbitrary measurable spaces, $X : \Omega \to S_1$ be $\mathcal{A}/\mathcal{S}_1$ measurable and $F : S_1 \to S_2$ be $\mathcal{S}_1/\mathcal{S}_2$ measurable. Then the composition $F \circ X = F(X)$ is $\mathcal{A}/\mathcal{S}_2$ measurable. We could also say in less precise terms that if X is a random object and F is a measurable function, then $F(X)$ is a random object again.

This claim is actually easy to check: By measurability of X, we have

$$X^{-1}(B) \in \mathcal{A} \tag{10}$$

for all $B \in \mathcal{S}_1$. In the same way the measurability of F means

$$F^{-1}(C) \in \mathcal{S}_1 \tag{11}$$

for all $C \in \mathcal{S}_2$. Importantly, one can also check that

$$\begin{aligned}(F \circ X)^{-1}(C) &= \{\omega : F \circ X(\omega) \in C\} \\ &= \{\omega : F(X(\omega)) \in C\} \\ &= X^{-1}(F^{-1}(C))\end{aligned} \tag{12}$$

Putting (10) (11) and (12) together immediately gives

$$(F \circ X)^{-1}(C) \in \mathcal{A}$$

for all $C \in \mathcal{S}_2$, or that $F(X)$ is $\mathcal{A}/\mathcal{S}_2$ measurable. $\Diamond$

Finally, note that we have just given meaning to something that is usually very ill-defined in a first (or even second and third) statistics course: Expressions like $P(X < 0)$ or $P(5 \leq X \leq 10)$ or even $P(X \in A)$, do not make any sense if not defined further. After all P is defined on subsets

of Ω, so what should $P(X > 0)$ even mean? Well, it is actually just a shorthand, for what we defined before, namely

$$P(X < a) := P(X^{-1}((-\infty, a)) = \mu_X((-\infty, a))$$

or in general

$$P(X \in A) := P(X^{-1}(A)) = \mu_X(A),$$

for any $A \in \mathcal{B}(\mathbb{R})$. We already used this in Example 2, without explicitly stating it.

We conclude this introduction into very general random elements by focusing on $\Omega = \mathbb{R}$ and (almost) fully proving a monumental important result in Theorem 3.2. Let $X : \Omega \to \mathbb{R}$ be a $\mathcal{A}/\mathcal{B}(\mathbb{R})$ measurable map (i.e. a random variable). We then know that X induces a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and that this is it's distribution. However this is again a very complicated object, defined on all of $\mathcal{B}(\mathbb{R})$! This is impractical, we would like to have a better characterization of such a distribution. For instance, it would be nice to just have one formula characterizing the whole probability measure (which we will see is what we use in almost all practical cases). There are in fact many ways of characterizing the distribution, the first (and one of the most important ones), is through the cummulative distribution function (cdf):

Definition 3.3 (cdf)**.** The function $F : \mathbb{R} \to [0, 1]$,

$$F(x) = P(X \leq x) = \mu_X((-\infty, x])$$

is called the cummulative distribution function (cdf) of X.

Theorem 3.1. *Any cdf $F : \mathbb{R} \to [0, 1]$, has*

(i) F is monotone increasing ($x \leq y$ implies $F(x) \leq F(y)$),

(ii) $\lim_{x \to +\infty} F(x) = 1$, $\lim_{x \to -\infty} F(x) = 0$,

(iii) F is right-continuous (for any $y \in \mathbb{R}$: $\lim_{x \downarrow y} F(x) = F(y)$*).*

Proof. Recall the continuity of the probability measure as proven in Theorem 2.2.

(i) Whenever $x \leq y$, $(-\infty, x] \subset (-\infty, y]$ and by monotonicity of μ_X also $F(x) \leq F(y)$.

(iii) Take an arbitrary sequence $x_n \downarrow x$ and $A_n := (-\infty, x_n]$. Then $(A_n)_n$ is decreasing with limit

$$\lim_n A_n = \bigcap_{n=1}^{\infty} (-\infty, x_n] = (-\infty, x].$$

Thus with the continuity of μ_X

$$F(x) = \mu_X((-\infty, x]) = \mu_X(\lim_n (-\infty, x_n]) = \lim_n \mu_X((-\infty, x_n]) = \lim_n F(x_n).$$

So for any sequence $(x_n)_n$ in $\mathbb{R}$ with $x_n \downarrow x$, $\lim_n F(x_n) = F(x)$. This is actually equivalent to saying that $\lim_{x \downarrow y} F(x) = F(y)$.

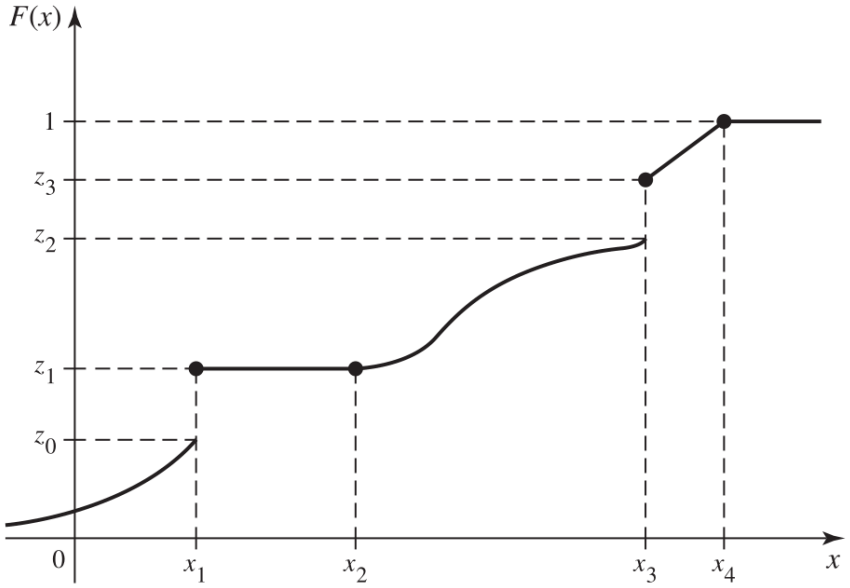

Figure 3: Example of a possible cdf. Source: Internet.

(ii) Same as (ii), but now using arbitrary sequences $(x_n)_n$ and $(y_n)_n$ in $\mathbb{R}$ with $x_n \uparrow +\infty$ and $y_n \downarrow -\infty$, so that $\lim_n (-\infty, x_n] = \bigcup_n (-\infty, x_n] = \mathbb{R}$ and $\lim_n (-\infty, y_n] = \bigcap_n (-\infty, y_n] = \emptyset$.

■

Note that since $F : \mathbb{R} \to \mathbb{R}$ is monotone increasing by (i), it is also measurable by Example 12! This in turn means with Example 13 that if $X : \Omega \to \mathbb{R}$ is a random variable, then the composition $F \circ X = F(X)$ is a random variable again. We will quickly study the distribution of this random variable in Section 3.2.

We will see many examples of important cdf's in the next section. The great advantage of this construction is that any random variable admits a cdf (contrary to a pdf which might not exist). The only other construction that is this general is the characteristic function (cf). What is more, any cdf-looking function (i.e. any function meeting conditions (i)-(iii) in 3.1) indeed completely characterizes a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Theorem 3.2 (Lebesgue-Stieljes)**.** *For any function $F : \mathbb{R} \to [0, 1]$ satisfying (i)-(iii) in Theorem 3.1 there exists a* unique *probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with:*

$$F(x) = \mu((-\infty, x]), \text{ for all } x \in \mathbb{R} \tag{13}$$

Proof. We will only proof existence here. Showing uniqueness is important, but not difficult, though we would need some auxiliary theorems for this. These can be found for instance in Durrett (2010).

Thus assume that we have a function $F : \mathbb{R} \to \mathbb{R}$ with properties (i) - (iii) from Theorem 3.1. The idea of the proof is to construct a new function on a suitable probability space and to verify that this is a r.v. Then μ is taken to be its distribution, and we also have to check that

indeed (13) holds. Take $\Omega = (0,1)$, $\mathcal{A} = \mathcal{B}((0,1))$, $P =$ Lebesgue measure on $(0,1)$. This in fact the same as saying that P has the *uniform distribution* as introduced in the next section, i.e. for $0 < a \leq b < 1$,

$$P((a,b)) = P((a,b]) = P([a,b)) = P([a,b]) = b - a.$$

Now define for $\omega \in (0,1)$:

$$X(\omega) = \sup\{y \in \mathbb{R} : F(y) < \omega\}, \tag{14}$$

so that $X : \Omega \to \mathbb{R}$. If F is strictly increasing (i.e. invertible) then this is just the inverse $F^{-1}(\omega)$. However this general formulation allows to make the proof for *any* random variable/distribution and any cdf F. Let us first study the set,

$$F^{-1}((-\infty, \omega)) = \{y \in \mathbb{R} : F(y) < \omega\},$$

for any $\omega \in (0,1)$. Then how does this set look like? As we have seen, this heavily depends on the properties of F: First F is increasing, so we know from the arguments in Example 12, that in fact $F^{-1}((-\infty, \omega))$ is an interval stretching to $-\infty$: $(-\infty, u)$ or $(-\infty, u]$, for $u = X(\omega)$. Now assume that $F^{-1}((-\infty, \omega)) = (-\infty, u]$. Then this means by definition that $F(u) < \omega$ and $F(y) \geq \omega$, for all $y > u$. However F is right-continuous as well, so there is a small $\varepsilon > 0$ such that also $F(u + \varepsilon) < \omega$ (this follows directly from the definition of right continuity). But this means that $u + \varepsilon \in F^{-1}((-\infty, \omega))$, a contradiction. So in fact the interval must be open:

$$F^{-1}((-\infty, \omega)) = (-\infty, u).$$

Now having this issue out of the way we can proof that X is indeed a random variable. The following equality makes this possible:

$$\forall y : \ \{\omega \in (0,1) : X(\omega) \leq y\} = \{\omega \in (0,1) : \omega \leq F(y)\}. \tag{15}$$

(Again for intuition, think of X as the inverse of F) Let us assume (15) to be true, before proving it. Then for all $y \in \mathbb{R}$,

$$\begin{aligned} X^{-1}((-\infty, y]) &= \{\omega \in (0,1) : X(\omega) \leq y\} \\ &= \{\omega \in (0,1) : \omega \leq F(y)\} \\ &= \begin{cases} (0, F(y)], & \text{if } F(y) < 1 \\ (0,1), & \text{if } F(y) = 1, \end{cases} \end{aligned}$$

i.e. X is indeed measurable (and thus a random variable), since all intervals in $(0,1)$ are elements of the Borel σ-algebra on $(0,1)$.[8] So we may define $\mu(A) = \mu_X(A) = P(X \in A)$ for all $A \in B(\mathbb{R})$ and get our probability measure. What is more, with the same argument as above we have for

[8] Again recall that this condition of measurability is sufficient, we do not need to check $X^{-1}(A) \in B((0,1))$ for all $A \in \mathcal{B}(\mathbb{R})$

all $x \in \mathbb{R}$,

$$\begin{aligned}\mu((-\infty, x]) = P(X \leq x) &= P(\{\omega \in (0,1) : X(\omega) \leq x\}) \\ &= P(\{\omega \in (0,1) : \omega \leq F(x)\}) \\ &= F(x),\end{aligned}$$

by the definition of P. So indeed also (13) is proven.

It remains to proof (15). For this it helps to remember that $F^{-1}((-\infty, \omega)) = (-\infty, X(\omega))$ as demonstrated above. Then for $y \in \mathbb{R}$ arbitrary:

"$\supset$" If $\omega \in \{\omega \in (0,1) : \omega \leq F(y)\}$, $0 \leq \omega \leq F(y)$, thus $y \notin \{y \in \mathbb{R} : F(y) < \omega\} = (-\infty, X(\omega))$ and therefore $X(\omega) \leq y$.

"$\subset$" We show $\{\omega \in (0,1) : X(\omega) \leq y\} \subset \{\omega \in (0,1) : \omega \leq F(y)\}$, by showing

$$\omega \notin \{\omega \in (0,1) : \omega \leq F(y)\} \implies \omega \notin \{\omega \in (0,1) : X(\omega) \leq y\}.$$

So if $\omega > F(y)$ then by definition $y \in F^{-1}((-\infty, \omega)) = (-\infty, X(\omega))$, or $y < X(\omega)$, which means $\omega \notin \{\omega \in (0,1) : X(\omega) \leq y\}$.

■

Let us put the above result in context: In Theorem 2.1 it was demonstrated that for Ω countable, all you need to find a new probability on $(\Omega, 2^\Omega)$ (and to completely characterize it), is a collection of real numbers $(p_\omega)_{\omega \in \Omega}$ with $p_\omega \geq 0$ and $\sum_{\omega \in \Omega} p_\omega = 1$. Theorem 3.2 on the other hand tells you that for $(\Omega, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ it is enough to find the cdf F to build a new probability measure (and to completely characterize it). In particular, if we have two probabilities μ_1 and μ_2 on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with cdfs F_1, F_2, then we know by the uniqueness statement in theorem 3.2, that: $F_1 = F_2 \implies \mu_1 = \mu_2$. So all we need to proof that $\mu_1(A) = \mu_2(A)$ for all $A \in \mathcal{B}(\mathbb{R})$, is to show that $F_1(x) = F_2(x)$ for all $x \in \mathbb{R}$. This will be used a few times below. In fact the cdf is also (sometimes) enough to simulate from this distribution:

Remark 2. The above proof is *constructive* (those are often the best proofs): It theoretically gives us a way to simulate a random variable for any given distribution, as soon as we have the cdf. After all, there is no actual randomness in computers, similarly as there is no actual randomness in mathematics. So simulating something random with (deterministic) computers is not at all trivial. However the above proof gives us a principled way of doing this, as long as we can simulate from a uniform distribution: Let F be the cdf of a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ we want to simulate from and $(\Omega, \mathcal{A}, P)$ the underlying probability space. Let also U be a uniform random variable, that is $U : \Omega \to (0,1)$ with $\mu_U =$ Lebesgue measure on $(0,1)$. Then we may simply use U to get to the probability space from above, that is U provides a bridge between $(\Omega, \mathcal{A}, P)$ and $((0,1), \mathcal{B}((0,1)), \mu_U)$. If we then define for each $\omega \in \Omega$, $U(\omega) = \tilde{\omega} \in (0,1)$, we may define

$$X(U(\omega)) = X(\tilde{\omega}) = \sup\{y \in \mathbb{R} : F(y) < \tilde{\omega}\}.$$

As we have demonstrated above, $X : ((0,1), \mathcal{B}((0,1))) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a random variable and its distribution is μ! Thus the proof of Theorem 3.2 tells us that as long as we can simulate from a uniform distribution, we are able to generate any distribution for which we are able to evaluate the inverse cdf. (*Every* distribution has a cdf, though we might not be able to write it analytically or invert it). $\lozenge$

Remark 3. The construction in (14) is also intimately related to the *quantile function*:

$$Q(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\} \tag{16}$$

for $p \in [0,1]$. In fact the proof above would have worked similarly using (16) instead of (14). If F is continuous and strictly increasing (i.e. invertible), then

$$\inf\{x \in \mathbb{R} : F(x) \geq p\} = \sup\{x \in \mathbb{R} : F(x) < p\} = F^{-1}(p).$$

The quantile function is of utmost importance in many (statistical) applications. In statistics one usually talks about the quartile to describe a distribution: The 25% and 75% quartile are Q(0.25) and Q(0.75) respectively. The *median* is the 50% quartile, Q(0.5). An important example in Finance is the Value at Risk (VaR) at level α, which is simply the negative of the α-quantile: $VaR_\alpha = -Q(\alpha)$. See e.g. McNeil et al. (2005).

Let us give a small example of the quantile function, in case of a cdf that is not invertible:[9] Define a measure on the finite measurable space $\Omega = \{0,1,2\}$, $\mathcal{A} = 2^\Omega$ as

$$P(\{\omega\}) = \begin{cases} 1/2, & \text{if } \omega = 0 \\ 1/4, & \text{if } \omega = 1 \\ 1/4, & \text{if } \omega = 2 \end{cases}.$$

Moreover we define $X : \Omega \to \mathbb{R}$, as the identity function: $X(0) = 0$, $X(1) = 1$, $X(2) = 2$. Then for all $A \in \mathcal{B}(\mathbb{R})$ (simply by the definition of μ_X),

$$\mu_X(A) = \sum_{i=1}^{3} P(\{\omega_i\})\delta_{X(\omega_i)}(A) = 1/2 \cdot \delta_0(A) + 1/4 \cdot \delta_1(A) + 1/4 \cdot \delta_2(A),$$

where $\delta_x(A)$ is the dirac measure at x, as in Example 7. The distribution function of μ_X is then given as

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1/2, & \text{if } 0 \leq x < 1 \\ 3/4, & \text{if } 1 \leq x < 2 \\ 1, & \text{if } x \geq 2 \end{cases}.$$

We now want to find $Q(0.2)$. The smallest x such that $F(x) \geq 0.2$ is given by $x = 0$, since $F(x) \geq 1/2$ for $x \geq 0$ and $F(x) = 0 < 0.2$ as soon as $x < 0$. $\lozenge$

[9] Adapted from the statlect site (`https://www.statlect.com/fundamentals-of-probability/quantile`).

3.2 Continuous Distributions

Recall the "intuitive" notion of a continuous random variable in Section 1. Formally speaking, a random variable X is continuous, if it admits a density:

Definition 3.4 (Definition 11.2 in Jacod and Protter (2004))**.** The density of a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a nonnegative $\mathcal{B}(\mathbb{R})/\mathcal{B}(\mathbb{R})$ measurable function $f : \mathbb{R} \to \mathbb{R}$ that has for all $x \in \mathbb{R}$:

$$\mu((-\infty, x]) = \int_{-\infty}^{x} f(y)dy. \tag{17}$$

If μ is the distribution of a random variable X, then f is called the *density* or *pdf* of X.

Remark 4. A note to the integral expression in (17): In this lecture you can think of this as an ordinary Riemann integral, which one can calculate with the tools used in first analysis courses. However it is actually a more general integral with respect to a *measure*: That is, let ν be an arbitrary measure on $(\Omega, \mathcal{A})$, and let $f : \Omega \to \mathbb{R}$ be $\mathcal{A}/\mathcal{B}(\mathbb{R})$ measurable and nonnegative. Then we are able to define the integral $\int_\Omega f d\nu$. We want to stress here, that measurability and nonnegativity are enough assumptions on f to define the integral, no continuity or any other strong assumptions are needed. If furthermore f is not necessarily nonnegative, then $|f|$ is, and we can define

$$\int_\Omega f d\nu,$$

as long as $\int_\Omega |f| d\nu < \infty$. All of this is detailed in Section 4.1, for ν being a probability measure P. In this lecture we usually take $(\Omega, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (or $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ later) and ν to be the very important Lebesgue measure λ, the unique measure assigning to any interval its length: $\lambda((a, b)) = \lambda([a, b]) = \lambda((a, b]) = \lambda([a, b)) = b - a$. In any case, the integral with respect to the Lebesgue measure luckily is the same than the Riemann integral whenever both exists. Since we will look only at densities f which are Riemann integrable, it is enough to see $\int_{-\infty}^{x} f(y)dy$ as a Riemann integral. However, there are a few nice properties of the Lebesgue integral that we will need throughout the course. For instance, if we look at the Lebesgue integral, then we can actually define the integration over *any set* $A \in \mathcal{B}(\mathbb{R})$, that is

$$\int_A f d\lambda = \int_A f(x)dx := \int_\Omega f \mathbb{I}_A d\lambda,$$

where $\mathbb{I}_A(x)$ is 1 if $x \in A$ and zero otherwise. It is then the case that we can find (in principle, it is not really clear how to calculate this in practice) the probability of $A \in \mathcal{B}(\mathbb{R})$ as:

$$\mu(A) = \int_A f d\lambda = \int_A f(x)dx = \int_\Omega f(x)\mathbb{I}_A(x)dx \tag{18}$$

So the measure $\mu(A)$ of any set $A \in \mathcal{B}(\mathbb{R})$ is just the integral over A. Let us quickly proof that this is true, if we take the fact that we can define an integral over A for granted. Define for all $A \in \mathcal{B}(\mathbb{R})$ the set function,

$$\mu_2(A) := \int_A f d\lambda = \int_A f(x)dx.$$

In Section 3.4 we show that $\mu_2(A)$ is indeed a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (i.e. $\mu_2(A) \geq 0$, $\mu_2(\mathbb{R}) = 1$ and countable additivity). Assuming this to be true, we would like to proof that $\mu(A) = \mu_2(A)$ for all $A \in \mathcal{B}(\mathbb{R})$, or $\mu = \mu_2$. But this equality is simple with Theorem 3.2: Let F be the cdf of μ and F_2 the one of μ_2 (both are probabilities on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ so they have a cdf). Then, for all $x \in \mathbb{R}$,

$$F_2(x) = \mu_2((-\infty, x)) = \int_{(-\infty,x)} f \, d\lambda = \int_{-\infty}^{x} f(y) \, dy = F(x).$$

So in fact the cdf's F, F_2 are the same, and we thus immmediately know that the measures μ, μ_2 are the same.

We stress again that (18) is well defined for f as in Definition 17, though it might be not clear how to actually calculate it in practice. Finally, note that we did not assume f to be continuous, but only measurable (again this comes out of the fact that we use this very general integral expression)! However we deal exclusively with densities that are continuous. In this case it turns out, as mentioned in Section 1, that indeed $F' = f$, i.e. the derivative of the cdf gives the density back. ◊

Theorem 3.3 (Theorem 11.3 in Jacod and Protter (2004)). *A nonnegative $\mathcal{B}(\mathbb{R})/\mathcal{B}(\mathbb{R})$ measurable function $f : \mathbb{R} \to \mathbb{R}$ is the density of a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ iff it satisfies*

$$\int_{-\infty}^{+\infty} f(y)dy = 1. \tag{19}$$

In this case f entirely characterizes the probability measure.

Proof. For the more interesting direction, assume f has $f \geq 0$ and $\int_{-\infty}^{+\infty} f(y)dy = 1$. Define

$$F(x) := \int_{-\infty}^{x} f(t)dt, \quad \text{for all } x \in \mathbb{R}. \tag{20}$$

We want to proof that there exists a probability μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with f being the density of μ. To do this, we can use the properties of the Lebesgue integral, treated in Section 4.1, to show that F meets condition (i)-(iii) in Theorem 3.1. First, we may use $f \geq 0$ and the monotonicity of the integral to show monotonicity of F. That is, for $x_1 \leq x_2$, we have $\mathbb{I}_{(-\infty,x_1]}(y) \leq \mathbb{I}_{(-\infty,x_2]}(y)$ and thus $\mathbb{I}_{(-\infty,x_1]}(t)f(y) \leq \mathbb{I}_{(-\infty,x_2]}(y)f(y)$ for all $y \in \mathbb{R}$ and,

$$F(x_1) := \int_{-\infty}^{x_1} f(y)dy = \int_{-\infty}^{+\infty} f(y)\mathbb{I}_{(-\infty,x_1]}(y)dy \leq \int_{-\infty}^{+\infty} f(y)\mathbb{I}_{(-\infty,x_2]}(y)dy = F(x_2).$$

Next, we can use the dominated convergence theorem (Theorem 4.4) to show that F is right-continuous: Let $(x_n)_n$ be a decreasing sequence converging to $x \in \mathbb{R}$. Then the sequence of functions $(f\mathbb{I}_{(-\infty,x_n]})_{n\in\mathbb{N}}$ converges pointwise to $f\mathbb{I}_{(-\infty,x]}$, that is

$$\lim_{n\to\infty} f(y)\mathbb{I}_{(-\infty,x_n]}(y) = f(y)\mathbb{I}_{(-\infty,x]}(y) \text{ for all } y \in \mathbb{R}.$$

Thus we have that,

$$\lim_n F(x_n) = \lim_n \int_{-\infty}^{x_n} f(y)dy = \lim_n \int_{-\infty}^{+\infty} f(y)\mathbb{I}_{(-\infty,x_n]}(y)dy = \int_{-\infty}^{+\infty} f(y)\mathbb{I}_{(-\infty,x]}(y)dy = F(x),$$

as a consequence of the dominated convergence theorem.[10] Thus we have shown that for any decreasing sequence $(x_n)_{n\in\mathbb{N}}$ with a limit $x \in \mathbb{R}$, $\lim_n F(x_n) = F(x)$, or F is right-continuous. With exactly the same procedure, we can also show that F is *left*-continuous: Let $(x_n)_n$ be an increasing sequence converging to $x \in \mathbb{R}$. Then the sequence of functions $(f\mathbb{I}_{(-\infty,x_n]})_{n\in\mathbb{N}}$ converges pointwise to $f\mathbb{I}_{(-\infty,x)}$. Thus we have that,

$$\lim_n F(x_n) = \lim_n \int_{-\infty}^{+\infty} f(y)\mathbb{I}_{(-\infty,x_n]}(y)dt = \int_{-\infty}^{+\infty} f(y)\mathbb{I}_{(-\infty,x)}(y)dy = \int_{-\infty}^{x} f(y)dy = F(x).$$

Using the same tricks again (even with the same dominating function $g = f$), we obtain that for an increasing sequence $(x_n)_n$, $x_n \to +\infty$,

$$\lim_n F(x_n) = \lim_n \int_{-\infty}^{x_n} f(y)dy = \int_{-\infty}^{+\infty} f(y)dy = 1,$$

and for a decreasing sequence with $(x_n)_n$, $x_n \to -\infty$,

$$\lim_n F(x_n) = \lim_n \int_{-\infty}^{x_n} f(y)dy = \int_{-\infty}^{+\infty} \lim_n f(y)\mathbb{I}_{(-\infty,x_n]}(y)dy = 0,$$

since $\mathbb{I}_{(-\infty,x_n]}(y) \to 0$ for all y. Thus, we have shown that f defines a cdf F through the integral in (20). Again by Theorem 3.2, this means there exists a unique measure μ, such that for all $x \in \mathbb{R}$:

$$\mu((-\infty, x]) = F(x) = \int_{-\infty}^{x} f(y)dy,$$

which is what we wanted to proof.

The other direction is simple: Let us assume that μ is a probability on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, with density f. Then by assumption,

$$F(x) = \mu((-\infty, x]) = \int_{-\infty}^{x} f(y)dy.$$

In particular, again using the dominated convergence theorem, for any increasing sequence $(x_n)_n$, $x_n \to +\infty$,

$$\int_{-\infty}^{\infty} f(y)dy = \int_{-\infty}^{\infty} \lim_{n\to\infty} f(y)\mathbb{I}_{(-\infty,x_n]}dy = \lim_{n\to\infty} F(x_n) = 1,$$

which is (19). So in fact the cdf can be expressed in terms of f and as we have seen in Theorem 3.2 that the cdf completely characterizes μ. That means we are in principle able to calculate $\mu(A)$ for any $A \in \mathcal{B}(\mathbb{R})$ from F alone. The pdf f now gives us a principled way of doing this, with the integral in (18). ■

Remark 5. We have shown that the cdf $F : \mathbb{R} \to \mathbb{R}$,

$$F(x) := \int_{-\infty}^{x} f(t)dt,$$

[10] Here we used the dominating function $g(y) = f(y)$, which has $g(y) \geq f(y)\mathbb{I}_{(-\infty,x_n]}(y)$ for all $y \in \mathbb{R}$ and $n \in \mathbb{N}$ and

$$\int_{-\infty}^{+\infty} g(y)dy = \int_{-\infty}^{+\infty} f(y)dy = 1 < +\infty.$$

is right-continuous and left-continuous. Thus it is also *continuous*, which is another nice feature of having a density. Note that this is a property solely derived from the nonnegativity of f and the properties of the Lebesgue integral, and holds even if f is not continuous. However it does not necessarily hold for measures that don't have a density. $\lozenge$

Remark 6. Note that in the above proof it was absolutely unnecessary to go through the cdf like we did. The reason is that we had already worked out in remark 4, that for all $A \in \mathcal{B}(\mathbb{R})$,

$$\mu(A) = \int_A f(x)dx.$$

This could be used to proof Theorem 3.3 more elegantly. This will in fact be done in Section 3.4, where we proof Theorem 3.3 on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (this is Theorem 3.9). Nonetheless we followed Jacod and Protter (2004) and used a longer proof using the cdf, to exemplify once again how to make good use of Theorem 3.2. $\lozenge$

Remark 7. If f is the pdf of a probability measure μ then it is almost uniquely implied by μ. That is it is characterizing μ (in the sense that for two densities $f_1 = f_2$ means $\mu_1 = \mu_2$ and that for any $A \in \mathcal{B}(\mathbb{R})$ $\mu(A)$ can be calculated as in (18)), yet f is itself only "almost" unique: It is actually possible that $f \neq g$ only a set $\Omega_0 \in \mathcal{A}$ with $\lambda(\Omega_0) = 0$. We then say that $f = g$ λ-almost everywhere in the sense that the set on which f is not equal g has a Lebesgue measure zero. For such two functions one can (quite easily) show that:

$$\int_A f(x)dx = \int_A g(x)dx.$$

So clearly f and g define the same probability measure, even though they are not completely the same. Consider the two densities:

$$f(x) = \mathbb{I}_{(0,1)}(x)$$
$$g(x) = \mathbb{I}_{[0,1]}(x)$$

Then $f = g$ on all of $\mathbb{R}$ except, at the point $\{0, 1\}$, since $f(0) = f(1) = 0$ and $g(0) = g(1) = 1$. However single points always have a measure of zero with respect to the Lebesgue measure (Remember=$\lambda([a, b]) = a - b$ for all $b \geq a \in \mathbb{R}$, and a single point $\{a\} = [a, a]$). So in fact f and g determine the same distribution, the uniform distribution! $\lozenge$

This means whenever we have found a function f fulfilling the above conditions, we have discovered and characterized a new distribution/probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$!

Example 14 (Toy Example from Statistics). The following example is not of great relevance (to my knowledge), but should serve as a toy example for us, before studying the very important cases: Let for some $c \in \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ be:

$$f(x) = c(1 - x)\mathbb{I}_{[0,1]}(x) = \begin{cases} c(1 - x), & \text{if } 0 \leq x \leq 1 \\ 0, & \text{else} \end{cases}$$

Proof that this is the density of a distribution for a certain c and use Remark 2 to simulate from this distribution with your favorite software.

We can choose c freely, so we notice first that $c \geq 0$ is necessary for $f(x) \geq 0$ for all x. Then we may integrate

$$\int_{-\infty}^{+\infty} f(x)dx = \int_0^1 c(1-x)dx = c\int_0^1 (1-x)dx = c\left[x - x^2/2\right]_0^1 = c - \frac{c}{2}$$

This is equal to one iff $c = 2 > 0$. By Theorem f is then a density of a distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For $x \in (0,1)$ cdf is given as

$$F(x) = \int_{-\infty}^{x} 2(1-t)\mathbb{I}_{[0,1]}(t)dt = 2x - x^2,$$

while for $x \leq 0$, $F(x) = 0$ and for $x \geq 1$, $F(x) = 1$. This simple cdf is already not invertible for x outside $[0,1]$, so we need to use the generalized inverse defined in (14). However remember that we only need to look at $y \in (0,1)$ (by Remark 2 we simulate from a uniform distribution to get to the probability space $\Omega = (0,1)$) and for such $y \in (0,1)$ this inverse simplifies to the regular one (since $F(x) \in (0,1)$ only for $x \in (0,1)$, and on this interval F is continuous and strictly increasing), so:

$$y = 2x - x^2 \iff -y + 1 = x^2 - 2x + 1 \iff 1 - y = (x-1)^2 \iff \sqrt{1-y} = |x-1|,$$

and since $x - 1 < 0$ for all $x \in (0,1)$,

$$\sqrt{1-y} = |x-1| \iff -\sqrt{1-y} = x - 1 \iff 1 - \sqrt{1-y} = x.$$

So for any $y \in (0,1)$: $\sup\{x \in \mathbb{R} : F(x) < y\} = 1 - \sqrt{1-y}$. This can then be used to simulate from the distribution by simulating $U \sim \text{Unif}(0,1)$ and calculating $1 - \sqrt{1-U}$, see Matlab example. ♢

Before we look at the different distributions, we need to introduce some more important notions. It should be clear from the above that evaluating integrals is an important part of working with continuous distributions. Unfortunately calculating integrals is an art and most integrals are actually not solvable analytically. To help with this task certain types of integrals are just "defined" to be functions. That means, sometimes the goal is simply to change a given integral until it resembles this specific form. Moreover, the specific structure of these integrals usually make a numerical approximation very easy. In that sense we may treat an integration problem as solved, if we can write it in terms of such a predefined function, even though the integration is technically not removed.[11] An important example of such a function defined through an integral is the gamma function:

$$\Gamma(a) := \int_0^\infty x^{a-1}e^{-x}\, dx, \quad a \in \mathbb{R}_{>0}. \tag{21}$$

[11] This should only motivate the main idea of these functions. Oftentimes they arise much more generally, for example as solutions to differential equations, and just happen to have an integral representation.

The gamma function is a smooth function (it is continuous as are all its derivatives) of one parameter, say a, on $\mathbb{R}_{>0}$. There exists no closed form expression for $\Gamma(a)$ in general, so that it must be computed using numerical methods. However,

$$\Gamma(a) = (a-1)\,\Gamma(a-1), \quad a \in \mathbb{R}_{>1}, \tag{22}$$

and, in particular,

$$\Gamma(n) = (n-1)!, \quad n \in \mathbb{N}. \tag{23}$$

Thus the gamma function generalizes the factorial, which is only defined for natural numbers, to the positive real line. The relation in (23) immediately follows if we use (22) for $n \in \mathbb{N}$, namely:

$$\Gamma(n) = (n-1)\Gamma(n-1) = \ldots = (n-1)(n-2)(n-3)\cdots 1,$$

as $\Gamma(0) = 1$. To prove (22), apply integration by parts with $u = x^{a-1}$ and $dv = e^{-x}\,dx$. This gives $du = (a-1)\,x^{a-2}\,dx$, $v = -e^{-x}$ and

$$\begin{aligned}\Gamma(a) &= \int_0^\infty x^{a-1}e^{-x}\,dx = uv|_{x=0}^{\infty} - \int_0^\infty v\,du = -e^{-x}x^{a-1}\big|_{x=0}^{\infty} \\ &\quad + \int_0^\infty e^{-x}\,(a-1)\,x^{a-2}\,dx \\ &= 0 + (a-1)\,\Gamma(a-1).\end{aligned}$$

It can also be shown that $\Gamma(1/2) = \sqrt{\pi}$. The *incomplete gamma function* is defined as

$$\Gamma_x(a) = \int_0^x t^{a-1}e^{-t}\,dt, \quad a, x \in \mathbb{R}_{>0} \tag{24}$$

and also denoted by $\gamma(x, a)$. The *incomplete gamma ratio* is the standardized version, given by

$$\bar{\Gamma}_x(a) = \Gamma_x(a)\,/\Gamma(a). \tag{25}$$

In general, both functions $\Gamma(a)$ and $\Gamma_x(a)$ need to be evaluated using numerical methods. The *beta function* is an integral expression of two parameters, denoted $B(\cdot,\cdot)$ and defined to be

$$B(a,b) := \int_0^1 x^{a-1}\,(1-x)^{b-1}\;dx, \quad a, b \in \mathbb{R}_{>0}.$$

Closed–form expressions do not exist for general a and b; however, the identity

$$B(a,b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}$$

can be used for its evaluation in terms of the gamma function.

Example 15. To express $\int_0^1 \sqrt{1-x^4}\,dx$ in terms of the beta function, let $u = x^4$ and $dx = (1/4)u^{1/4-1}\,du$, so that

$$\int_0^1 \sqrt{1-x^4}dx = \frac{1}{4}\int_0^1 u^{-3/4}\,(1-u)^{1/2}\;du = \frac{1}{4}B\left(\frac{1}{4}, \frac{3}{2}\right).$$

$\Diamond$

Similar to the incomplete gamma function, the *incomplete beta function* is

$$B_x(p,q) = \mathbb{I}_{[0,1]}(x) \int_0^x t^{p-1}(1-t)^{q-1} \ dt. \tag{26}$$

Let us again define expectation, still not in full generality, but in a way that we are able to use it for the upcoming calculations. Define for any random variable X,

$$X^+ = \max(X, 0)$$
$$X^- = -\min(X, 0)$$

the positive and negative parts of X. In other words if say $X(1) = 5$ and $X(2) = -10$, then $X^+(1) = 5, X^+(2) = 0$ and $X^-(1) = 0, X^-(2) = 10$. It then holds that for all ω:

$$X(\omega) = X^+(\omega) - X^-(\omega)$$
$$|X(\omega)| = X^+(\omega) + X^-(\omega)$$

which may be checked by a case by case analysis. Moreover it is not hard to show that both X^+ and X^- will be $\mathcal{A}/\mathcal{B}(\mathbb{R})$-measurable if X is $\mathcal{A}/\mathcal{B}(\mathbb{R})$-measurable to begin with. Now a very nice property of the Lebesgue integral is that for any measurable function $g : \mathbb{R} \to \mathbb{R}$:

$$\int g \ d\lambda = \int g(x)dx$$

is well-defined whenever $g \geq 0$. It could be $+\infty$, but similar to the case of limits of sequences this does not really bother us usually. The key is that it could in principle be precisely determined. As an aside, this does not hold true at all for the Riemann integral. However if it turns out that g is also Riemann integrable and

$$\int |g(x)|dx = \int g(x)dx < +\infty,$$

then Riemann and Lebesgue integral again can be used interchangeable.

Now in full generality, we could define

$$\mathbb{E}[X^+] = \int_\Omega X^+(\omega)dP(\omega), \quad \mathbb{E}[X^-] = \int_\Omega X^-(\omega)dP(\omega)$$

which again is valid, because X^+ and X^- are measurable and nonnegative. Then if $\mathbb{E}[X^+] < +\infty$ *or* $\mathbb{E}[X^-] < +\infty$, we would define

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

This definition is possible for *any* random variable, discrete, continuous or none of the two. However, since we did not yet talk about integrals with respect to general (probability) measures P, we will instead do this in less general terms, namely, for f being the density of X:

$$\mathbb{E}[X^+] = \int_{\mathbb{R}} \max(x, 0)f(x)dx$$
$$\mathbb{E}[X^-] = \int_{\mathbb{R}} -\min(x, 0)f(x)dx.$$

Those are well defined since f is $\mathcal{B}(\mathbb{R})/\mathcal{B}(\mathbb{R})$-measurable by assumption and, again, since $\max(x,0)f(x) \geq 0$ and $-\min(x,0)f(x) \geq 0$ for all $x \in \mathbb{R}$. Now if *both* $\mathbb{E}[X^+] < +\infty$ and $\mathbb{E}[X^-] < +\infty$, then we define the expecation of X to be

$$\mathbb{E}[X] := \mathbb{E}[X^+] - \mathbb{E}[X^-] = \int_{\mathbb{R}} xf(x)dx, \tag{27}$$

otherwise we say $\mathbb{E}[X]$ does not exist. This approach of defining expectation only works for continuous random variables (i.e. r.v. with a density f). One can also show that the general approach discussed above "simplifies" to (27) for continuous random variables, so that the two approaches are consistent.

Similarly we define for any measurable function $g : \mathbb{R} \to \mathbb{R}$:

$$\mathbb{E}[g(X)] := \mathbb{E}[g(X)^+] - \mathbb{E}[g(X)^-] = \int_{\mathbb{R}} g(x)f(x)dx, \tag{28}$$

whenever $\mathbb{E}[g(X)^+] < +\infty$ and $\mathbb{E}[g(X)^-] < +\infty$. In the case where we know that $\mathbb{E}[g(X)]$ exists, we will just directly calculate it as the integral in (28).

Finally, we will often deal with symmetric distributions:

Definition 3.5. A random variable X is symmetric around zero if $X \overset{D}{=} -X$, that is X and $-X$ have the same distribution, or $\mu_X = \mu_{-X}$.

For example, consider $X \sim N(0,1)$. As we will mention again below, the Gaussian distribution is completely characterized by its mean and variance (in this case 0 and 1). Now $-X$ has $\mathbb{E}[-X] = 0$ and $\mathbb{V}(-X) = \mathbb{V}(X) = 1$ and it actually holds that $-X$ is Gaussian as well (this can for example be demonstrated using Remark 9 below). So indeed $X \overset{D}{=} -X$ and X is symmetric. With this example in mind, one might surmise that the symmetry of a distribution or random variable has a lot to do with the symmetry of its density. This is indeed true, as shown in Theorem 3.4. Before we attend it and its proof, we need some more remarks however

Remark 8. Two important remarks regarding the consequences of continuity (for cdf and density):

1. So far we only tiptoed around the fact that if the density f is itself continuous, then the cdf is differentiable and $F' = f$, as mentioned in Remark 4. Now we will make this slightly more precise (though without proof): At every $x \in \mathbb{R}$ at which $f(x)$ is continuous, F is *continuously* differentiable and $F'(x) = f(x)$. Thus if f is continuous for all $x \in A$, for some set $A \in \mathcal{B}(\mathbb{R})$, then F is continuously differentiable on A and $F'(x) = f(x)$ for all $x \in A$. This is not just nice for finding a pdf from a cdf, it also immediately has the nice benefit, that if $g : \mathbb{R} \to \mathbb{R}$ is another *continuous* density for μ_X, i.e. $f = g$ λ-almost everywhere, then in fact $F'(x) = f(x) = g(x)$ for all $x \in A$.

   In particular, if $A = \mathbb{R}$, then F is continuously differentiable everywhere and $F'(x) := f(x)$ is the unique continuous density of μ_X. That is any other continuous density g, as $g = f$.[12]

[12] One could still construct a different density g that is only continuous almost everywhere and for which it holds that $f = g$ only λ-almost everywhere. But then g is not continuous at all points for which $g(x) \neq f(x)$.

2. Below we want to find the cdf of a simple function of the r.v. X, namely of $-X$. To distinguish them, let's call the cdf of X, F_X and the cdf of $-X$ simply F_{-X}. Then by definition, for any random variable X, the cdf of $-X$ can be expressed as:

$$F_{-X}(x) = P(-X \leq x) = P(X \geq -x) = 1 - P(X < -x).$$

Now notice that $P(X < -x)$ is the left limit $\lim_{y \uparrow -x} F_X(y)$. Indeed, take any arbitrary monotone sequence $(y_n)_n$ with $y_n \uparrow -x$. Then $F_X(y_n)$ is itself a monotone sequence, since $y_n \leq y_{n+1}$ implies $F_X(y_n) \leq F_X(y_{n+1})$ and by the monotonicity of μ_X:

$$\lim_n F_X(y_n) = \lim_n \mu_X((-\infty, y_n]) = \mu_X((-\infty, x)),$$

since $\lim_n (-\infty, y_n] = \bigcup_n (-\infty, y_n] = (-\infty, x)$. Since the sequence $(y_n)_{n \in \mathbb{N}}$ was arbitrary, we have that

$$\lim_{y \uparrow -x} F_X(y) = \mu_X((-\infty, -x)) = P(X < -x).$$

So far this was all true, for any random variable $X : \Omega \to \mathbb{R}$. But now, if X admits a density f_X, we have shown above that F_X is in fact continuous (whether or not f_X is continuous) we have that the left limit not only exists, but $\lim_{y \uparrow -x} F_X(y) = F_X(-x)$, or:

$$F_{-X}(x) = 1 - P(X < -x) = 1 - \lim_{y \uparrow -x} F_X(y) = 1 - F_X(-x). \tag{29}$$

In general beware the simply lessons: The cdf F is always increasing, right-continuous and $\lim_{x \to \infty} F(x) = 1$, $\lim_{x \to -\infty} F(x)$, and:

$$\begin{aligned} \text{density exists (continuous or not)} &\implies F \text{ continuous} \\ \text{density continuous} &\implies F \text{ differentiable} \end{aligned}$$

$\Diamond$

Theorem 3.4. *If X possesses a density $f_X : \mathbb{R} \to \mathbb{R}$ which is itself continuous, then X is symmetric around zero iff $f_X(-x) = f_X(x)$ for all $x \in \mathbb{R}$, i.e. the density is symmetric. Furthermore for any k odd for which $\mathbb{E}[X^k]$ exists, we have $\mathbb{E}[X^k] = 0$.*

Proof. First, we want to show that X being symmetric around zero, implies that the density f_X is symmetric. So by assumption $\mu_X = \mu_{-X}$, and in particular we immediately know that μ_{-X} also has a continuous density, namely f_X, since then for all x:

$$\mu_{-X}((-\infty, x]) = \mu_X((-\infty, x]) = \int_{-\infty}^{x} f_X(t)dt.$$

By Remark 8, this immediately implies that F'_{-X} is a valid and even continuous density for $-X$ and,

$$F'_{-X}(x) = f_X(x) \text{ for all } x \in \mathbb{R}.$$

As also shown in Remark 8, F_{-X} can be express through (29). But this means we may just differentiate (29) to obtain the desired result, for each $x \in \mathbb{R}$:

$$f_X(x) = F'_{-X}(x) = (1 - F_X(-x))' = f_X(-x).$$

Now assume $f_X(x) = f_X(-x)$ for all $x \in \mathbb{R}$. Then with $t = -s$ and $dt = -ds$,

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(t)dt \\ &= -\int_{\infty}^{-x} f_X(-s)ds \\ &= \int_{-x}^{\infty} f_X(s)ds \\ &= 1 - F_X(-x) \\ &= F_{-X}(x), \end{aligned}$$

using (29) in the last step. So in fact $F_X = F_{-X}$ and thus we immediately know that $\mu_X = \mu_{-X}$.

Finally if $\mathbb{E}[X^k]$ exists for any $k \in \mathbb{N}$ odd, then we can use the substitution $y = -x$, st. $dy = -dx$ and

$$\begin{aligned} \mathbb{E}[X^k] &= \int_{\mathbb{R}} x^k f_X(x)dx \\ &= \int_{-\infty}^{0} x^k f_X(x)dx + \int_{0}^{+\infty} x^k f_X(x)dx \\ &= -\int_{+\infty}^{0} (-y)^k f_X(-y)dy + \int_{0}^{+\infty} x^k f_X(x)dx \\ &\overset{(*)}{=} -\int_{0}^{\infty} y^k f_X(y)dy + \int_{0}^{+\infty} x^k f_X(x)dx \\ &= 0. \end{aligned}$$

The crucial facts in $(*)$ are that $(-y)^k = -y^k$, which is only true for k odd, and that $f(x) = f(-x)$. ■

Remark 9. A similar idea as in the above simple proof also allows to find the pdf of a function of the continuous random variable X. Say $g : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable and invertible function with nonzero derivative, i.e. $g'(x) > 0$ or $g'(x) < 0$ for all $x \in \mathbb{R}$. First assume that $g'(x) > 0$, for all x. In particular, this implies that g is strictly increasing. We wish to find the distribution μ_Y of $Y = g(X)$. Then if we assume the density of X, f_X, to be continuous, we have

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

Taking derivatives:

$$\frac{dF_Y(y)}{dy} = f_X(g^{-1}(y))\frac{dg^{-1}(y)}{dy}.$$

If $g'(x) < 0$ on the other hand,

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)).$$

with derivative

$$\frac{dF_Y(y)}{dy} = -f_X(g^{-1}(y))\frac{dg^{-1}(y)}{dy}.$$

Combining the two cases we can say that if g is continuously differentiable and strictly monotone on $\mathbb{R}$ ($g'(x) > 0$ or $g'(x) < 0$ for all $x \in \mathbb{R}$):

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|. \tag{30}$$

An important example of this is the following: If X has some distribution with continuous pdf f_X, then for $a \in \mathbb{R}$ and $\sigma > 0$, $Y = a + \sigma X$ has pdf:

$$f_Y(y) = f_X(g^{-1}(y))\frac{dg^{-1}(y)}{dy} = \frac{1}{\sigma}f_X\left(\frac{y-a}{\sigma}\right), \tag{31}$$

since $g^{-1}(y) = (y-a)/\sigma$ and $g'(x) = \sigma > 0$ for all x.

$\Diamond$

Finally, assume we have a random variable X with a continuous density which is symmetric around zero, and we want to find out whether $\mathbb{E}[X^k]$ exists. We need to check that

$$\mathbb{E}[(X^k)^+] = \int_{-\infty}^{\infty} \max(x^k, 0)f(x)dx < \infty$$
$$\mathbb{E}[(X^k)^-] = \int_{-\infty}^{\infty} -\min(x^k, 0)f(x)dx < \infty.$$

But in this case for $k \in \mathbb{N}$ *odd*, since $x^k \geq 0$ iff $x \geq 0$:

$$\mathbb{E}[(X^k)^+] = \int_0^{\infty} x^k f(x)dx,$$

and with $y = -x$, $dy = -dx$ and $-\min(x^k, 0) = \max(-(x^k), 0) = \max((-x)^k, 0)$:

$$\mathbb{E}[(X^k)^-] = \int_{-\infty}^{\infty} -\min(x^k, 0)f(x)dx = \int_{-\infty}^{\infty} \max(y^k, 0)f(y)dy = \int_0^{\infty} x^k f(x)dx,$$

due to $f(x) = f(-x)$. So for k odd, it is enough to check $\int_0^{\infty} x^k f(x)dx < \infty$. Now consider $k \in \mathbb{N}$ *even*. Then $x^k \geq 0$ for all $x \in \mathbb{R}$. Thus:

$$\mathbb{E}[(X^k)^+] = \int_{-\infty}^{\infty} \max(x^k, 0)f(x)dx = \int_{-\infty}^{\infty} x^k f(x)dx$$
$$\mathbb{E}[(X^k)^-] = \int_{-\infty}^{\infty} -\min(x^k, 0)f(x)dx = 0.$$

So for k even, we need to check $\int_{-\infty}^{\infty} x^k f(x)dx < \infty$. Adding again the symmetry of $f(x)$, the expectation becomes

$$\mathbb{E}[(X^k)^+] = \int_{-\infty}^{\infty} x^k f(x)dx = \int_0^{\infty} x^k f(x)dx + \int_{-\infty}^{0} x^k f(x)dx = 2\int_0^{\infty} x^k f(x)dx,$$

since with $y = -x$ and $f(x) = f(-x)$,

$$\int_{-\infty}^{0} x^k f(x)dx = \int_{-\infty}^{0} (-x)^k f(-x)dx = \int_0^{+\infty} y^k f(y)dx.$$

In other words whether k is odd or not, we only need to check

$$\int_0^{\infty} x^k f(x)dx < +\infty \tag{32}$$

in case of X having a symmetric continuous pdf. If it is finite, $\mathbb{E}[X^k]$ exists, if it is infinite, $\mathbb{E}[X^k]$ does not exist. This will be used mainly to demonstrate Theorem 3.5 in case of the t-distribution below.

For the upcoming two subsections we take a lot from Paolella (2006, Chapter 7):

3.2.1 Uniform Distribution

For $b > a \in \mathbb{R}$, we denote $X \sim U(a, b)$, or $\mu_X = U(a, b)$, if μ_X has the density $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = \frac{1}{b-a}\mathbb{I}_{(a,b)}(x), \tag{33}$$

with $\mathbb{I}_A(x) = 1$, if $x \in A$ and 0 else. Note that by the example in Remark 7, it does not matter whether we include a, b or not. If $a = 0$ and $b = 1$, then in fact $\mu_X = \lambda_{|(0,1)}$, the Lebesgue measure restricted to $(0, 1)$. If $a \neq 0$ or $b \neq 1$, we need to renormalize the density however, so that

$$\int_{-\infty}^{\infty} f(x)\ dx = \int_{-\infty}^{\infty} \frac{1}{b-a}\mathbb{I}_{(a,b)}(x)\ dx = \frac{1}{b-a}\int_a^b\ dx = \frac{b-a}{b-a} = 1.$$

For $x \in (a, b)$, cdf is then given as

$$F(x) = \int_{-\infty}^{x} \frac{1}{b-a}\mathbb{I}_{(a,b)}(t)\ dt = \frac{x-a}{b-a},$$

while $F(x) = 0$, for $x \leq a$ and $F(x) = 1$, for $x \geq b$. In other words it can be expressed as:

$$F(x) = \frac{x-a}{b-a}\mathbb{I}_{(a,b)}(x) + \mathbb{I}_{[b,+\infty)}(x). \tag{34}$$

If $(a, b) = (0, 1)$, then $\mu_X((0, x)) = F(x) = x$, as we have mentioned before in Theorem 3.2. So indeed $\mu_X = \lambda_{|(0,1)}$ on $(0, 1)$. For any $k \in \mathbb{N}$, the kth moment is

$$\mathbb{E}[X^k] = \frac{1}{b-a}\int_a^b x^k dx = \frac{1}{b-a}\frac{b^{k+1} - a^{k+1}}{k+1}.$$

Despite its simplicity, the distribution has important applications. For example:

Example 16. Recall again the discussion in Section 3.1, showing that for any random variable X, $F(X)$ is again a random variable. We have also seen in theorem 3.2 that if $Y \sim \text{Unif}(0,1)$, and F is the cdf of some probability μ, then $F^{-1}(Y)$, with F^{-1} defined as in (14), has distribution μ. If F is continuous and strictly increasing ($x < y \implies F(x) < F(y)$), this translates into the fact that

$$F(X) \sim \text{Unif}(0,1).$$

Indeed this is easy to check: Let $Y = F(X)$, so that $Y : \Omega \to \mathbb{R}$ is a random variable with cdf F_Y, then for $y \in (0,1)$:

$$F_Y(y) = P(\{\omega : Y(\omega) \leq y\}) = P(\{\omega : F(X(\omega)) \leq y\}) = P(\{\omega : X(\omega) \leq F^{-1}(y)\}) = F(F^{-1}(y)) = y,$$

while $F_Y(y) = 0$ for $y \leq 0$ and $F_Y(y) = 1$ for $y \geq 1$. This is just the cdf of a uniform distribution on $(0,1)$. Since the cdf uniquely characterizes a distribution, this means $F(X) \sim \text{Unif}(0,1)$. $\Diamond$

A final note about this distribution; So far we did not explicitly talk about this, but with the density in (33) we obtain a distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and not on $((a,b), \mathcal{B}((a,b)))$. This is why we need to consider values $x \leq a$ and $x \geq b$ when looking at the cdf. In that sense $\mu_X = \lambda_{|(0,1)}$ is not entirely correct, we should say $\mu_{X|(0,1)} = \lambda_{|(0,1)}$, i.e. if both μ_X and λ are restricted to $(0,1)$, then they are the same. To solve this apparent paradox notice that $f(x) > 0$ only iff $x \in (a,b)$. That is, we only have positive probabilities for subsets of (a,b) or on $\mathcal{B}((a,b))$. The closure of this set, $[a,b]$, is what we call the support of X. To define this more generally, let for any set A in a metric or topological space, $\bar{A}$ be its closure, i.e. the smallest closed set including A. Similarly the interior of A is the largest open set in A and denoted A^{o}. For instance if $A = [a,b) \subset \mathbb{R}$, then $\overline{A} = [a,b]$ and $A^{\text{o}} = (a,b)$. Define

Definition 3.6. Let $(\Omega, \mathcal{A}, P)$ be a probability space and T a topological space with Borel σ-algebra $\mathcal{B}(T)$. The support of a $\mathcal{A}/\mathcal{B}(T)$ measurable function $X : \Omega \to T$, denoted $\text{supp}(X)$, is defined as the smallest closed set $\overline{C} \subset T$ such that

$$P(X \in \overline{C}) = \mu_X(\overline{C}) = 1.$$

This definition is very general, however note that we didn't talk about a general measurable space $(S, \mathcal{S})$ here and instead used a topological space T. The reason is that we do not want to allow for an arbitrary σ-algebra $\mathcal{S}$, but instead only consider the Borel σ-algebra (which is based on open sets and thus needs a topology). In this context the above definition makes sense, since $\overline{C}$ closed is certainly in $\mathcal{B}(T)$. One can show that for continuous random variables like $X \sim U(a,b)$, the support of X is also the closure of the set on which $f(x) > 0$:[13]

$$\text{supp}(X) = \overline{\{x \in \mathbb{R} : f(x) > 0\}}.$$

[13] At least there exists *a version* of the density such that this is true; Recall that if f is the density of X, then all $g : \mathbb{R} \to \mathbb{R}$, such that $g = f$ almost everywhere are also densities for X! So we have a choice in densities and it is simple to show that we can choose a density f which is zero outside of $\text{supp}(X)$ and for this f the equality holds true.

As an important remark, whenever we assumed in Section 3.2 that the density f is continuous on $\mathbb{R}$, it is actually enough that it is continuous on the *interior of the support of* X. In other words on the open set $\text{supp}(X)^{\text{o}} = \{x \in \mathbb{R} : f(x) > 0\}$.[14] More concretely, say we know the density f of X to be continuous on $\text{supp}(X)^{\text{o}}$, but we don't know what it is. We can then obtain it as:

$$f(x) = \begin{cases} F'(x), & x \in \text{supp}(X)^{\text{o}} \\ 0, & x \notin \text{supp}(X)^{\text{o}}, \end{cases}$$

since $F'(x) = f(x)$ holds for all x for which $f(x)$ is continuous. Clearly for any integration we undertake, only the set where $f(x) > 0$ matters. In the same spirit, for Theorem 3.4 and Remark 9 it is enough that the density f is continuous on $\text{supp}(X)^{\text{o}}$.

3.2.2 Gaussian Distribution

We denote $X \sim N(0,1)$, or $\mu_X = N(0,1)$, if μ_X has the density $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{35}$$

There is no closed form solution for the cdf available, yet there is no shortage of implemented numerical approximations to it. It plays itself an important role, for instance in modelling probabilities as in a probit regression. One can proof that for any $k \in \mathbb{N}$, $\mathbb{E}[X^k]$ exists.[15] Since also the density is obviously symmetric, we have that X is symmetric around zero. This means for any k odd, it holds by Theorem 3.4 that $\mathbb{E}[X^k] = 0$. If k is even on the other hand, let $r = k/2 \in \mathbb{N}$ (since k is even) so that,

$$\begin{aligned} \mathbb{E}[X^k] = \mathbb{E}[X^{2r}] &= \int_{-\infty}^{\infty} x^{2r} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\, dx \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x^{2r} \exp\left(-\frac{x^2}{2}\right)\, dx \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} (2u)^r \exp(-u) \frac{1}{\sqrt{2u}}\, du \\ &= \frac{2^{1+r-1/2}}{\sqrt{2\pi}} \int_0^{\infty} u^{r-1/2} \exp(-u)\, du \\ &= \frac{2^r \Gamma(r+1/2)}{\sqrt{\pi}}, \end{aligned}$$

with $u = x^2/2$ st. $x = (2u)^{1/2}$ and $du = xdx$ or $dx = 1/\sqrt{2u}\, du$. One could actually simplify this further, as done for instance in Paolella (2006, p. 257), but we will content ourselves with the above expression. Consider as an example $k = 2$:

$$\mathbb{E}[X^2] = \frac{2^1 \Gamma(1+1/2)}{\sqrt{\pi}} = \frac{21/2\Gamma(1/2)}{\sqrt{\pi}} = 1,$$

[14] As a not exam relevant comment; The set $M = \{x \in \mathbb{R} : f(x) > 0\}$ is open for f continuous, because any $x \in M$ has some small open interval around it with $f(y) > 0$ for all y in this interval (by continuity of f at x). This corresponds to the definition of M being open.

[15] In fact the distribution has so called "light" tails, which in particular implies that all positive moments exist.

since $\Gamma(a) = (a-1)\Gamma(a-1)$, for $a > 1$ and $\Gamma(1/2) = \sqrt{\pi}$.

For $a \in \mathbb{R}$ and $\sigma > 0$, from Remark 9, one can surmise that then $Y = a + \sigma X$ has pdf

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-a)^2}{\sigma^2}\right), \tag{36}$$

and we say $Y \sim N(a, \sigma^2)$, or $\mu_Y = N(a, \sigma^2)$. For any $k \in \mathbb{N}$ the moments are given as,

$$\mathbb{E}[Y^k] = \mathbb{E}[(a + \sigma X)^k] = \sum_{l=0}^{k} \binom{k}{l} a^{k-l}\sigma^l \mathbb{E}[X^l].$$

For instance,

$$\mathbb{E}[Y] = a^1 + \sigma\mathbb{E}[X^1] = a,$$

and

$$\mathbb{V}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[(\sigma X)^2] = \sigma^2\mathbb{E}[X^2] = \sigma^2.$$

Many things in statistics are defined in relation to the Gaussian distribution. For instance, the kurtosis of a given random variable Z, $\mathbb{E}[(Z - \mathbb{E}[Z])^4]/(\mathbb{E}[(Z - \mathbb{E}[Z])^2])^2$ (a measure of tail thickness) if often defined as the difference from the kurtosis of a Gaussian random variable $Y \sim N(a, \sigma^2)$, which is

$$\frac{\mathbb{E}[(Y - \mathbb{E}[Y])^4]}{\mathbb{E}[(Y - \mathbb{E}[Y])^2]^2} = \mathbb{E}\left[\left(\frac{Y-a}{\sigma}\right)^4\right] = \mathbb{E}[X^4] = \frac{2^2\Gamma(5/2)}{\sqrt{\pi}} = \frac{2^2 3/2\Gamma(3/2)}{\sqrt{\pi}} = \frac{2^2 3/4\Gamma(1/2)}{\sqrt{\pi}} = 3,$$

with $X = (Y - a)/\sigma \sim N(0, 1)$ and since $\Gamma(1/2) = \sqrt{\pi}$. The (excess) kurtosis of Z is then

$$\frac{\mathbb{E}[(Z - \mathbb{E}[Z])^4]}{\mathbb{E}[(Z - \mathbb{E}[Z])^2]^2} - 3.$$

The Gaussian distribution arises in a plethora of applications. Informally speaking, this seems to be related to the central limit theorem; the sum of many independent random variables has a distribution which is approximately Gaussian. One might surmise that things like the height of a person are the result of a combination of many independent factors, genetic as well as environmental influences. However there are situations, in finance for instance, where the assumption of a Gaussian distribution is not appropriate. In particular it is confined to symmetry, which does not make sense for returns data for instance. Additionally the tails of the normal distribution are not heavy, meaning that very extreme events (such as great losses) are deemed very unlikely.

We will wait with an example until we reach the multivariate case.

3.2.3 Student's t distribution

For $\nu \in \mathbb{R}_{>0}$, we denote $X \sim t_\nu$, or $\mu_X = t_\nu$, if μ_X has the density $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)\nu^{\nu/2}}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(\nu + x^2\right)^{-(\nu+1)/2} = K_\nu\left(\nu + x^2\right)^{-(\nu+1)/2} \tag{37}$$

We again have $f(x) = f(-x)$, so X is symmetric around zero. There are several ways to arrive at the above density. We will quickly mention the two most important once, without derivations:

(a) Assume we have a collection of n iid Gaussian random variables $(X_1, \ldots, X_n)$ with mean a and variance σ^2 all defined on the probability space $(\Omega, \mathcal{A}, P)$. This means they are independent (a notion we did not define yet) and identically distributed, i.e. $\mu_{X_1} = \mu_{X_2} = \ldots = \mu_{X_n}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Define the mean $\bar{X} = 1/n \sum_{i=1}^n X_i \sim N(a, \sigma^2/n)$ and estimated variance $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$. Then both $\bar{X}, S^2$ are again $\mathcal{B}(\mathbb{R})/\mathcal{B}(\mathbb{R})$-measurable functions from Ω to $\mathbb{R}$ (i.e. random variables) and if σ were known

$$\frac{(\bar{X} - a)}{\sqrt{\sigma^2/n}} \sim N(0, 1).$$

However, it turns out that if we replace the fixed σ^2 with the random variable S^2 instead,

$$\frac{(\bar{X} - a)}{\sqrt{S^2/n}} \sim t_{n-1}. \tag{38}$$

Imagine you have a "zero" hypothesis about the mean of your data, say you are measuring some effect with a Gaussian distribution and you want to test: $H_0 : a = a_0$ vs. $H_1 : a \neq a_0$. Oftentimes the sample at hand for such tests will have a distribution close to a Gaussian (or at least that was assumed in classical statistics for some time). Then if H_0 *were true*

$$T := \frac{(\bar{X} - a_0)}{\sqrt{S^2/n}} \sim t_{n-1},$$

since we would in this case subtract the correct mean. Thus we know the distribution (under H_0) of our so-called test statistic $T : \Omega \to \mathbb{R}$, which is again a measurable function. If we now find out that for a given realization $t = T(\omega)$, $P(T > |t|)$ is "very small", we reject the null hypothesis. This is the basis of the relatively simple hypothesis tests from classical statistics.[16]

(b) Like many interesting distribution, $\mu_X = t_\nu$ can be attained by a so called continuous mean-variance mixture. Namely

$$f(x) = \int_0^\infty \phi(x; 0, 1/g) f_G(g) dg$$

where $\phi(x; 0, 1/g)$ is the density of a Gaussian distribution with mean zero and variance $1/g$ and f_G is the density of a gamma random variable with parameters $\nu/2$ and $2/\nu$. This can be interpreted as follows: Given a realization of the gamma random variable $G(\omega) = g$, $X \sim N(0, 1/g)$, or $X|G = g \sim N(0, 1/g)$. Now G is "latent", that means we do not observe it and therefore want to integrate it out: As we will see in Section 3.4, $(x, g) \mapsto \phi(x; 0, 1/g) f_G(g)$ is actually the joint density of $Z = (X, G)$ and integrating over G gives the marginal density of X. One can then go one to show that this marginal density is the t distribution.

[16] Through the magic of the Central limit theorem and the law of large numbers, this is also approximately valid, if $(X_1, \ldots, X_n)$ are iid with some unknown distribution (under some moment conditions) for "large" n.

An analytical solution to the cdf is available in terms of the so called Gaussian hypergeometric function. It would not be hard to derive this, but to refrain from having to introduce another such function we will not do this here. What is really nice about the t-distribution, especially from the point of view of finance, is that the thickness of tails can be varied (or estimated) with the parameter ν. For $\nu = 1$, we obtain the Cauchy distribution discussed below, a distribution with such heavy tails that not even the expected value exists. If we let $\nu \in (2, 3]$, $\mathbb{E}[X^2]$ exists, but $\mathbb{E}[X^3]$ does not. To proof this formally, let us introduce the following notion: If $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are two functions, we say g is asymptotically equivalent to h for $|x| \to \infty$, denoted $g \sim h$, if

$$\lim_{|x|\to\infty} \frac{h(x)}{g(x)} = 1 \iff \lim_{x\to+\infty} \frac{h(x)}{g(x)} = 1 \text{ and } \lim_{x\to-\infty} \frac{h(x)}{g(x)} = 1.$$

If $g, h : [0, +\infty)$, then this condition simplifies to

$$\lim_{x\to+\infty} \frac{h(x)}{g(x)} = 1.$$

It then holds that:

Theorem 3.5. *Let $X \sim t_\nu$. Then for $k \in \mathbb{N}$, $\mathbb{E}[X^k]$ exists if $k < \nu$ and does not exist for $k \geq \nu$. In the former case $\mathbb{E}[X^k] = 0$ for all k odd and*

$$\mathbb{E}[X^k] = \frac{\nu^{k/2}}{B\left(\frac{\nu}{2}, \frac{1}{2}\right)} B\left(\frac{k+1}{2}, \frac{\nu-k}{2}\right),$$

for all k even.

Proof. We will make an argument by studying the "asymptotic behavior" of the density f and proof that

$$\int_0^\infty x^k f(x)dx \begin{cases} < +\infty, & \text{if } k < \nu \\ = +\infty, & \text{if } k \geq \nu \end{cases}$$

This is exactly condition (32), so it is indeed enough to look at this integral as we have demonstrated above.

Looking at the density in (37) it is clear that $f(x) = f(|x|)$ for all $x \in \mathbb{R}$ (which is just a different way of saying f is symmetric) and $f(x) = f(|x|) \sim K_\nu |x|^{-(\nu+1)}$, since

$$\frac{K_\nu \left(\nu + |x|^2\right)^{-(\nu+1)/2}}{K_\nu |x|^{-(\nu+1)}} = \frac{(|x|^2)^{(\nu+1)/2}}{(\nu + |x|^2)^{(\nu+1)/2}} = \left(\frac{|x|^2}{\nu + |x|^2}\right)^{(\nu+1)/2} = \left(\frac{1}{\nu/|x|^2 + 1}\right)^{(\nu+1)/2} \longrightarrow 1,$$

as $|x| \to +\infty$. So asymptotically speaking $f(x)$ behaves the same as $K_\nu |x|^{-(n+1)}$. But this also means that $|x|^k f(|x|) \sim K_\nu |x|^{k-(n+1)}$. Why could this help? First note that to check condition (32), it is enough to look at the domain $[0, +\infty)$, so we can drop the absolute values and just state that $x^k f(x) \sim K_\nu x^{k-(n+1)}$ for $x \to +\infty$. Then $x^k f(x) \sim K_\nu x^{k-(\nu+1)}$ as $x \to +\infty$ means

by definition of convergence that for any $\varepsilon > 0$ there exists some $x_\varepsilon > 0$, such that for all x with $x > x_\varepsilon$,

$$\begin{aligned}
&\left| \frac{x^k \left(\nu + x^2\right)^{-(\nu+1)/2}}{x^{k-(\nu+1)}} - 1 \right| \leq \varepsilon \\
\iff & -\varepsilon \leq \frac{x^k \left(\nu + x^2\right)^{-(\nu+1)/2}}{x^{k-(\nu+1)}} - 1 \leq \varepsilon \\
\iff & x^{k-(\nu+1)}(1-\varepsilon) \leq x^k \left(\nu + x^2\right)^{-(\nu+1)/2} \leq x^{k-(\nu+1)}(1+\varepsilon) \qquad (39)
\end{aligned}$$

Looking at the integral of interest we have

$$\begin{aligned}
\int_0^\infty x^k f(x)dx &= \int_0^{x_\varepsilon} x^k f(x)dx + \int_{x_\varepsilon}^{+\infty} x^k f(x)dx \\
&= K_\nu \int_0^{x_\varepsilon} x^k \left(\nu + x^2\right)^{-(\nu+1)/2} dx + K_\nu \int_{x_\varepsilon}^{+\infty} x^k \left(\nu + x^2\right)^{-(\nu+1)/2} dx \\
&= K_\nu((I) + (II)).
\end{aligned}$$

Let us first consider (I). For $x \in [0, x_\varepsilon]$, $h(x) := x^k \left(\nu + x^2\right)^{-(\nu+1)/2}$ and $x^{k-(\nu+1)}$ might be very far apart. However this does not bother us, since $h : [0, x_\varepsilon] \to \mathbb{R}$ is a *continuous function on a compact set*. This means it attains its maximum, say $M < \infty$, on that set, or:

$$x^k \left(\nu + x^2\right)^{-(\nu+1)/2} \leq M \text{ for all } x \in [0, x_\varepsilon].$$

But then

$$(I) = \int_0^{x_\varepsilon} x^k \left(\nu + x^2\right)^{-(\nu+1)/2} dx \leq \int_0^{x_\varepsilon} M dx = x_\varepsilon M < \infty.$$

So we can disregard (I) for the purposes of this theorem, as it is certainly finite. Using (39), we can estimate (II) as

$$(1-\varepsilon) \int_{x_\varepsilon}^\infty x^{k-(\nu+1)} dx \leq \int_{x_\varepsilon}^\infty x^k \left(\nu + x^2\right)^{-(\nu+1)/2} dx \leq (1+\varepsilon) \int_{x_\varepsilon}^\infty x^{k-(\nu+1)} dx,$$

by the monotonicity of integrals. So clearly (remember all involved integrals are well-defined),

$$\int_{x_\varepsilon}^{+\infty} x^k f(x)dx < \infty \iff \int_{x_\varepsilon}^\infty x^{k-(\nu+1)} dx < \infty \qquad (40)$$

Now we can check for an arbitrary sequence $(x_n)_n$, with $x_n \uparrow +\infty$ and $x_n \geq x > 0$, that for all n,

$$\int_{x_\varepsilon}^{x_n} x^{k-(\nu+1)} dx = \begin{cases} [\log(x)]_{x_\varepsilon}^{x_n} = \log(x_n) - \log(x_\varepsilon), \text{ if } k = \nu \\ \left[x^{k-\nu}/(k-\nu)\right]_{x_\varepsilon}^{x_n} = (x_n^{k-\nu} - x_\varepsilon^{k-\nu})/(k-\nu), \text{ if } k \neq \nu, \end{cases}$$

which follows because $x^{k-(\nu+1)}$ is Riemann integrable on (x_ε, x_n) with a finite integral value, so the Lebesgue and Riemann integral coincide in this case. If $n \to \infty$ (and thus $x_n \to \infty$) the limit exists in both cases (since $(x_n)_n$ is monotone, so are the sequences $(\log(x_n))_n$ and $(x_n^{k-\nu})_n$), but they will only be finite iff $k - \nu < 0$, or $k < \nu$. In this case $\lim_n x^{k-\nu} = 0$ and thus

$$\int_{x_\varepsilon}^\infty x^{k-(\nu+1)} dx = \lim_n \int_{x_\varepsilon}^{x_n} x^{k-(\nu+1)} dx = \lim_n \frac{x_n^{k-\nu} - x_\varepsilon^{k-\nu}}{k-\nu} = -\frac{x_\varepsilon^{k-\nu}}{k-\nu} < \infty.$$

Then combining everything we have

$$\int_0^\infty x^k f(x)dx = K_\nu((I) + (II)) < \infty$$

iff $k < \nu$.

Since the density f is symmetric, we then immediately know that $\mathbb{E}[X^k] = 0$, for k odd with $k < \nu$. For k even, first note that, as usual by symmetry

$$\mathbb{E}[X^k] = 2\int_0^{+\infty} x^k f(x)dx.$$

Furthermore for simplicity of notation, we write the density in (37) slightly different, in terms of the $B(.,.)$ function:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)\nu^{\nu/2}}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(\nu + x^2\right)^{-(\nu+1)/2} = \frac{\nu^{-1/2}}{B\left(\frac{\nu}{2},\frac{1}{2}\right)}\left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

Starting from those two remarks, we then use the (not so obvious) substitution:

$$y = \frac{x^2}{\nu + x^2}, \quad x = +\sqrt{\nu\frac{y}{1-y}}, \quad dx = \frac{\nu^{1/2}}{2}y^{-1/2}(1-y)^{-3/2}dy,$$

to get:

$$\begin{aligned}
\mathbb{E}[X^k] &= 2\frac{\nu^{-1/2}}{B\left(\frac{\nu}{2},\frac{1}{2}\right)}\int_0^{+\infty} x^k\left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}dx \\
&= 2\frac{\nu^{-1/2}}{B\left(\frac{\nu}{2},\frac{1}{2}\right)}\int_0^{+\infty} x^k\left(\frac{x^2}{x^2}\frac{\nu}{\nu+x^2}\right)^{(\nu+1)/2}dx \\
&= 2\frac{\nu^{-1/2}}{B\left(\frac{\nu}{2},\frac{1}{2}\right)}\int_0^{+\infty} \nu^{(\nu+1)/2}x^{k-2(\nu+1)/2}\left(\frac{x^2}{\nu+x^2}\right)^{(\nu+1)/2}dx \\
&= 2\frac{\nu^{-1/2}}{B\left(\frac{\nu}{2},\frac{1}{2}\right)}\int_0^{1} \nu^{(\nu+1)/2}\nu^{(k-(\nu+1))/2}\left(\frac{y}{1-y}\right)^{(k-(\nu+1))/2}y^{(\nu+1)/2}\frac{\nu^{1/2}}{2}y^{-1/2}(1-y)^{-3/2}dy \\
&= 2\frac{\nu^{-1/2}}{B\left(\frac{\nu}{2},\frac{1}{2}\right)}\int_0^{1} \frac{\nu^{(\nu+1)/2}\nu^{(k-\nu-1)/2}\nu^{1/2}}{2}y^{(\nu+1)/2-1/2+(k-\nu-1)/2}(1-y)^{-3/2-(k-\nu-1)/2}dy \\
&= 2\frac{\nu^{-1/2}}{B\left(\frac{\nu}{2},\frac{1}{2}\right)}\int_0^{1} \frac{\nu^{(k-1)/2}}{2}y^{(k+1)/2}(1-y)^{(\nu-k-2)/2}dy \\
&= \frac{\nu^{k/2}}{B\left(\frac{\nu}{2},\frac{1}{2}\right)}\int_0^{1} y^{(k+1)/2}(1-y)^{(\nu-k-2)/2}dy \\
&= \frac{\nu^{k/2}}{B\left(\frac{\nu}{2},\frac{1}{2}\right)}B\left(\frac{k+1}{2},\frac{\nu-k}{2}\right).
\end{aligned}$$

■

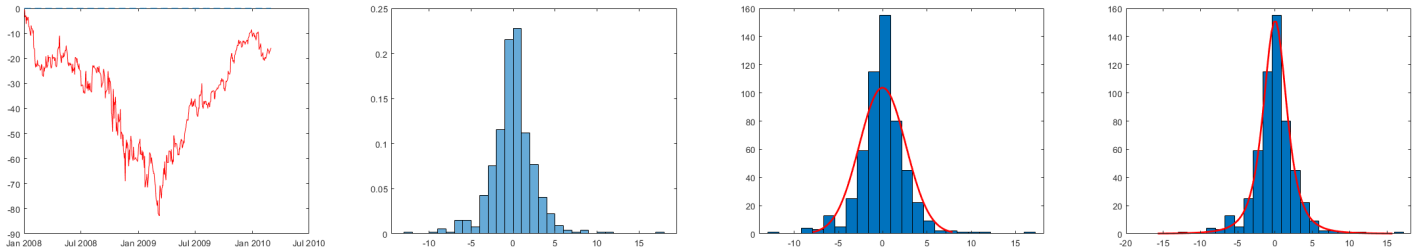

Figure 4: The t and Gaussian distribution fitted to actual logged returns data.

Example 17. It has been shown that (univariate) returns data or portfolio returns can be accurately modeled with t distributions and noncentral versions of it. This is especially true in times of crisis where the Gaussian or other types of light-tail distributions are no longer able to capture the high probability of extreme events. As a (more or less random) example consider the logged returns of the Microsoft Corporation from January 2008 to March 2010. The first picture in Figure 4 displays histograms of those logged returns over the whole period. The histogram admits clear signs of a heavy tailed distribution, which is not surprising given this time of crisis. In the next pictures, a Gaussian and t-distribution were fitted. As expected the Gaussian is not able to capture the more extreme events and deems it too rare to happen. Note however that,

(i) we implicitly assume to observe an iid. sample if we fit a distribution like this. In this example this is certainly not true, the daily return obviously depends on the return on earlier days and the distribution might not be the same each day. However though the heavy-tailedness is usually less pronounced when accounting for this unrealistic assumption (say by using a GARCH filter first), a similar pattern often remains.

(ii) the estimated degrees of freedom ν given by Matlab's "fitdist" routine are still 2.94. If this value were to be correct, it is at least high enough for both mean and variance to exists. However $\mathbb{E}[X^3]$ or $\mathbb{E}[X^4]$ already would not be defined, in particular we could not calculate a measure for skewness or kurtosis in this case.

$\Diamond$

For $\nu = 1$, (37) simplifies to the *Cauchy* density:

$$f(x) = \frac{\Gamma(1)}{\sqrt{\pi}\Gamma\left(\frac{1}{2}\right)}\left(1+x^2\right)^{-1} = \frac{1}{\pi\left(1+x^2\right)}, \tag{41}$$

since $\Gamma(1) = (1-1)! = 1$ and $\Gamma(1/2) = \sqrt{\pi}$. Instead of $X \sim t_1$ we may also say $X \sim$ Cauchy in this case. The cdf is then given as

$$F(x) = \frac{1}{\pi}\arctan(x) + 1/2, \tag{42}$$

as $\arctan'(x) = 1/(1+x^2)$. Since Theorem 3.5 still holds and $\nu = 1$, we see that $\mathbb{E}[X^k]$ does not exist for any $k \geq 1$, i.e. the distribution does not even have a mean. If there is reason to assume

that a sample $(X_1, \ldots, X_n)$ is iid. Cauchy, then calculating mean or variance does not make any sense, since the "population" values of these estimates do not exist. In particular, the CLT or LLN break down and the distribution of $\bar{X}$ will never approach a Gaussian one.

3.2.4 Gamma Distribution

For $\alpha > 0$, we denote $X \sim \text{Gam}(\alpha, 1)$, or $\mu_X = \text{Gam}(\alpha, 1)$, if μ_X has the density $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x) \mathbb{I}_{(0,\infty)}(x). \tag{43}$$

Thus the support of X is $\text{supp}(X) = [0, \infty)$ and on the set $\text{supp}(X)^{\text{o}} = (0, \infty)$ f is continuous. Usually one extends (43) with a scale term $\beta > 0$, which, with Remark 9, results in the density:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp(-x/\beta) \mathbb{I}_{(0,\infty)}(x). \tag{44}$$

We then say $X \sim \text{Gam}(\alpha, \beta)$. For $x \in (0, +\infty)$, using the substitution $s = t/\beta$, $t = s\beta$, $dt = \beta ds$, the cdf is given as

$$\begin{aligned}
F(x) &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_{-\infty}^{x} t^{\alpha-1} \exp(-t/\beta) \mathbb{I}_{(0,\infty)}(t)\ dt \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_{0}^{x} t^{\alpha-1} \exp(-t/\beta)\ dt \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_{0}^{x/\beta} (s\beta)^{\alpha-1} \exp(-s) \beta\ ds \\
&= \frac{1}{\Gamma(\alpha)} \int_{0}^{x/\beta} s^{\alpha-1} \exp(-s)\ ds \\
&= \frac{\Gamma_{x/\beta}(\alpha)}{\Gamma(\alpha)},
\end{aligned} \tag{45}$$

and $F(x) = 0$, for $x \leq 0$. Clearly $\lim_{x \to +\infty} F(x) = 1$. To calculate the moments with the scale term included, we perform the substitution $u = x/\beta$, so that $dx = \beta du$. Then for any $k \in \mathbb{N}$:

$$\begin{aligned}
\mathbb{E}[X^k] &= \int_0^\infty x^k \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp(-x/\beta)\ dx \\
&= \int_0^\infty \frac{x^\alpha}{\beta^\alpha \Gamma(\alpha)} x^{\alpha+k-\alpha-1} \exp(-x/\beta)\ dx \\
&= \int_0^\infty \frac{u^\alpha}{\Gamma(\alpha)} (\beta u)^{k-1} \exp(-u) \beta\ du \\
&= \frac{\beta^k}{\Gamma(\alpha)} \int_0^\infty u^{\alpha+k-1} \exp(-u)\ du \\
&= \frac{\beta^k \Gamma(\alpha + k)}{\Gamma(\alpha)}.
\end{aligned}$$

In particular the expected value is given as $\mathbb{E}[X] = \beta\Gamma(\alpha + 1)/\Gamma(\alpha) = \beta\alpha$ and since

$$\mathbb{E}[X^2] = \frac{\beta^2 \Gamma(\alpha + 2)}{\Gamma(\alpha)} = \frac{\beta^2 (\alpha + 1)\Gamma(\alpha + 1)}{\Gamma(\alpha)} = \frac{\beta^2 (\alpha + 1)\alpha\Gamma(\alpha)}{\Gamma(\alpha)} = \beta^2 \alpha(1 + \alpha),$$

the variance is given as:

$$\mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \beta^2\alpha(1+\alpha) - (\beta\alpha)^2 = \beta^2(\alpha + \alpha^2 - \alpha^2) = \beta^2\alpha.$$

The gamma distribution is extremely flexible and useful in modeling random phenomena with positive outputs. Moreover as we have seen for the t-distribution (and will see in the next subsection), one can combine a gamma distribution with Gaussian ones in a mixture, thereby creating powerful new distributions. Moreover, while there is no straightforward generalization of the gamma to the multivariate case, these mixtures easily extend to $\mathbb{R}^d$, as we will see in Section 3.4. In fact the gamma distribution is so general, that two important and often used distributions arise as a special case:

- X follows an *exponential distribution*, denoted $X \sim \text{Exp}(\lambda)$ or $\mu_X = \text{Exp}(\lambda)$, if $X \sim \text{Gam}(1, 1/\lambda)$. From (44) its density is given as,

$$f(x) = \lambda \exp(-x\lambda)\mathbb{I}_{(0,\infty)}(x). \tag{46}$$

From above

$$\mathbb{E}[X^k] = \frac{\Gamma(1+k)}{\lambda^k},$$

and in particular

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{\lambda} \\
\mathbb{E}[X^2] &= \frac{2}{\lambda^2} \\
\mathbb{V}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.
\end{aligned}$$

One application of the exponential distribution is the modeling of lifetimes. In basic statistics one usually talks about the lifetime of components, say a light bulb. In marketing it is used extensively to model the "lifetime" of a customer (the "death" of the customer is the moment he breaks ties with the firm). This is incorporated in probabilistic models of customer behavior, like the ParetoNBD model. However the exponential distribution has an interesting feature that has to be considered in the decision whether or not it should be used as a modeling tool. This is the property of "memorylessness". Consider $t, s \geq 0$, and imagine we want to find the conditional probability of the event $A = \{\omega : T(\omega) > s + t\}$ given $B = \{\omega : T(\omega) > s\}$. Using the definition of conditional probability in Section (2.3):

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

Now for any $x \geq 0$,

$$P(X > x) = \int_x^\infty f(t)dt = \int_x^\infty \lambda \exp(-t\lambda)dt = [-\exp(-t\lambda)]_x^\infty = \exp(-x\lambda).$$

So:

$$P(X > s+t|X > s) = \frac{\exp(-(s+t)\lambda)}{\exp(-s\lambda)} = \exp(-t\lambda) = P(X > t).$$

To make this more concrete, assume $X \sim \text{Exp}(\lambda)$ is used to model human life span. Then the memorylessness property means that the probability that an individual lives for 20 years more is always the same; The probability that she lives longer than 110, given that she is already over 90, is the same as the probability that she lives to be over 60, given that she is over 40, is the same as just the unconditional probability that she lives past 20. In this case this assumption is clearly not tenable. However in other applications, such as modeling the lifespan of mechanical components, this might be a good-enough approximation.

- X follows an χ^2 *distribution with* ν *degrees of freedom*, denoted $X \sim \chi^2_\nu$ or $\mu_X = \chi^2_\nu$, if $X \sim \text{Gam}(\nu/2, 2)$. From (44) its density is given as,

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} \exp(-x/2)\mathbb{I}_{(0,\infty)}(x). \tag{47}$$

From above

$$\mathbb{E}[X^k] = \frac{2^k\Gamma(\nu/2+k)}{\Gamma(\nu/2)},$$

and in particular

$$\begin{aligned}
\mathbb{E}[X] &= \nu \\
\mathbb{E}[X^2] &= \nu(2+\nu) \\
\mathbb{V}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \nu(2+\nu) - \nu^2 = 2\nu.
\end{aligned}$$

The χ^2 is important, because it arises in a very peculiar situation; If $Y \sim N(0,1)$, then $X = Y^2$ has $X \sim \chi^2_1$. Let us quickly proof this with the help of Remark 9. Unfortunately $g : \mathbb{R} \to \mathbb{R}$, $g(y) = y^2$ is not monotone and also not directly invertible (if $g(y) = x > 0$, then there are two possibilities, $y = +\sqrt{x}$ or $y = -\sqrt{x}$). However, g is strictly decreasing and invertible on $(-\infty, 0)$ and strictly increasing and invertible on $(0, +\infty)$, so:

$$\begin{aligned}
g_{|(-\infty,0)} : (-\infty, 0) \to (0,\infty), \quad g^{-1}_{|(-\infty,0)}(x) = -\sqrt{x}, \quad \frac{dg^{-1}_{|(-\infty,0)}(x)}{dx} = -\frac{x^{-1/2}}{2} < 0 \\
g_{|(0,+\infty)} : (0, +\infty) \to (0,\infty), \quad g^{-1}_{|(0,+\infty)}(x) = \sqrt{x}, \quad \frac{dg^{-1}_{|(0,+\infty)}(x)}{dx} = \frac{x^{-1/2}}{2} > 0.
\end{aligned}$$

Moreover the density of $Y \sim N(0,1)$ is continuous on these intervals. So we can use the same approach as in Remark 9, to obtain for $x \in (0,\infty)$:

$$F_X(x) = P(Y^2 \le x) = P(|Y| \le \sqrt{x}) = P(-\sqrt{x} \le Y \le \sqrt{x})) = F_Y(\sqrt{x}) - F_Y(-\sqrt{x})$$

where the last equality follows from the continuity of F_Y on $\mathbb{R}$. In other words for $x \in (0,\infty)$

$$F_X(x) = F_Y(g^{-1}_{|(0,+\infty)}(x)) - F_Y(g^{-1}_{|(-\infty,0)}(x)),$$

while for $x \leq 0$, $F_X(x) = 0$. Since f_Y is continuous everywhere we can differentiate this with respect to x to obtain for $x \in (0, \infty)$:

$$\begin{aligned} f_X(x) &= f_Y(g^{-1}_{|(0,+\infty)}(x)) \frac{dg^{-1}_{|(0,+\infty)}(x)}{dx} - f_Y(g^{-1}_{|(-\infty,0)}(x)) \frac{dg^{-1}_{|(-\infty,0)}(x)}{dx} \\ &= f_Y(\sqrt{x}) \frac{x^{-1/2}}{2} + f_Y(-\sqrt{x}) \frac{x^{-1/2}}{2} \\ &= 2 f_Y(\sqrt{x}) \frac{x^{-1/2}}{2}, \end{aligned}$$

where we also used that $f_Y(y) = f_Y(-y)$ for all $y \in \mathbb{R}$. For $x \leq 0$, once again $f_X(x) = 0$. Putting the density of Y from (35) into this equation finally gives:

$$\begin{aligned} f_X(x) &= 2 \frac{1}{\sqrt{2\pi}} \exp\left(-(\sqrt{x})^2/2\right) \frac{x^{-1/2}}{2} \mathbb{I}_{(0,+\infty)} \\ &= \frac{1}{2^{1/2}\Gamma(1/2)} x^{1/2-1} \exp\left(-x/2\right) \mathbb{I}_{(0,+\infty)}, \end{aligned}$$

which is the density of a χ^2_1 distribution (and therefore uniquely identifies the distribution of X to be χ^2_1). More generally, if $Y_1, \ldots, Y_n$ are independent $N(a_i, \sigma_i^2)$ random variables, then with

$$Z_i = \frac{Y_i - a}{\sigma}, \quad \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n.$$

This could for instance easily be proven using characteristic functions.

The importance of the χ^2 distribution then often arises not in the direct modeling of real world phenomena, but instead in the process of constructing statistical tests. For instance, if we have an iid. sample of Gaussian r.v. $(X_1, \ldots, X_n)$ with variance σ^2, then it can be shown that $(n-1)S^2/\sigma^2 = \sum_{i=1}^n (X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$. Intuitively speaking, estimating the mean with $\bar{X}$ costs us one degree of freedom. Related to this, the omnipresent "Wald-test" in Econometrics is approximately χ^2 distributed under H_0. This in turn is strongly related to the F-test. In fact, the F distribution arises as a ratio of two independent χ^2 random variable. See for instance Paolella (2006, Chapter 9).

3.2.5 Variance–Gamma distribution

For $\lambda > 0$, $\alpha > 0$, $\beta \in (-\alpha, \alpha)$ and $a \in \mathbb{R}$, we denote $X \sim \mathrm{VG}(\lambda, \alpha, \beta, a)$, or $\mu_X = \mathrm{VG}(\lambda, \alpha, \beta, a)$, if μ_X has the density $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = \frac{2\left(\frac{\alpha^2-\beta^2}{2}\right)^{\lambda}}{\sqrt{2\pi}\Gamma(\lambda)} \left(\frac{|x-a|}{\alpha}\right)^{\lambda-\frac{1}{2}} K_{\lambda-\frac{1}{2}}(\alpha|x-a|) e^{\beta(x-a)}, \tag{48}$$

where

$$K_\nu(x) := 1/2 \int_0^{+\infty} t^{\nu-1} \exp(-1/2x(t + t^{-1}))\, dt, \quad x > 0$$

is the modified Bessel function of the third kind, as given for example in Paolella (2007, p. 300). This distribution is a special case of the even more general "Generalized hyperbolic" (Ghype)

distribution. As might be surmised from the number of parameters, the VG distribution is extremely flexible. In fact the more general Ghype distribution is almost too flexible, leading to estimation problems. This is one of the reasons, why one instead focuses on special cases, such as the VG distribution, which is easier to estimate and still able to capture complex empirical distributions.[17] One can again show that all positive moments of this distribution exists. Though we will not go into this, this distribution has so called *semi-heavy* tails; It means the tails do not decay exponentially fast as in the Gaussian distribution (light tails), but are also not so heavy to have some positive moments not existing (such as the t-distribution). Much more interesting however, is how the distribution arises: Similar as in the case of the t-distribution, let $G \sim \text{Gam}(\lambda, 2/(\alpha^2 - \beta^2))$ (that is where the condition $\beta \in (-\alpha, \alpha)$ comes from). Additionally given a realization $g = G(\omega)$, let $X \sim N(a+\beta g, g)$, or somewhat imprecisely: $X|G \sim N(a+\beta G, G)$. Then, integrating out the latent variable G, gives us the marginal distribution of $X \sim \text{VG}(\lambda, \alpha, \beta, a)$:

$$f(x) = \int_0^\infty \phi(x; a + \beta g, g) f_G(g) dg.$$

We will not do these here, but it is not very difficult to show that (48) is the result of the above integration. This immediately gives a way of simulating variance–gamma random variables if one can already simulate from the Gaussian and gamma distribution (which could be done with the method presented in Remark 2): For a given set of parameters, simply draw $G \sim \text{Gam}(\lambda, 2/(\alpha^2 - \beta^2))$ to give a realization g. Based on this realization draw $X \sim N(a + \beta g, g)$. Doing this say N times gives a sample of n iid. variance gamma distributed random variables. Though we did not mention this in Section 3.2.3, the same principle can of course be used to simulate from a t-distribution.

Let us make an informal argument to show what the moments are (the argument itself is correct, however we did not properly define the involved quantities, so in that sense it is informal): Let for simplicity $a = 0$. As mentioned $X|G \sim N(\beta G, G)$, and using the *law of total expectation*, for $k \in \mathbb{N}$:

$$\begin{aligned}
\mathbb{E}[X^k] &= \mathbb{E}[\mathbb{E}[X^k|G]] \\
&= \mathbb{E}\left[\sum_{l=0}^{k} \binom{k}{l} (\beta G)^{k-l} (\sqrt{G})^l \frac{2^{l/2}\Gamma(l/2+1/2)}{\sqrt{\pi}} \mathbb{I}_{\mathbb{N}\cup\{0\}}(l/2)\right] \\
&= \sum_{l=0}^{k} \binom{k}{l} \beta^{k-l} \frac{2^{l/2}\Gamma(l/2+1/2)}{\sqrt{\pi}} \mathbb{I}_{\mathbb{N}\cup\{0\}}(l/2) \mathbb{E}\left[G^{k-l} G^{l/2}\right] \\
&= \sum_{l=0}^{k} \binom{k}{l} \beta^{k-l} \frac{2^{l/2}\Gamma(l/2+1/2)}{\sqrt{\pi}} \mathbb{I}_{\mathbb{N}\cup\{0\}}(l/2) \mathbb{E}\left[G^{k-l/2}\right], \\
&= \sum_{l=0}^{k} \binom{k}{l} \frac{2^{k-l/2}\Gamma(\lambda+k-l/2)}{(\alpha^2-\beta^2)^{k-l/2}\Gamma(\lambda)} \beta^{k-l} \frac{2^{l/2}\Gamma(l/2+1/2)}{\sqrt{\pi}} \mathbb{I}_{\mathbb{N}\cup\{0\}}(l/2),
\end{aligned}$$

using both the expression for the kth moment of a Gaussian and a gamma distribution. So as one might have expected, the kth moment of X is a mix of the kth moment of a Gaussian and a

[17] Note that in standard Matlab, this distribution is not even implemented!

gamma distribution. In particular

$$\mathbb{E}[X] = \beta \frac{2\Gamma(\lambda+1)}{(\alpha^2-\beta^2)\Gamma(\lambda)} = \frac{2\beta\lambda}{\alpha^2-\beta^2}$$
$$\mathbb{E}[X^2] = \frac{2^2\Gamma(\lambda+2)}{(\alpha^2-\beta^2)^2\Gamma(\lambda)}\beta^2 + \frac{2\Gamma(\lambda+1)}{(\alpha^2-\beta^2)\Gamma(\lambda)}\frac{2\Gamma(1+1/2)}{\sqrt{\pi}} = \frac{4(\lambda+1)\lambda}{(\alpha^2-\beta^2)^2}\beta^2 + \frac{2\lambda}{\alpha^2-\beta^2}$$

Before we change to discrete random variables, let us make a few remarks:

- Moments also exist sometimes for k not necessarily in $\mathbb{N}$. For instance, we could calculate $\mathbb{E}[\sqrt{X}]$ or any other exponent of X. In particular, $k < 0$ is possible, so that $\mathbb{E}[X^k] = \mathbb{E}[(1/X)^\kappa]$, with $\kappa = -k > 0$. For $k = 0$, we obtain

$$\mathbb{E}[X^k] = \mathbb{E}[1] = \int_{-\infty}^{\infty} 1 f(x) dx = 1.$$

- Note that with the densities defined in this section we have essentially reduced the problem of characterizing a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ to finding a finite set of parameters: If we know the two parameters a and σ we completely characterized μ if μ is the Gaussian distribution and analogously for any other distribution we looked at.

- Also note that the common practice to estimate the mean and variance of a given dataset (in finance and many other areas) loses a lot of its appeal in the case of non-Gaussian data. If the data were Gaussian, then calculating say $\bar{X} = 1/n \sum_{i=1}^n X_i$ and $S^2 = 1/(n-1) \sum_{i=1}^n (X_i - \bar{X})^2$ as estimators of a and σ^2 indeed makes sense. However if the data would be t distributed say, or worse Cauchy distributed, then trying to estimate the mean or variance does not make sense (either of them might not even exist). In Finance the Markowitz approach, which still seems to be in ample use, assumes a multivariate Gaussian distribution of asset returns. With this assumption it is indeed enough to focus on the mean and variance, and trying to minimize the latter as a measure of risk. However non-Gaussianity is common in Stock Market data, especially during times of crisis. Once the Gaussianity assumption is lifted in favor of more realistic distributions (such as the multivariate Variance–Gamma distribution below), focussing on mean and variance can no longer be justified.

3.3 Discrete Distributions

3.3.1 Bernoulli and Binomial

3.3.2 Geometric and negative binomial

3.3.3 Poisson

3.4 Multivariate Distributions

We will now change from $\mathbb{R}$ to $\mathbb{R}^d$ and study random vectors $\mathbf{X} : \Omega \to \mathbb{R}^d$. In other words

$$\mathbf{X}(\omega) = (X_1(\omega), \ldots, X_d(\omega)),$$

with $X_1, \ldots, X_d$ defined on Ω. First note that this highlights another great advantage of the underlying sample space: We can generally take all random variables / vectors / elements under consideration to be defined on the same measurable space $(\Omega, \mathcal{A})$. Since we never actually need to describe this space, imagination is basically unlimited here. This is important also in the study of asymptotic behavior, it turns out that one may even define a countably infinite sequence of random elements on Ω. For instance one can look at iid. random vectors $(\mathbf{X}_i)_{i\in\mathbb{N}}$.

Observing the above equation, we can look at $\mathbf{X}$ in at least two ways: We can regard it as a measurable map into the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, or we can look at the respective constituents $X_1, \ldots, X_d$ and regard $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ as a Cartesian product:

Definition 3.7 (Cartesian Product)**.** Let $\Omega_1, \Omega_2, \ldots \Omega_d$, $1 \leq d \leq +\infty$, be a countable collection of sets (i.e. finite or countably infinite). The Cartesian product is then then defined as

$$\prod_{i=1}^{d} \Omega_i = \{(\omega_1, \ldots, \omega_d) : \omega_i \in \Omega_i \text{ for all } i\}.$$

So the Cartesian product is just the sets of all ordered tuples, where the element at position i is in the space Ω_i. One can then also define such a product of the σ-algebra:

Definition 3.8. Let $\mathcal{A}_1, \mathcal{A}_2, \ldots \mathcal{A}_d$, $1 \leq d < +\infty$, be a countable collection of σ-algebras on the spaces $\Omega_1, \Omega_2, \ldots \Omega_d$. The product σ-algebra is then defined as the smallest σ-algebra containing the set

$$\mathcal{A}_0 = \{A_1 \times A_2 \times \ldots A_d : A_i \in \mathcal{A}_i \text{ for all } i\}. \tag{49}$$

It turns out that:

Theorem 3.6. *The Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ (Remember this is the smallest σ-algebra containing all open subsets of $\mathbb{R}^d$) has*

$$\mathcal{B}(\mathbb{R}^d) = \prod_{i=1}^{d} \mathcal{B}(\mathbb{R}).$$

In other words it is also the smallest σ-algebra containing

$$\{A_1 \times \ldots \times A_d : A_i \in \mathcal{B}(\mathbb{R}) \text{ for all } i\}$$

Recall that $\mathbb{R}^d$ is simply defined to be the Cartesian product over $d \in \mathbb{N}$ copies of $\mathbb{R}$. This is however not self-evident for the Borel σ algebra, making the above example interesting. We will however not present a proof here, see for instance Dudley (2002, Proposition 4.1.7). We also note, that this σ-algebra does exactly what it is supposed to in the following sense:

Theorem 3.7. *If $X_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, d$ are $\mathcal{A}/\mathcal{B}(\mathbb{R})$ measurable maps, then the map $\omega \mapsto \mathbf{X}(\omega) = (X_1(\omega), \ldots, X_d(\omega))$ is $\mathcal{A}/\prod_{i=1}^{d} \mathcal{B}(\mathbb{R})$ measurable.*

Proof. We want to show that

$$\mathbf{X}^{-1}(B) \in \mathcal{A}$$

for all $B \in \prod_{i=1}^{d} \mathcal{B}(\mathbb{R})$. Let us frame this differently, and instead define the set of sets

$$\mathcal{C} = \{B \subset \mathbb{R}^d : \mathbf{X}^{-1}(B) \in \mathcal{A}\}.$$

If we can proof that $\prod_{i=1}^{d} \mathcal{B}(\mathbb{R}) \subset \mathcal{C}$, we have automatically shown that the measurability condition holds! First, we check that

$$\mathcal{B}_0 = \{B_1 \times B_2 \ldots \times B_d : B_i \in \mathcal{B}(\mathbb{R})\} \subset \mathcal{C}.$$

In this case for $B_i \in \mathcal{B}(\mathbb{R})$ arbitrary

$$\begin{aligned}
\mathbf{X}^{-1}(B_1 \times B_2 \ldots \times B_d) &= \{\omega \in \Omega : \mathbf{X}(\omega) \in B_1 \times B_2 \ldots \times B_d\} \\
&= \{\omega \in \Omega : (X_1(\omega), \ldots, X_d(\omega)) \in B_1 \times B_2 \ldots \times B_d\} \\
&= \bigcap_{i=1}^{d} \{\omega \in \Omega : X_i(\omega) \in B_i\} \\
&= \bigcap_{i=1}^{d} X_i^{-1}(B_i) \in \mathcal{A}.
\end{aligned}$$

So indeed $\mathcal{B}_0 \subset \mathcal{C}$. However $\prod_{i=1}^{d} \mathcal{B}(\mathbb{R})$ is the smallest σ-algebra containing $\mathcal{B}_0$. Furthermore, with the properties of the inverse image it can easily be shown that $\mathcal{C}$ is a σ-algebra itself (remember the conditions from definition 2.1). But then $\mathcal{B}_0 \subset \mathcal{C}$ immediately implies that $\prod_{i=1}^{d} \mathcal{B}(\mathbb{R}) \subset \mathcal{C}$ as well. ■

With Theorem 3.6 this also means that $\mathbf{X} = (X_1, \ldots, X_d)$ is $\mathcal{A}/\mathcal{B}(\mathbb{R}^d)$ measurable whenever each X_i is $\mathcal{A}/\mathcal{B}(\mathbb{R})$ measurable. On the other hand it can also be shown that if $\mathbf{X}$ is $\mathcal{A}/\mathcal{B}(\mathbb{R}^d) = \prod_{t=1}^{d} \mathcal{B}(\mathbb{R})$ measurable, then each X_i, $i = 1, \ldots, d$ is $\mathcal{A}/\mathcal{B}(\mathbb{R})$ measurable. This is easily demonstrated by defining for each i the *projection* $\pi_i : \mathbb{R}^d \to \mathbb{R}$, $\pi_i(\mathbf{x}) = x_i$. In other words this function just takes out the ith element from $\mathbf{x}$.[18] It is easy to show that π_i is continuous and therefore $\mathcal{B}(\mathbb{R}^d) \to \mathcal{B}(\mathbb{R})$ measurable. Since also

$$X_i = \pi_i \circ \mathbf{X},$$

we have that X_i is a composition of two measurable functions, and thus itself measurable. Since this holds for all $i = 1, \ldots, d$, we have with Theorem 3.7:

$$\mathbf{X} = (X_1, \ldots, X_d) \ \mathcal{A}/\mathcal{B}(\mathbb{R}^d)\text{-measurable} \iff X_i \ \mathcal{A}/\mathcal{B}(\mathbb{R})\text{-measurable for all } i.$$

A particular interesting example of a Cartesian product is the d-dimensional interval (a cube for $d = 3$): Consider for $j = 1, \ldots, d$, the open intervals (a_j, b_j) with $b_j > a_j$. Then the d-dimensional open interval is given as

$$(\mathbf{a}, \mathbf{b}) = \prod_{j=1}^{d} (a_j, b_j) = \{\mathbf{x} \in \mathbb{R}^d : a_j < x_j < b_j \text{ for all } j\}.$$

[18] In this case this is fairly simple, it is just the dot product of two vectors. It gets however more interesting in infinite-dimensional products for which this is also possible.

Similarly we define $(\mathbf{a}, \mathbf{b}]$, $[\mathbf{a}, \mathbf{b})$ and $[\mathbf{a}, \mathbf{b}]$ as the product of intervals $(a_j, b_j]$, $[a_j, b_j)$ and $[a_j, b_j]$ respectively. The lebesgue measure λ on $\mathbb{R}^d$ is then the unique measure assigning each such interval it's "length":

$$\lambda((\mathbf{a}, \mathbf{b})) = \lambda((\mathbf{a}, \mathbf{b}]) = \lambda([\mathbf{a}, \mathbf{b})) = \lambda([\mathbf{a}, \mathbf{b}]) = \prod_{j=1}^{d} (b_j - a_j).$$

In other words, for a rectangle in $\mathbb{R}^3$, we just get back the volume. For a $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R})$ measurable function $f : \mathbb{R}^d \to \mathbb{R}$ we can again define the Lebesgue integral

$$\int_A f d\lambda = \int_A f(\mathbf{x}) d\mathbf{x},$$

for any $A \in \mathcal{B}(\mathbb{R}^d)$. To help connect this back to the case on $\mathbb{R}$, the powerful Fubini-Tonelli theorem is available:

Theorem 3.8 (Adaptation of Theorem 1.7.2 in Durrett (2010))**.** *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a $\mathcal{B}(\mathbb{R}^2)/\mathcal{B}(\mathbb{R})$ measurable function with $f \geq 0$ or $\int_{\mathbb{R}^2} |f| \, d\lambda < \infty$. Then for any $A_1, A_2 \in \mathcal{B}(\mathbb{R})$,*

$$\int_{A_1 \times A_2} f \, d\lambda = \int_{A_1} \int_{A_2} f(x, y) \, dxdy = \int_{A_2} \int_{A_1} f(x, y) \, dydx$$

Through induction it is easy to check that the above theorem is still true if $d > 2$, so that for instance

$$\int_{(\mathbf{a}, \mathbf{b}]} f(\mathbf{x}) d\mathbf{x} = \int_{a_d}^{b_d} \dots \int_{a_1}^{b_1} f(x_1, \dots, x_d) dx_1 \dots dx_d.$$

Similar as on $\mathbb{R}$ one can show that there exists a cdf $F : \mathbb{R}^d \to [0, 1]$ for any random vector $\mathbf{X} : \Omega \to \mathbb{R}^d$. However this is much more cumbersome, both to proof this result and to actually use the cdf, see for instance Durrett (2010). In Section 4.2 we will encounter an additional tool that exists for any random vector on $\mathbb{R}^d$, the characteristic function. Here we instead look at distributions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with a density $f : \mathbb{R}^d \to \mathbb{R}$. Both cdf and pdf again completely characterize the probability, though the former exists for all random vectors.

Definition 3.9 (Definition 12.2 in Jacod and Protter (2004))**.** The density of a probability measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is a nonnegative $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R})$ measurable function $f : \mathbb{R}^d \to \mathbb{R}$ that has for all $A \in \mathcal{B}(\mathbb{R}^d)$:

$$\mu(A) = \int_A f(\mathbf{x}) d\mathbf{x}. \tag{50}$$

If μ is the distribution of a random vector $\mathbf{X}$, then f is called the *joint density* or *joint pdf* of $\mathbf{X} = (X_1, \dots, X_d)$.

Note that this definition is also valid for the case $d = 1$ (as usual) and that it indeed consistent with the one given in Definition 3.4. To proof the next theorem as an analogue to Theorem 3.3 in the case $d = 1$, we need the following fact: If $f : \mathbb{R}^d \to \mathbb{R}$, with $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^d$ (i.e. f is nonnegative) has the condition in (51) fulfilled, then the set function $\mu_2 : \mathcal{B}(\mathbb{R}^d) \to [0, 1]$

$$\mu_2(A) = \int_A f(\mathbf{x}) d\mathbf{x}$$

is a probability measure. Let us quickly demonstrate this: We need to check that μ_2 has (i) $\mu_2(A) \geq 0$ for all $A \in \mathcal{B}(\mathbb{R})$, (ii) $\mu(\mathbb{R}^d) = 1$ and (iii) countable additivity. (i), (ii) are quite easy, since $f \geq 0$, clearly also means $\int_A f(\mathbf{x})d\mathbf{x} \geq 0$ for any $A \in \mathcal{B}(\mathbb{R}^d)$, and since (ii) holds by the assumption in Equation (51). Let now $(A_n)_n$ be a countable disjoint family of sets in $\mathcal{B}(\mathbb{R}^d)$. Then:

$$\begin{aligned}\mu_2\left(\bigcup_n A_n\right) &= \int_{\bigcup_n A_n} f(\mathbf{x})d\mathbf{x} \\ &= \int_{\mathbb{R}^d} f(\mathbf{x})\mathbb{I}_{\bigcup_n A_n}(\mathbf{x})d\mathbf{x}\end{aligned}$$

In fact this relation is how $\int_A f(\mathbf{x})d\mathbf{x}$ is defined in the first place, just the integral over the whole space with the function $f\mathbb{I}_A$. Further, because the A_n's are disjoint, we have that

$$\mathbb{I}_{\bigcup_n A_n}(\mathbf{x}) = \sum_{n=1}^{\infty} \mathbb{I}_{A_n}(\mathbf{x}).$$

Again this holds, because $\mathbf{x} \in \bigcup_n A_n$ and $(A_n)_n$ disjoint means $\mathbf{x} \in A_n$ for *exactly one* n, similar as in Section 2.2. Thus,

$$\begin{aligned}\mu_2\left(\bigcup_n A_n\right) &= \int_{\mathbb{R}^d} f(\mathbf{x}) \sum_{n=1}^{\infty} \mathbb{I}_{A_n}(\mathbf{x})d\mathbf{x} \\ &= \int_{\mathbb{R}^d} f(\mathbf{x}) \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{I}_{A_n}(\mathbf{x})d\mathbf{x}\end{aligned}$$

Now we want to exchange this limit with the integral. This is always a dangerous thing to do, but luckily we can use either the monotone or dominated convergence theorems, which were already quickly mentioned in the univariate case. Without going into details, they both tell us that limit and integral can be exchanged in this case, so that by the linearity of the integral:

$$\begin{aligned}\mu_2\left(\bigcup_n A_n\right) &= \lim_{N\to\infty} \int_{\mathbb{R}^d} f(\mathbf{x}) \sum_{n=1}^{N} \mathbb{I}_{A_n}(\mathbf{x})d\mathbf{x} \\ &= \lim_{N\to\infty} \sum_{n=1}^{N} \int_{\mathbb{R}^d} f(\mathbf{x})\mathbb{I}_{A_n}(\mathbf{x})d\mathbf{x} \\ &= \sum_{n=1}^{\infty} \int_{\mathbb{R}^d} f(\mathbf{x})\mathbb{I}_{A_n}(\mathbf{x})d\mathbf{x} \\ &= \sum_{n=1}^{\infty} \mu_2(A_n).\end{aligned}$$

So indeed, since $(A_n)_n$ was disjoint but otherwise an arbitrary sequence in $\mathcal{B}(\mathbb{R}^d)$, we have shown countable additivity and finally that μ_2 is indeed a probability on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. This leads to the following analogue of of Theorem 3.3:

Theorem 3.9. *[Theorem 12.1 in Jacod and Protter (2004)] A nonnegative $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R})$ measurable function $f : \mathbb{R}^d \to \mathbb{R}$ is the density of a probability measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ iff it satisfies*

$$\int_{\mathbb{R}^d} f(\mathbf{y})d\mathbf{y} = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(y_1, \dots, y_d)\ dy_1, \dots, dy_d = 1. \tag{51}$$

In this case f entirely characterizes the probability measure.

Proof. As in the case $d = 1$, if $f \geq 0$ is the density of a probability μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, then by Definition 3.9,

$$1 = \mu(\mathbb{R}^d) = \int_{\mathbb{R}^d} f(\mathbf{y})d\mathbf{y}.$$

f then uniquely characterizes μ through (50). In particular if f_1 is the density of μ_1 and f_2 of μ_2 and $f_1 = f_2$, then $\mu_1 = \mu_2$.

Now assume $f : \mathbb{R}^d \to \mathbb{R}$ is nonnegative and has (51). We have shown above that in this case we can define the new set function $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, 1]$

$$\mu(A) := \int_A f(\mathbf{y})d\mathbf{y},$$

and it will indeed be a probability measure. But then the proof is already done, because we have now found a unique probability measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ induced by f (similar as in the proof of theorem 3.2, where we constructed a measure μ_X from the cdf F alone). By definition μ has f as its density. ■

Remark 10. Once again we have that a given probability μ only defines its density up to sets of measure zero. That is if f is the density of μ and $g = f$ on a set $A \in \mathcal{B}(\mathbb{R}^d)$ with $\lambda(A^c) = 0$ (i.e. $g = f$ almost everywhere), then g is also a density of μ, since for all $A \in \mathcal{B}(\mathbb{R}^d)$:

$$\int_A g d\lambda = \int_A f d\lambda.$$

♢

It is of high interest to determine what relation a joint density has to the univariate densities of each component of $\mathbf{X}$. The following is for simplicity shown for $d = 2$, but it can again be easily extended to the case $d > 2$ by an induction argument.

Theorem 3.10 (Theorem 12.2 in Jacod and Protter (2004))**.** *Assume $\mathbf{X} = (X_1, X_2)$ has joint density $f : \mathbb{R}^2 \to \mathbb{R}$. Then*

a) Both X_1 and X_2 have densities on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by :

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2 \quad f_{X_2}(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_1$$

b) X_1 and X_2 are independent iff

$$f(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2) \quad \lambda\text{-almost everyhwere}$$

c) We can define another density denoted $f_{X_1=x_1} : \mathbb{R} \to \mathbb{R}$ *as*

$$f_{X_1=x_1}(x_2) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)},$$

for all x_1 *st.* $f_{X_1}(x_1) \neq 0$.

Proof. a) First note that if for any probability measure P on $(\Omega, \mathcal{A})$, if $P(A) = 1$, for some $A \in \mathcal{A}$, then for any $B \in \mathcal{A}$:

$$P(B) = P(B \cap A) + P(B \cap A^c) = P(B \cap A),$$

since $P(B \cap A^c) \leq P(A^c) = 0$. So indeed $P(B)$ and $P(B \cap A)$ are the same in this case. Now, for each $B \in \mathcal{B}(\mathbb{R})$, it holds that (Remember, both X_1 and X_2 are defined on $(\Omega, \mathcal{A}, P)$),

$$\mu_{X_1}(B) = P(X_1^{-1}(B) \cap X_2^{-1}(\mathbb{R})) = P(\mathbf{X}^{-1}(B \times \mathbb{R})) = \mu_{\mathbf{X}}(B \times \mathbb{R}).$$

But since $\mu_{\mathbf{X}}$ has density f this means that

$$\begin{aligned}
\mu_{X_1}(B) &= \mu_{\mathbf{X}}(B \times \mathbb{R}) \\
&= \int_{B \times \mathbb{R}} f(\mathbf{x}) dx_2 dx_1 \\
&= \int_B \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 dx_1,
\end{aligned}$$

where we used Fubini in the first step, which we can do since $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^2$. If we now define the function

$$f_{X_1}(x_1) := \int_{-\infty}^{\infty} f(x_1, x_2) dx_2,$$

it follows immediately from the above and Definition 3.9 that this is a valid candidate for the density of X_1.

b) We will wait with this part of the proof, until we properly defined independence below.

c) Clearly $f_{X_1=x_1} : \mathbb{R} \to \mathbb{R}$ is nonnegative. Furthermore

$$\begin{aligned}
\int_{-\infty}^{\infty} f_{X_1=x_1}(x_2) dx_2 &= \int_{-\infty}^{\infty} \frac{f(x_1, x_2)}{f_{X_1}(x_1)} dx_2 \\
&= \frac{1}{f_{X_1}(x_1)} \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \\
&= \frac{1}{f_{X_1}(x_1)} f_{X_1}(x_1) \\
&= 1
\end{aligned}$$

So it is indeed a valid density.

■

An example of c) was given above: For a Gaussian mean variance mixture, we take f_{X_1} to be the density of a gamma distribution and $f_{X_1=x_1}(x_2)$ to be the density of a Gaussian distribution whose variance and/or mean depend on x_1! Since $f_{X_1}(x_1) > 0$ for all $x_1 > 0$, $f_{X_1=x_1}(x_2)$ is defined for all $x_1 > 0$. We then calculated

$$\begin{aligned}
&\int_0^{+\infty} f_{X_1=x_1}(x_2) f_{X_1}(x_1) dx_1 \\
&= \int_0^{+\infty} f(x_1, x_2) dx_1 \\
&= f_{X_2}(x_2).
\end{aligned}$$

This leads to the following important definition

Definition 3.10. Let $a, \beta \in \mathbb{R}$ and $x \mapsto \phi(x; b, \sigma^2)$ be the density of a Gaussian random variable with mean b and variance σ^2. The random variable X is said to be a continuous mean–variance mixture if its density f_X has representation

$$f_X(x) = \int_0^{\infty} \phi(x; a + \beta g, g) f_G(g) dg \tag{52}$$

for some continuous random variable $G : \Omega \to \mathbb{R}$ with nonnegative support. Equivalently X has stochastic representation

$$X \stackrel{D}{=} a + \beta G + \sqrt{G} Z, \tag{53}$$

with $Z \sim N(0, 1)$ and G, Z independent.

For example for the VG distribution we said:

$$f(x) = \int_0^{\infty} \phi(x; a + \beta g, g) f_G(g) dg,$$

with $G \sim \text{Gam}(\lambda, 2/(\alpha^2 - \beta^2))$.

Remark 11. The density $f_{X_1=x_1}(x_2)$ is often denoted $f(x_1|x_2)$, the conditional density of x_1 given x_2. This is justified partly by the following fact: Consider $\mathbf{X} = (X_1, \ldots, X_d)$ a random vector with joint density f and a subset say $\mathbf{Y} = (X_{i_1}, \ldots, X_{i_m})$ for $m < d$. Then from the above theorem we may "integrate out" all elements that are not part of $\mathbf{Y}$ to obtain $f_{\mathbf{Y}}$. It can then be shown that for any bounded function $g : \mathbb{R}^d \to \mathbb{R}$ the *conditional expectation* can be calculated as,

$$\mathbb{E}[g(\mathbf{X})|\mathbf{Y} = \mathbf{y}] = \int_{\mathbb{R}^{d-m}} g(\mathbf{x}) f(\mathbf{x}|\mathbf{y}) d(x_j)_{j \notin \{i_1, \ldots, i_m\}}. \tag{54}$$

In particular for $d = 2$:

$$\mathbb{E}[g(X_1, X_2)|X_1 = x_1] = \int_{-\infty}^{\infty} g(x_1, x_2) f(x_2|x_1) dx_2 \tag{55}$$

See for instance Jacod and Protter (2004, Chapter 23) or Paolella (2006, Chapter 8.2.3).

Conditional expectations and probabilities of random vectors are actually a generalization to "regular" expectations and need a surprising amount of work if one wants to define them in

full generality. In particular, they are defined as random variables, i.e. as measurable functions on a sub-σ-algebra $\mathcal{F} \subset \mathcal{A}$ on the probability space $(\Omega, \mathcal{A})$. Then $\mathbb{E}[X|\mathcal{F}]$ is the expected value of X given the information we have in $\mathcal{F}$. An important example of this is when $(\mathcal{F}_n)_n$ with $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{A}$ are so called *filtrations*. These are σ-algebras with some additional properties and they model the information available at time point n. They get larger, as we assume to learn more about the process under consideration with each new time point n. For a $\mathcal{A}/\mathcal{B}(\mathbb{R})$ measurable random variable X, we might then ask what $\mathbb{E}[X|\mathcal{F}_n]$ is.

Finally one can (actually quite easily) show the *law of iterated expectations*:

$$\mathbb{E}_Y[\mathbb{E}[X|Y]] = \mathbb{E}[X], \tag{56}$$

where $\mathbb{E}_Y$ means we are taking the expectation with respect to the measure induced by Y. $\Diamond$

We will now make a slight change in notation and not only look at $\mathbf{x} \in \mathbb{R}^d$ as a tupel, but instead define it to be a column vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} x_1 & \ldots & x_d \end{pmatrix}^T$$

In particular now $\mathbf{X}(\omega) = (X_1(\omega), \ldots, X_d(\omega))^T$. This is the language of linear algebra, and though we could have continued using a more general notation we change it here for convenience.

Since we are only considering distributions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with densities, we can define expected values equivalently as in the case $d = 1$: For $g : \mathbb{R}^d \to \mathbb{R}$ a measurable function, define

$$\mathbb{E}[g(\mathbf{X})^+] = \int_{\mathbb{R}^d} \max(g(\mathbf{x}), 0) f(\mathbf{x}) d\mathbf{x}$$
$$\mathbb{E}[g(\mathbf{X})^-] = \int_{\mathbb{R}^d} -\min(g(\mathbf{x}), 0) f(\mathbf{x}) d\mathbf{x}.$$

If both expectations are finite we define

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

exactly as before. Moreover, the expected value of $\mathbf{X}$ itself is defined as

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X^d] \end{pmatrix}$$

if all involved univariate expected values exist. This is in fact the natural extension of the Lebesgue integral to vector-valued functions and is also used more generally.

An additional complication when talking about distributions on $\mathbb{R}^d$, is that we not just look at $\mathbb{E}[X_j^k]$, for $j = 1, \ldots, d$, but also at things like $\mathbb{E}[X_j^{k_1} X_i^{k_2}]$ say. Or even $\mathbb{E}[X_1^{k_1} X_2^{k_2} \cdots X_d^{k_d}]$. Of special interest thereby is the case of $\mathbb{E}[X_j X_i]$, which properly standardized is the covariance between X_j, X_i:

Definition 3.11. Assume X_1, X_2 are two random variables with $\mathbb{E}[X_1^2], \mathbb{E}[X_2^2]$ existing. Then we call

$$\mathrm{Cov}(X_1, X_2) := \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])],$$

the covariance between X_1 and X_2.

One might notice that the variance is just the covariance between X and itself, so that we also have a proper definition of the variance now. It is furthermore simple to check that

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] \tag{57}$$

and in particular $\mathbb{V}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. If we write down the covariance of each combination of X_i, X_j we obtain the covariance matrix in the d-dimensional case:

$$\mathbb{V}(\mathbf{X}) := \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] = \begin{pmatrix} \mathbb{V}(X_1) & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \dots & \mathrm{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \dots & \mathbb{V}(X_d) \end{pmatrix}$$

From the definition of the covariance, this matrix is obviously symmetric. Moreover it is also *positive semidefinite*. Indeed let $\mathbf{b} \in \mathbb{R}^d$ be arbitrary. We want to show that $\mathbf{b}^T \Sigma \mathbf{b} \geq 0$. However, $\mathbf{b}^T \Sigma \mathbf{b}$ is simply the variance of the univariate random variable $\mathbf{b}^T \mathbf{X}$. That is:

$$0 \leq \mathbb{V}(\mathbf{b}^T \mathbf{X}) = \mathbb{E}[(\mathbf{b}^T \mathbf{X} - \mathbb{E}[\mathbf{b}^T \mathbf{X}])(\mathbf{b}^T \mathbf{X} - \mathbb{E}[\mathbf{b}^T \mathbf{X}])^T] = \mathbf{b}^T \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T]\mathbf{b} = \mathbf{b}^T \Sigma \mathbf{b},$$

proving the claim.

We will now discuss various multivariate distributions. In doing this we will use a somewhat different presentation as in the univariate case: Instead of directly defining a given distribution over its density, we will first show how to construct the distribution from univariate ones and then state the density which comes out of this construction (and which then of course characterizes the probability measure). This should illustrate that, while it is relatively easy on $\mathbb{R}$ to start with a density when defining a new probability measure, things are more complex in $\mathbb{R}^d$. In particular in most approaches we need to use univariate distributions in one way or another to construct multivariate ones.

3.4.1 Independence Distribution

The first multivariate construction we look at is in fact a class of distributions with a certain dependence structure. That is for *any collection* of d random variables $X_1, \dots, X_d$ we can define the dependence structure to be, well, independent. This idea of treating dependency structure and marginal random variables seperately has an important generalization in Copulas, which we will however not discuss here. See for instance McNeil et al. (2005). The convenient definition of independence we give here has its roots in the following most important theorem:

Theorem 3.11. *Let $\mu_1, \ldots, \mu_d$ be probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then there exists a unique probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, denoted $\prod_{j=1}^d \mu_j$ or $\mu_1 \times \mu_2 \times \ldots \mu_d$ such that*

$$\mu_1 \times \mu_2 \times \ldots \mu_d(A_1 \times \ldots \times A_d) = \prod_{j=1}^d \mu_j(A_j)$$

for all $A_j \in \mathcal{B}(\mathbb{R})$, $j = 1, \ldots, d$.

$\mu_1 \times \mu_2 \times \ldots \mu_d$ is often referred to as the *product* (probability) measure. Remember the discussion in the beginning of this section, where we said that $\mathcal{B}(\mathbb{R}^d)$ is the smallest σ-algebra containing the set of sets $\mathcal{A}_0 = \{A_1 \times \ldots \times A_d : A_j \in \mathcal{B}(\mathbb{R}) \text{ for all } j\}$. This theorem now just says that we can find a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ which factors into the "univariate" probability on this set $\mathcal{A}_0$. Additionally this should have a familiar ring to it by now; We generally like to define measures on a "small" and well-behaved set like $\mathcal{A}_0$, or say $I_0 = \{(-\infty, a], a \in \mathbb{R}\}$ and then extend it from there to the whole of $\mathcal{B}(\mathbb{R}^d)$ or $\mathcal{B}(\mathbb{R})$. For instance, the result that the cdf $F(x) = \mu((-\infty, x])$ completely characterizes the probability measure μ is akin to saying that it is enough to define the measure on the set I_0 and that it can be uniquely extended to all of $\mathcal{B}(\mathbb{R})$.

Now independence is tightly tied to Theorem 3.11 as mentioned:

Definition 3.12. Let $X_1, \ldots, X_d$ be random variables with distributions $\mu_1, \ldots, \mu_d$. Then $X_1, \ldots, X_d$ are called independent if $\mathbf{X} = (X_1, \ldots, X_d)$ has distribution $\prod_{j=1}^d \mu_d$ on $\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right)$.

Let us first proof b) from Theorem 3.10 above, namely that X_1, X_2 with joint density f are independent iff $f(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$ λ-almost everyhwere. Indeed if X_1, X_2 are independent, $\mu_{(X_1, X_2)} = \mu_1 \times \mu_2$. Since we assumed (X_1, X_2) to have a joint density, the marginal densities exists and can be found as in a) above. We then consider the new set function

$$\mu_{\text{new}}(A) = \int_A f_{X_1}(x_1) f_{X_2}(x_2)\, dx_1 dx_2.$$

Since $f_{X_1}(x_1) f_{X_2}(x_2) \geq 0$ the above integral is not only well-defined but we can also use Fubini to see that

$$\mu_{\text{new}}(A_1 \times A_2) = \int_{A_1 \times A_2} f_{X_1}(x_1) f_{X_2}(x_2)\, dx_1 dx_2 = \int_{A_1} f_{X_1}(x_1) dx_1 \int_{A_2} f_{X_2}(x_2) dx_2 = \mu_1(A_1)\mu_2(A_2),$$

for all $A_1, A_2 \in \mathcal{B}(\mathbb{R})$. In particular $\mu_{\text{new}}(\mathbb{R}^2) = 1$, so μ_{new} is indeed a probability measure. By theorem 3.11 we immediately know that $\mu_{\text{new}} = \mu_1 \times \mu_2 = \mu_{(X_1, X_2)}$. So in fact f and $f_{X_1} \cdot f_{X_2}$ define the same distribution and thus $f(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$ λ-almost everyhwere. On the other hand if $f(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$ λ-almost everyhwere, then for $A_1, A_2 \in \mathcal{B}(\mathbb{R})$,

$$\mu_{(X_1, X_2)}(A_1 \times A_2) = \int_{A_1 \times A_2} f(x_1, x_2)\, dx_1 dx_2 = \int_{A_1} f_{X_1}(x_1) dx_1 \int_{A_2} f_{X_2}(x_2) dx_2 = \mu_{X_1}(A_1)\mu_{X_2}(A_2),$$

using Fubini. So again we immediately know that $\mu_{(X_1, X_2)}$ is the independence distribution.

In fact, of the crucial peculiarities of the independence distribution is that we can get from the univariate densities to the joint density. This is a slight reformulation of 3.10 b):

Theorem 3.12. *Let $\mathbf{X} = (X_1, \ldots, X_d)$ be independent. Then $\mathbf{X}$ has a joint density $f_{\mathbf{X}}$ iff each X_i has a density f_{X_i}. In this case $f_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^d f_{X_i}(x_j)$ λ-almost everywhere.*

Proof. If $\mathbf{X}$ has a joint density $f_{\mathbf{X}}$, we can use exactly the same proof as above for Theorem 3.10 b) to show that $f_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^d f_{X_i}(x_j)$ λ-almost everywhere. On the other hand if each X_i has density f_{X_i}, define

$$f_{\mathbf{X}}(\mathbf{x}) := \prod_{j=1}^d f_{X_i}(x_j).$$

Then again by the same argument as above, we see that $f_{\mathbf{X}}$ is the density of the independence distribution. So indeed $f_{\mathbf{X}}$ is a density for $\mathbf{X}$. ■

Having defined the distribution, we turn to moments. It turns out that, for $k_j \in \mathbb{N} \cup \{0\}$

$$\mathbb{E}[X_1^{k_1} X_2^{k_2} \cdots X_d^{k_d}] = \mathbb{E}[X_1^{k_1}]\mathbb{E}[X_2^{k_2}] \cdots \mathbb{E}[X_d^{k_d}]. \tag{58}$$

if all invovled moments exist. This is in fact true in full generality, that is for any type of random vector $\mathbf{X}$. However we will only show this for the case that $\mathbf{X}$ has a joint density f. Then if $\mathbb{E}[X_j^{k_j}]$ exists for all k_j (for $k_j = 0$ this is again just 1), we have first of all that

$$\mathbb{E}[(X_j^{k_j})^+] < +\infty, \mathbb{E}[(X_j^{k_j})^-] < +\infty$$

by definition. In particular we also have that

$$\mathbb{E}[|X_j^{k_j}|] = \mathbb{E}[(X_j^{k_j})^+] + \mathbb{E}[(X_j^{k_j})^-] < +\infty.$$

Thus

$$\begin{aligned}
\mathbb{E}[|X_1^{k_1} X_2^{k_2} \cdots X_d^{k_d}|] &= \mathbb{E}[|X_1^{k_1}||X_2^{k_2}| \cdots |X_d^{k_d}|] \\
&= \int_{\mathbb{R}^d} |x_1^{k_1}||x_2^{k_2}| \cdots |x_d^{k_d}| \prod_{j=1}^d f_{X_i}(x_j) d\mathbf{x} \\
&= \int_{\mathbb{R}} |x_1^{k_1}| f_{X_1}(x_1) dx_1 \cdots \int_{\mathbb{R}} |x_d^{k_d}| f_{X_d}(x_d) dx_d < +\infty
\end{aligned}$$

where we again used Fubini in the last step, which is possible because the integrand is nonnegative. So

$$\mathbb{E}[|X_1^{k_1} X_2^{k_2} \cdots X_d^{k_d}|] < +\infty$$

and in particular $\mathbb{E}[X_1^{k_1} X_2^{k_2} \cdots X_d^{k_d}]$ exists. Using the same simple calculations again on $\mathbb{E}[X_1^{k_1} X_2^{k_2} \cdots X_d^{k_d}]$, we also immediately get (58).

Independence plays a hugely important role in statistics, not least due to mathematical convenience. In classical statistics one often deals with iid. sequences such as encountered in the examples in Section 3.2. We can now define what this means in full generality:

Definition 3.13. A collection of random variables $X_1, \ldots, X_d$ is identically and independently distributed (iid.) if $\mu_{X_1} = \mu_{X_2} = \ldots = \mu_{X_d}$ and if they are independent.

We will forego another example here, though we will encounter independence a few more times in the upcoming sections. Even when working with dependence (such as time series model) a good proxy is to choose a model allowing to transform a dependent series into an independent one. For instance, using a GARCH or ARMA filter should in theory result in residual terms that are iid.

3.4.2 Multivariate Gaussian Distribution

There are many ways of actually constructing (the same) multivariate Gaussian distribution. For instance, in Jacod and Protter (2004, Chapter 16) it is defined as the (unique) distribution, such that all linear combinations of a random vector with that distribution are univariate Gaussian again, i.e. $\mathbf{X}$ is said to have multivariate Gaussian distribution if $\mathbf{a}^T\mathbf{X}$ is univariate Gaussian for all $\mathbf{a} \in \mathbb{R}^d$. We will start differently, but then show that this holds indeed true also for our definition.

Definition 3.14. Let $Z_1, \ldots, Z_d$ be iid. $N(0,1)$ and $\mathbf{Z} = (Z_1, \ldots, Z_d)^T$. Let $\mathbf{a} \in \mathbb{R}^d$ and Σ be a symmetric positive semi-definite $d \times d$ matrix. Then we say $\mathbf{Y} \sim N(\mathbf{a}, \Sigma)$ or $\mu_{\mathbf{Y}} = N(\mathbf{a}, \Sigma)$, if

$$\mathbf{Y} \stackrel{D}{=} \mathbf{a} + \Sigma^{1/2}\mathbf{Z}.$$

By the above definition: $\mathbf{Z} \sim N(\mathbf{0}, I)$, with I being the matrix with all ones on the diagonal.

Remark 12. $\Sigma^{1/2}$ is a $d \times d$ matrix such that $\Sigma^{1/2}(\Sigma^{1/2})^T = \Sigma$. How would we find such a matrix? If Σ is diagonal, i.e. $\Sigma = \operatorname{diag}(\sigma_{11}, \ldots, \sigma_{dd})$, for $\sigma_{ii} \geq 0$, then the answer is easy, simply take:

$$\Sigma^{1/2} = \operatorname{diag}(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{dd}}).$$

In general, we need to use tools from linear algebra: Since Σ is symmetric it has an *eigendecomposition*:

$$\Sigma = U\Lambda U^T,$$

where U is an orthogonal matrix (i.e. $UU^T = U^TU = I$) containing the eigenvectors of Σ, and Λ is a diagonal matrix with the eigenvalues of Σ as its diagonal. Then the fact that Σ is positive semi-definite translates into the fact that $\Lambda_{ii} \geq 0$ for all $i = 1, \ldots, d$. But that means we may take $\Sigma^{1/2}$ to be

$$\Sigma^{1/2} = U\Lambda^{1/2} = U \operatorname{diag}(\sqrt{\lambda_{11}}, \ldots, \sqrt{\lambda_{dd}}),$$

so that

$$\Sigma^{1/2}(\Sigma^{1/2})^T = U\Lambda^{1/2}\Lambda^{1/2}U^T = U\Lambda U^T = \Sigma.$$

Continuing this, if Σ is also *positive definite* (which is stronger than positive semi-definite), then in fact $\Lambda_{ii} > 0$ for all $i = 1, \ldots, d$. In this case we can also easily find an inverse of Σ, as

$$\Sigma^{-1} = (U\Lambda U^T)^{-1} = U\Lambda^{-1}U^T,$$

has $\Sigma^{-1}\Sigma = \Sigma\Sigma^{-1} = I$ (again because $UU^T = U^TU = I$). Similarly, since also $\sqrt{\Lambda_{ii}} > 0$, we can do exactly the same thing to get the inverse of $\Sigma^{1/2}$ in this case, i.e.

$$\Sigma^{-1/2} := (U\Lambda^{1/2})^{-1} = \Lambda^{-1/2}U^T = \text{diag}(\lambda_{11}^{-1/2}, \dots, \lambda_{dd}^{-1/2})U^T,$$

so that $\Sigma^{-1/2}\Sigma^{1/2} = \Sigma^{1/2}\Sigma^{-1/2} = I$. Additionally it also holds that

$$(\Sigma^{-1/2})^T\Sigma^{-1/2} = U\Lambda^{-1}U^T = \Sigma^{-1},$$

which we will use in a minute. Both facts again hold since $U^TU = UU^T = I$. $\Diamond$

Finding the density of $\mathbf{Y}$ according to the above definition is then actually not very hard with the right tools: It derives from a very general and powerful transformation theorem, which is the multivariate analog of the formula presented in Remark 9. Using this it turns out that for B being a *positive definite* (and thus invertible) matrix and $\mathbf{b} \in \mathbb{R}^d$, $\mathbf{Y} = \mathbf{b} + B\mathbf{Z}$ has density:

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\det(B)} f_{\mathbf{Z}}\left(B^{-1}(\mathbf{y} - \mathbf{b})\right), \tag{59}$$

which indeed simplifies to (31) in the case $d = 1$. Additionally we know that by construction (i.e. independence)

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{d} f_{Z_i}(z_i) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\sum_{i=1}^{d} \frac{z_i^2}{2}\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\mathbf{z}^T\mathbf{z}}{2}\right).$$

Using (59) with $B = \Sigma^{1/2}$ we then get the density of a multivariate Gaussian random vector:

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{1}{(2\pi)^{d/2}\det(\Sigma^{1/2})} \exp\left(-\frac{(\Sigma^{-1/2}(\mathbf{y} - \mathbf{a}))^T\Sigma^{-1/2}(\mathbf{y} - \mathbf{a})}{2}\right) \\ &= \frac{1}{(2\pi)^{d/2}\det(\Sigma^{1/2})} \exp\left(-\frac{(\mathbf{y} - \mathbf{a})^T(\Sigma^{-1/2})^T\Sigma^{-1/2}(\mathbf{y} - \mathbf{a})}{2}\right) \\ &= \frac{1}{(2\pi)^{d/2}\det(\Sigma^{1/2})} \exp\left(-\frac{(\mathbf{y} - \mathbf{a})^T\Sigma^{-1}(\mathbf{y} - \mathbf{a})}{2}.\right) \end{aligned} \tag{60}$$

Note that this generalizes the case $d = 1$ in (36), as then $\Sigma = \det(\Sigma) = \sigma^2$. Also note that this makes sense since indeed:

$$\mathbb{E}[\mathbf{Y}] = \mathbb{E}[\mathbf{a} + \Sigma^{1/2}\mathbf{Z}] = \mathbf{a} + \Sigma^{1/2}\mathbb{E}[\mathbf{Z}] = \mathbf{a},$$

and

$$\mathbb{V}(\mathbf{Y}) = \mathbb{E}[(\mathbf{Y} - \mathbf{a})(\mathbf{Y} - \mathbf{a})^T] = \mathbb{E}[\Sigma^{1/2}\mathbf{Z}(\Sigma^{1/2}\mathbf{Z})^T] = \Sigma^{1/2}\mathbb{E}[\mathbf{Z}\mathbf{Z}](\Sigma^{1/2})^T = \Sigma^{1/2}(\Sigma^{1/2})^T = \Sigma.$$

Finally, from how we constructed the distribution (and from (60)) it can be seen that the elements in $\mathbf{Y}$ are independent iff $\Sigma = \mathbb{V}(Y)$ is diagonal. In other words if we consider d=2 and if (X_1, X_2) are jointly Gaussian:

$$X_1, X_2 \text{ independent} \iff \text{Cov}(X_1, X_2) = 0.$$

However it should be clear that this need not hold in general, and in fact for mixture distributions defined below it is not true.

There are numerous example of the multivariate Gaussian distribution in action. We will focus on a very practical and direct applicaton:

Example 18 (Gaussian Bayes Classifier). Assume we observe a discrete random variable $Y : \Omega \to \mathbb{R}$, with support $\text{supp}(Y) = \{1, \ldots, K\}$, for some $K \in \mathbb{N}$. In this example, these are the group labels of K different groups. Given a realization $Y(\omega) = y$ we define the conditional density as in Theorem 3.10, as

$$f_{Y=y}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\det(\Sigma_y)} \exp\left(-\frac{(\mathbf{x}-\mathbf{a}_y)^T\Sigma_y^{-1}(\mathbf{x}-\mathbf{a}_y)}{2}\right), \tag{61}$$

i.e. $\mathbf{X}|Y = y \sim N(\mathbf{a}_y, \Sigma_y)$ for $y \in \{1, \ldots, K\}$. Further assume that for each group y, we have an iid. sample $\mathbf{X}_{1,y}, \ldots, \mathbf{X}_{n_y,y}$ with $\mathbf{X}_{i,y} \sim N(\mathbf{a}_y, \Sigma_y)$. So for each y, there is a sample of n_y, giving a total of

$$n = \sum_{y=1}^{K} n_y$$

observations.

Usually the parameters $(\mathbf{a}_y, \Sigma_y)_{y=1}^K$ are not known. However they can easily be estimated using the standard approach for each group, that is for each y:

$$\begin{aligned}\bar{\mathbf{X}}_y &= \frac{1}{n_y}\sum_{i=1}^{n_y}\mathbf{X}_{i,y}\\ \mathbf{S}_y &= \frac{1}{n_y}\sum_{i=1}^{n_y}\left(\mathbf{X}_{i,y}-\bar{\mathbf{X}}_y\right)\left(\mathbf{X}_{i,y}-\bar{\mathbf{X}}_y\right)^T.\end{aligned}$$

Now for a new point (i.e. a so far unclassified realization) $\mathbf{X}(\omega) = \mathbf{x}$ the classifier uses the Bayes approach to find the most likely group y $\mathbf{x}$ belongs to. That is we want to maximize the *posterior probability*:

$$P_{\mathbf{x}}(Y = y) := \frac{f_{Y=y}(\mathbf{x})P(Y = y)}{f(\mathbf{x})}. \tag{62}$$

Now $f(\mathbf{x})$ is the marginal density of $\mathbf{X}$, which is given as:

$$f(\mathbf{x}) = \sum_{y=1}^{K} P(Y = y)f_{Y=y}(\mathbf{x}),$$

i.e. a finite mixture of Gaussian distributions. However we do not need this, since in the optimization with respect to y it is just a constant. That is:

$$\arg\max_y P_{\mathbf{x}}(Y = y) = \arg\max_y f_{Y=y}(\mathbf{x})P(Y = y).$$

Now with the estimated parameters, we know what $f_{Y=y}(\mathbf{x})$ is, but we do not know $P(Y = y)$ yet. However we can also easily estimate it from the data! The reason is that actually $P(Y =$

$y) = P(\{\omega : Y(\omega) = y\}) = \mathbb{E}[\mathbb{I}_{\{\omega:Y(\omega)=y\}}]$, so that for n realizations $(y_i)_{i=1}^n$ of Y, a consistent estimator, say p_y, is simply

$$p_y = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{y\}}(y_i) = \frac{n_y}{n}.$$

Thus finally we have all ingredients together and can find the label of the point $\mathbf{x}$, say $y_{\mathbf{x}}$ as:

$$y_{\mathbf{x}} = \underset{y\in\{1,\dots,K\}}{\arg\max}\, f_{Y=y}(\mathbf{x})p_y.$$

This simple classifier is easy to implement and works extremely well. The special case where $\Sigma_y = I$ for all y is called naive Bayes classifier. It is naive, because it assumes independence between the elements of $\mathbf{X}_{i,y}$, which is often not true. However as usual in statistics, a simple model can outperform a more complex one, especially if there is no abundance of data (complexity of the model vs estimation error). Indeed naive Bayes is often used and works remarkably well, even compared to more sophisticated algorithms! ◊

Finally, having defined the multivariate Gaussian, we can also define multivariate mixture densities:

Definition 3.15. Let $\mathbf{a}, \boldsymbol{\beta} \in \mathbb{R}^d$ and Σ be a positive definite matrix. Let furthermore $\mathbf{x} \mapsto \phi(\mathbf{x}; \mathbf{a}, \Sigma)$ be the density of a Gaussian random vector with mean $\mathbf{a}$ and variance Σ. The random vector $\mathbf{X}$ is said to be a multivariate continuous mean–variance mixture if its density $f_{\mathbf{X}}$ has representation

$$f_{\mathbf{X}}(\mathbf{x}) = \int_0^\infty \phi(\mathbf{x}; \mathbf{a} + \boldsymbol{\beta}g, g\Sigma) f_G(g) dg \tag{63}$$

for some continuous random variable $G : \Omega \to \mathbb{R}$ with nonnegative support. Equivalently $\mathbf{X}$ has stochastic representation

$$\mathbf{X} \overset{D}{=} \mathbf{a} + \boldsymbol{\beta}G + \sqrt{G}\Sigma^{1/2}\mathbf{Z}, \tag{64}$$

with $\mathbf{Z} \sim N(\mathbf{0}, I)$ and $G, \mathbf{Z}$ independent.

This highlights a convenient way of getting complex multivariate distributions. In fact, both remaining subsections of this script are not much more than a direct application of Definition 3.15.

3.4.3 Multivariate t-Distribution

Let $\nu > 0$, $\mathbf{a} \in \mathbb{R}^d$ and Σ a positive definite matrix, as before. We say $\mathbf{X} \sim t_\nu(\mathbf{a}, \Sigma)$ or $\mu_{\mathbf{X}} = t_\nu(\mathbf{a}, \Sigma)$ if $\mathbf{X}$ is a multivariate continuous mean–variance mixture with G following an *inverse gamma distribution* and $\boldsymbol{\beta} = \mathbf{0}$. In other words

$$\mathbf{X} \overset{D}{=} \mathbf{a} + \sqrt{G}\Sigma^{1/2}\mathbf{Z},$$

with G being "inverse gamma" with parameters $\alpha = \beta = \nu/2$. We did not talk about the inverse gamma distribution, though we can define it like this: For $\alpha, \beta > 0$, we say $Y \sim \mathrm{IGam}(\alpha, \beta)$, if

$Y^{-1} \sim \text{Gam}(\alpha, 1/\beta)$. With the tools we described for finding the density of a transformation of a random variable, one could then go on to find the density of this new distribution on $(0, +\infty)$. However since we derived everything for the gamma distribution, we instead use this definition to get another stochastic representation of $\mathbf{X}$, namely:

$$\mathbf{X} \overset{D}{=} \mathbf{a} + \tilde{G}^{-1/2}\Sigma^{1/2}\mathbf{Z},$$

with $\tilde{G} \sim \text{Gam}(\nu/2, 2/\nu)$. This is exactly the setting we presented in the univariate case, see Section 3.2.3. In fact for $d = 1$ we obtain a t_ν distribution which is augmented by a location a and scale term σ (just as we did in case of the Gaussian distribution). That is if $Y \sim t_\nu$, then $X = a + \sigma Y$.

Since $\mathbb{E}[\mathbf{X}]$ exists iff $\mathbb{E}[X_i]$ for $i = 1, \ldots, d$ exists, we can use Theorem 3.5 to see that for $\nu > 1$, $\mathbb{E}[\mathbf{X}]$ exists and from either one of the two stochastic representation:

$$\begin{aligned}
\mathbb{E}[\mathbf{X}] &= \mathbb{E}[\mathbf{a} + \sqrt{G}\Sigma^{1/2}\mathbf{Z}] \\
&= \mathbb{E}[\mathbb{E}[\mathbf{a} + \sqrt{G}\Sigma^{1/2}\mathbf{Z}|G]] \\
&= \mathbb{E}[\mathbf{a} + \sqrt{G}\Sigma^{1/2}\mathbb{E}[\mathbf{Z}|G]] \\
&= \mathbf{a}.
\end{aligned}$$

With a similar argument, we see that $\mathbb{V}(\mathbf{X})$ exists if $\nu > 2$ and in this case:[19]

$$\begin{aligned}
\mathbb{V}(\mathbf{X}) &= \mathbb{E}[(\mathbf{X} - \mathbf{a})(\mathbf{X} - \mathbf{a})^T] \\
&= \mathbb{E}[\mathbb{E}[\sqrt{G}\Sigma^{1/2}\mathbf{Z}(\sqrt{G}\Sigma^{1/2}\mathbf{Z})^T|G]] \\
&= \mathbb{E}[G\Sigma^{1/2}\mathbb{E}[\mathbf{Z}\mathbf{Z}^T|G]\Sigma^{1/2}] \\
&= \mathbb{E}[G]\Sigma \\
&= \frac{\nu}{\nu - 2}\Sigma,
\end{aligned}$$

since for G being $\text{IGam}(\nu/2, \nu/2)$, $\mathbb{E}[G] = \nu/(\nu - 2)$. To see the latter with what we derived, simply take again $\tilde{G} \sim \text{Gam}(\nu/2, 2/\nu)$, so that $\mathbb{E}[G] = \mathbb{E}[\tilde{G}^{-1}]$, and use the general expression of moments we derived for the Gamma distribution:

$$\mathbb{E}[Y^k] = \frac{\beta^k \Gamma(\alpha + k)}{\Gamma(\alpha)},$$

if $Y \sim \text{Gam}(\alpha, \beta)$. This was originally only for $k \in \mathbb{N}$. However for $k < 0$ one can show that as long as $\nu/2 > -k$, $\mathbb{E}[G^k]$ exists (this is of course tightly connected to the condition derived for the t distribution). The preceding expression then arises from the same calculations as we did in Section 3.2.4. In our case $k = -1$, and by assumption $\nu/2 > 1 = -k$, and thus we obtain

$$\mathbb{E}[\tilde{G}^k] = \mathbb{E}[\tilde{G}^{-1}] = \frac{(2/\nu)^{-1}\Gamma(\nu/2 - 1)}{\Gamma(\nu/2)} = \frac{\nu\Gamma(\nu/2 - 1)}{2(\nu/2 - 1)\Gamma(\nu/2 - 1)} = \frac{\nu}{2(\nu/2 - 1)} = \frac{\nu}{\nu - 2}.$$

[19] Remember: $\mathbb{E}[X_i^2]$ exists for all $i = 1, \ldots, d$, iff $\nu > 2$. This means the variance exists and one can show that it also implies that $\mathbb{E}[X_i X_j]$ exists for all $i \neq j$. But this immediately implies that the covariance between X_i and X_j exists as well.

Note that the univariate mixing variable induces dependencies between the elements of $\mathbf{X}$, even if $\Sigma = I$, i.e. even if Σ is diagonal! This is an important feature of the multivariate t (and of such mixture distributions in general). It is also a nice case for which we see that $\mathrm{Cov}(X_1, X_2) = 0$ does not at all imply that X_1, X_2 are also independent.[20]

Finally solving the integral

$$f(\mathbf{x}) = \int_0^\infty \phi(\mathbf{x}; \mathbf{a}, g\Sigma) f_G(g) dg$$

one obtains the joint density of $\mathbf{X}$ as:

$$f(\mathbf{x}) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}\det(\Sigma^{1/2})} \left(1 + \frac{1}{\nu}(\mathbf{x}-\mathbf{a})^T \Sigma^{-1} (\mathbf{x}-\mathbf{a})\right)^{-(\nu+d)}. \tag{65}$$

We will prove an important and very useful fact about this distribution once we get to characteristic functions. This is the fact that if $\mathbf{X} \sim t_\nu(\mathbf{a}, \Sigma)$ and $\mathbf{w} \in \mathbb{R}^d$, then $\mathbf{w}^T\mathbf{X} \sim t_\nu(\mathbf{w}^T\mathbf{a}, \mathbf{w}^T\Sigma\mathbf{w})$.

3.4.4 Multivariate Variance–Gamma Distribution

Let $\lambda > 0$, $\alpha > 0$, $\boldsymbol{\beta} \in \mathbb{R}^d$ such that $\sqrt{\boldsymbol{\beta}^T\boldsymbol{\beta}} \in (-\alpha, \alpha)$, $\mathbf{a} \in \mathbb{R}^d$ and Σ a positive definite matrix. We denote $\mathbf{X} \sim \mathrm{MVG}(\lambda, \alpha, \boldsymbol{\beta}, \mathbf{a}, \Sigma)$ or $\mu_{\mathbf{X}} = \mathrm{MVG}(\lambda, \alpha, \boldsymbol{\beta}, \mathbf{a}, \Sigma)$ if $\mathbf{X}$ is a multivariate continuous mean–variance mixture with G following a Gamma distribution. In other words

$$\mathbf{X} \overset{D}{=} \mathbf{a} + G\boldsymbol{\beta} + \sqrt{G}\Sigma^{1/2}\mathbf{Z},$$

with $G \sim \mathrm{Gam}\left(\lambda, 2/(\alpha^2 - \boldsymbol{\beta}^T\boldsymbol{\beta})\right)$.

Since all positive moments of the univariate VG distribution exist, we also know that $\mathbb{E}[\mathbf{X}]$ and $\mathbb{V}(\mathbf{X})$ exists and we can find them as in Section 3.4.3. That is, for the expected value:

$$\begin{aligned}
\mathbb{E}[\mathbf{X}] &= \mathbb{E}[\mathbf{a} + \boldsymbol{\beta}G + \sqrt{G}\Sigma^{1/2}\mathbf{Z}] \\
&= \mathbb{E}[\mathbb{E}[\mathbf{a} + \boldsymbol{\beta}G + \sqrt{G}\Sigma^{1/2}\mathbf{Z}|G]] \\
&= \mathbb{E}[\mathbf{a} + \boldsymbol{\beta}G + \sqrt{G}\Sigma^{1/2}\mathbb{E}[\mathbf{Z}|G]] \\
&= \mathbf{a} + \boldsymbol{\beta}\mathbb{E}[G] \\
&= \mathbf{a} + \frac{2\lambda\boldsymbol{\beta}}{\alpha^2 - \boldsymbol{\beta}^T\boldsymbol{\beta}}.
\end{aligned}$$

Similarly one could calculate

$$\mathbb{V}(\mathbf{X}) = \mathbb{E}\left[\left(\mathbf{X} - \mathbf{a} - \frac{2\lambda\boldsymbol{\beta}}{\alpha^2 - \boldsymbol{\beta}^T\boldsymbol{\beta}}\right)\left(\mathbf{X} - \mathbf{a} - \frac{2\lambda\boldsymbol{\beta}}{\alpha^2 - \boldsymbol{\beta}^T\boldsymbol{\beta}}\right)^T\right].$$

[20] Though the other way around is true, since (if all involved moments exist)

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] = 0.$$

Note that for $\boldsymbol{\beta} = \mathbf{0}$, the exact same steps as in Section 3.4.3 give:

$$\begin{aligned}\mathbb{V}(\mathbf{X}) &= \mathbb{E}[(\mathbf{X}-\mathbf{a})(\mathbf{X}-\mathbf{a})^T] \\ &= \mathbb{E}[G]\Sigma \\ &= \lambda\frac{2}{\alpha^2}\Sigma.\end{aligned}$$

Again solving the integral

$$f(\mathbf{x}) = \int_0^\infty \phi(\mathbf{x}; \mathbf{a} + g\boldsymbol{\beta}, g\Sigma) f_G(g) dg$$

one obtains the joint density of $\mathbf{X}$ as:

$$\begin{aligned}f(\mathbf{x}) =& \frac{2\left(\frac{\alpha^2-\boldsymbol{\beta}^T\boldsymbol{\beta}}{2}\right)^\lambda}{(2\pi)^{d/2}\det(\Sigma^{1/2})\Gamma(\lambda)}\left(\frac{(\mathbf{x}-\mathbf{a})^T\Sigma^{-1}(\mathbf{x}-\mathbf{a})}{\alpha^2}\right)^{(\lambda-d/2)/2} \\ & K_{\lambda-d/2}\left(\alpha\sqrt{(\mathbf{x}-\mathbf{a})^T\Sigma^{-1}(\mathbf{x}-\mathbf{a})}\right)\exp(\boldsymbol{\beta}^T(\mathbf{x}-\mathbf{a})),\end{aligned} \tag{66}$$

which is a very natural generalization of the density in (48).

4 Selected Topics

4.1 Integration with respect to a Probability Measure (optional)

We construct the integral with respect to a probability measure (i.e. the Lebesgue integral) as in Jacod and Protter (2004, Chapter 9): Let as usual $(\Omega, \mathcal{A}, P)$ be a probability space. We first define expectation for a simple random variable or function:

Definition 4.1. A r.v. $X : \Omega \to \mathbb{R}$ is *simple* if it takes on only a finite number of values in $\mathbb{R}$.

A simple r.v. X can always be written in the form:

$$X(\omega) = \sum_{i=1}^n a_i \mathbb{I}_{A_i}(\omega) \quad \forall \omega \in \Omega$$

with $a_i \in \mathbb{R}$ and $A_i \in \mathcal{A}$, $(A_i)_{i=1}^n$ a disjoint sequence. Indeed since X is simple it only takes on say n real values $(a_i)_{i=1}^n$. If we take $A_i = \{\omega \in \Omega : X(\omega) = a_i\} = X^{-1}(\{a_i\})$, then, since $\{a_i\} \in \mathcal{B}(\mathbb{R})$ and X is measurable, $A_i \in \mathcal{A}$. It is then easy to check that the above equality holds for each ω. Note however that there are many different such sum representations for X. As an example, consider $X = a$, i.e. X takes on only one value. Then for some $A \in \mathcal{A}$:

$$X = a\mathbb{I}_\Omega \text{ and } X = a\mathbb{I}_A + a\mathbb{I}_{A^c},$$

are valid representations. Also note that we can without loss of generality always assume that $\bigcup_{i=1}^n B_i = \Omega$ (in addition to $(B_i)_{i=1}^n$ being disjoint). If this is not the case, i.e.

$$\bigcup_{i=1}^n B_i \subset \Omega,$$

with strict inclusion, we can just take $b_{n+1} = 0$ and $B_{n+1} = \Omega \setminus \bigcup_{i=1}^{n} B_i$ and have

$$X = \sum_{i=1}^{n+1} b_i \mathbb{I}_{B_i}.$$

We now define the integral with respect to P over this simple function:

Definition 4.2. The integral of a simple r.v. X with respect to P is given as

$$\int_{\Omega} X(\omega) dP(\omega) = \sum_{i=1}^{n} a_i P(A_i).$$

Thus P enters in this integral definition as the measure of the sets A_i. Let us now check that Definition 4.2 expression makes sense; That is if we have two sum representations of X, say

$$X = \sum_{i=1}^{n} a_i \mathbb{I}_{A_i} \text{ and } X = \sum_{j=1}^{m} b_j \mathbb{I}_{B_j},$$

again with $a_i, b_i \in \mathbb{R}$ and $A_i, B_i \in \mathcal{A}$, then we would like the resulting integral to be the same (otherwise it depends on the representation of X, and is thus not "stable")! Using what we have said above, we can assume $(A_i)_{i=1}^n$ and $(B_i)_{i=1}^n$ disjoint to have Ω as their union (in other words they partition Ω). Then, whenever we have that $A_l \cap B_r \neq \emptyset$ for some $l \in \{1, \ldots, n\}$, $r \in \{1, \ldots, m\}$, it must be that $a_l = b_r$. Indeed, if $\omega \in A_l \cap B_r$,

$$a_l = \sum_{i=1}^{n} a_i \mathbb{I}_{A_i}(\omega) = X(\omega) = \sum_{j=1}^{m} b_j \mathbb{I}_{B_j}(\omega) = b_r.$$

Additionally for any i, $A_i = \bigcup_{j=1}^{m} (B_j \cap A_i)$, and symmetrically for any j, $B_j = \bigcup_{i=1}^{n} (B_j \cap A_i)$. Thus

$$\begin{aligned}
\sum_{i=1}^{n} a_i P(A_i) &= \sum_{i=1}^{n} a_i P\left(\bigcup_{j=1}^{m} (B_j \cap A_i) \right) \\
&= \sum_{i=1}^{n} a_i \sum_{j=1}^{m} P\left(B_j \cap A_i\right) \\
&= \sum_{j=1}^{m} \sum_{i=1}^{n} a_i P\left(B_j \cap A_i\right)
\end{aligned}$$

Now either $P\left(B_j \cap A_i\right) = P(\emptyset) = 0$, in which case we can set $a_i = b_j$ without changing anything or $P\left(B_j \cap A_i\right) > 0$ and in this case $a_i = b_j$ anyway, so

$$\begin{aligned}
\sum_{i=1}^{n} a_i P(A_i) &= \sum_{j=1}^{m} b_j \sum_{i=1}^{n} P\left(B_j \cap A_i\right) \\
&= \sum_{j=1}^{m} b_j P\left(B_j\right).
\end{aligned}$$

So indeed the integral has the same value, no matter what kind of characterization we choose.

Since it does not matter for the integral what kind of sum representation we choose for X, we can play around with this to show some interesting things: For example if we have two simple r.v. X, Y with

$$X = \sum_{i=1}^{n} a_i \mathbb{I}_{A_i} \text{ and } Y = \sum_{j=1}^{m} b_j \mathbb{I}_{B_j},$$

simply choose a new representation for X and Y, by using the collection of sets

$$C_{i,j} = A_i \cap B_j,\ i = 1, \ldots, n,\ j = 1, \ldots, m.$$

This is still a finite collection and now:

$$X = \sum_{i,j} a_{i,j} \mathbb{I}_{C_{i,j}} \text{ and } Y = \sum_{i,j} b_{i,j} \mathbb{I}_{C_{i,j}},$$

i.e. X, Y still have different values, but are now represented by the same sets.[21] But then

$$\begin{aligned}\int_\Omega X(\omega) + Y(\omega) dP(\omega) &= \sum_{i,j} (a_{i,j} + b_{i,j}) P(C_{i,j}) \\ &= \sum_{i,j} a_{i,j} P(C_{i,j}) + \sum_{i,j} b_{i,j} P(C_{i,j}) \\ &= \int_\Omega X(\omega) dP(\omega) + \int_\Omega Y(\omega) dP(\omega).\end{aligned}$$

Also it quite obviously always holds that for any number $s \in \mathbb{R}$:

$$\int_\Omega sX(\omega) dP(\omega) = s \int_\Omega X(\omega) dP(\omega).$$

In summary, the integral we defined is *linear* for simple functions, that is

$$\int_\Omega sX(\omega) + tY(\omega) dP(\omega) = s \int_\Omega X(\omega) dP(\omega) + t \int_\Omega Y(\omega) dP(\omega),$$

for X, Y simple r.v. and any $s, t \in \mathbb{R}$. Finally if

$$X(\omega) \leq Y(\omega) \text{ for all } \omega \in \Omega,$$

then in fact (if we take again a finite sum representation with the same sets), $a_i \leq b_i$ for all $i = 1, \ldots, n$ and

$$\int_\Omega X(\omega) dP(\omega) \leq \int_\Omega Y(\omega) dP(\omega),$$

so the integral is monotone.

Next we define the integral for *nonnegative* random variables X: If $X : \Omega \to \mathbb{R}$ is has $X(\omega) \geq 0$ for all $\omega \in \Omega$, we define:

$$\int_\Omega X(\omega) dP(\omega) = \sup \left\{ \int_\Omega Y(\omega) dP(\omega) : Y \text{ is a simple r.v. with } 0 \leq Y \leq X \right\}. \tag{67}$$

[21] Note that again the sets $(C_{i,j})_{i,j}$ are disjoint because the $(A_i)_i = 1^n$ and $(B_j)_{j=1}^n$ are. In general, if A_1, A_2 and B_1, B_2 are disjoint, then so are $A_1 \cap B_1$ and $A_2 \cap B_2$.

So we take the supremum over the set of integral values (which are simply values in $\mathbb{R}$) over simple functions smaller than X. Since the integral over the simple function $Y = 0$ is always part of the above set, it is never empty and the supremum well-defined![22] This is the reason that, while the supremum might be $+\infty$, it and thus $\int_\Omega X(\omega)dP(\omega)$ is always defined. From this we can also immediately define the integral for general random variables (or equivalently $\mathcal{A}/\mathcal{B}(\mathbb{R})$ measurable functions). In fact we have seen this definition already in Section 3.2:

Definition 4.3. Let $X : \Omega \to \mathbb{R}$ be a r.v. and define (pointwise for each $\omega \in \Omega$)

$$X^+ = \max(0, X)$$
$$X^- = -\min(0, X).$$

Then $X^+, X^- \geq 0$ are $\mathcal{A}/\mathcal{B}(\mathbb{R})$ measurable as well and we say the integral of X with respect to P exists if

$$\int_\Omega X^+(\omega)dP(\omega) = \int_\Omega \max(0, X(\omega))dP(\omega) < \infty$$
$$\int_\Omega X^-(\omega)dP(\omega) = \int_\Omega -\min(0, X(\omega))dP(\omega) < \infty.$$

In this case, we define

$$\int_\Omega X(\omega)dP(\omega) = \int_\Omega X^+(\omega)dP(\omega) - \int_\Omega X^-(\omega)dP(\omega). \tag{68}$$

As a remark; it would actually be enough to assume that just one of the two conditions

$$\int_\Omega X^+(\omega)dP(\omega) = \int_\Omega \max(0, X(\omega))dP(\omega) < \infty$$
$$\int_\Omega X^-(\omega)dP(\omega) = \int_\Omega -\min(0, X(\omega))dP(\omega) < \infty.$$

is true. So one of the two could be infinity and (68) would still be valid. We follow a different convention however and dictate that *both* need to be finite. If it exists then the *expectation* of $X : \Omega \to \mathbb{R}$, is simply the integral over X with respect to P:

$$\mathbb{E}[X] := \int_\Omega X(\omega)dP(\omega) = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

We can then state the beautiful properties of this integral notion (doing this using the expectation notation to make things shorter). First we need an important result:

Theorem 4.1. *If $X : \Omega \to \mathbb{R}$ is a nonnegative random variable, then there exists a monotone sequence of simple random variables $(X_n)_{n\in\mathbb{N}}$ with*

$$X_n(\omega) \uparrow X(\omega) \quad \forall \omega \in \Omega.$$

Furthermore, for any *increasing sequence of nonnegative simple r.v. $(X_n)_n$ with $X_n \uparrow X$ pointwise it is true that*

$$\mathbb{E}[X_n] \uparrow \mathbb{E}[X].$$

[22] Recall that any nonempty subset of $\mathbb{R}$ has a supremum, which is either itself part $\mathbb{R}$ or $+\infty$.

We will not prove this, but just say that a candidate for such an X_n is given in Jacod and Protter (2004, p. 53) as:

$$X_n(\omega) = \begin{cases} k2^{-n} & \text{if } k2^{-n} \leq X(\omega) < (k+1)2^{-n} \text{ for a } k,\ 0 \leq k \leq n2^n - 1 \\ n & \text{if } X(\omega) \geq n \end{cases}.$$

(there either is exactly one $k \leq n2^n - 1$ such that the first condition is true, or $X(\omega) \geq n$). One can then show that for all $\omega \in \Omega$: $X_n(\omega) \leq X_{n+1}(\omega)$ for all n, that $X_n(\omega) \uparrow X(\omega)$ and that $\mathbb{E}[X_n] \uparrow \mathbb{E}[X]$, as $n \to \infty$. In fact, as the theorem states, for any sequence of simple functions with these properties $\mathbb{E}[X_n] \uparrow \mathbb{E}[X]$ holds. However much more interesting then the proof is the result itself! It allows to extend results for simple random variables to general ones. This is what it is sometimes referred to as "measure theoretic induction":

Step 1: Show that something holds for the simplest random variable $X = \mathbb{I}_A$, for some $A \in \mathcal{A}$.

Step 2: Show that it holds for linear combinations of indicator functions,

$$\sum_{i=1}^{n} a_i \mathbb{I}_{A_i},\ a_i \in \mathbb{R},\ A_i \in \mathcal{A},$$

in other words that it holds for all simple functions.

Step 3: Show that it holds for all nonnegative r.v. by utilizing Theorem 4.1.

Step 4: Show that it holds for any r.v. X with existing expectation, by taking $X = X^+ - X^-$ and using the fact that X^+, X^- are nonnegative.

This technique will in part be used in the upcoming Theorem 4.2 and especially in Theorem 4.5.

Theorem 4.2 (Theorem 9.1 in Jacod and Protter (2004))**.** *Let throughout X and Y be two random variables with existing expectation/integral and $a \in \mathbb{R}$ be arbitrary. Then*

(a) $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ *and* $\mathbb{E}[aX] = a\mathbb{E}[X]$*, i.e. expectation is a linear map. Furthermore it is monotone, that is if $Y \leq X$, then $\mathbb{E}[Y] \leq \mathbb{E}[X]$.*

(b) The expectation of X exists iff $\mathbb{E}[|X|]$ exists and $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$

(c) If $X = Y$ P-as., i.e. if there exists a set $A \in \mathcal{A}$ with $P(A) = 1$ and such that for all $\omega \in A$, $X(\omega) = Y(\omega)$, then $\mathbb{E}[X] = \mathbb{E}[Y]$.

Proof. Proof missing

■

We can even state the beautiful convergence theorems now, that make the Lebesgue integral so powerful:

Theorem 4.3. *If $X_n : \Omega \to \mathbb{R}$, $n \in \mathbb{N}$ are a collection of r.v. with $(X_n)_n$*

(i) nonnegative (i.e. $X_n(\omega) \geq 0$ for all ω and n)

(ii) increasing (i.e. $X_n(\omega) \leq X_{n+1}(\omega)$ for all ω and n),

then

$$\lim_n \int_\Omega X_n(\omega) dP(\omega) = \int_\Omega \lim_n X_n(\omega) dP(\omega).$$

This is true even if $\mathbb{E}[X] = +\infty$.

Theorem 4.4. *Let $X_n : \Omega \to \mathbb{R}$, $n \in \mathbb{N}$, be a sequence of r.v. with $(X_n)_n$ converging pointwise to some r.v. X, i.e. $\lim_n X_n(\omega) = X(\omega)$ for all $\omega \in \Omega$. Let furthermore $Y \geq 0$ be a r.v. with existing expectation, i.e. $\mathbb{E}[Y] < +\infty$ and*

$$|X_n| \leq Y \text{for all } n \in \mathbb{N}.$$

Then

$$\lim_n \int_\Omega X_n(\omega) dP(\omega) = \int_\Omega X(\omega) dP(\omega).$$

Remark 13. One can show that if $\lim_n X_n(\omega) = X(\omega)$ for all $\omega \in \Omega$ the measurability of X is automatically guaranteed. So in both the monotone and the dominated convergence theorem, measurability of $\lim_n X_n$ is not an issue. ◊

Now one interesting question that remains is the following: Say we are interested in the expected value of a function $g : \mathbb{R} \to \mathbb{R}$ of X. Then $g \circ X : \Omega \to \mathbb{R}$ is a r.v. and we defined its expectation to be

$$\mathbb{E}[g(X)] = \mathbb{E}[g \circ X] = \int_\Omega g \circ X(\omega) dP(\omega).$$

Also recall that we defined the probability on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ induced by X, as $\mu_X(B) = P(X^{-1}(B))$ for all $B \in \mathcal{B}(\mathbb{R})$. One of the big advantages of this change in measure, was that we could forget about Ω and focus on the nicer space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. So the looming question is: Can we now make the above integration over Ω into one over $\mathbb{R}$? It turns out that indeed we can:

Theorem 4.5. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(S, \mathcal{S})$ be a measurable space, $X : \Omega \to S$ be $\mathcal{A}/\mathcal{S}$ measurable and $g : S \to \mathbb{R}$, be $\mathcal{S}/\mathcal{B}(\mathbb{R})$ measurable. Further, let as usual μ_X be the distribution of X on S. Then*

$$\mathbb{E}[g(X)] = \int_\Omega g \circ X(\omega) dP(\omega) \text{ exists} \iff \int_S g(x) d\mu_X(x) \text{ exists}$$

and in this case

$$\mathbb{E}[g(X)] = \int_\Omega g \circ X(\omega) dP(\omega) = \int_S g(x) d\mu_X(x). \tag{69}$$

Proof. Proof missing ■

If one takes $(S, \mathcal{S}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and substitutes $\mathbf{X} : \Omega \to \mathbb{R}^d$ for X in the above theorem, it follows that if $\mathbb{E}[g(\mathbf{X})]$ exists,

$$\mathbb{E}[g(\mathbf{X})] = \int_\Omega g(\mathbf{X}(\omega)) dP(\omega) = \int_{\mathbb{R}^d} g(\mathbf{x}) d\mu_{\mathbf{X}}(\mathbf{x}),$$

which is Equation (70) in the next section. Finally we can also state the Fubini Theorem in full generality, for any product measure:

Theorem 4.6 (Theorem 1.7.2 in Durrett (2010))**.** *Let $(S_1, \mathcal{S}_1, P_1)$, $(S_2, \mathcal{S}_2, P_2)$ be arbitrary probabilty spaces and $P_1 \times P_2$ the product probability measure as in Theorem 3.11. Further, let $X : S_1 \times S_2 \to \mathbb{R}$ be a $\mathcal{S}_1 \times \mathcal{S}_2 / \mathcal{B}(\mathbb{R})$ measurable function with $X \geq 0$ or $\int_{S_1 \times S_2} |X| \; dP_1 \times P_2 < \infty$. Then*

$$\int_{S_1 \times S_2} X \; dP_1 \times P_2 = \int_{S_2} \int_{S_1} X \; dP_1 dP_2 = \int_{S_1} \int_{S_2} X \; dP_2 dP_1.$$

In fact all that was said above remains valid in the same way if we consider the Lebesgue measure λ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ instead of a probability measure (the main difference between a probability measure and the measure λ being that the latter has $\lambda(\mathbb{R}) = \infty$). So the Lebesgue integral with the Lebesgue measure we used throughout the lecture is also constructed in exactly this way and has the same properties with essentially the same proofs. The key thing here is that though $\lambda(\mathbb{R}) = \infty$, the Lebesgue measure is still "σ-finite", a concept we will however not discuss here.

4.2 Characteristic Functions

We note that the Lebesgue integral we used so far can in fact not only be defined for the Lebesgue measure λ, but also for any other (probability measure). In this spirit we will from now on sometimes use the notation

$$\mathbb{E}[g(\mathbf{X})] \overset{(1)}{=} \int_\Omega g(\mathbf{X}(\omega)) dP(\omega) \overset{(2)}{=} \int_{\mathbb{R}^d} g(\mathbf{x}) d\mu_{\mathbf{X}}(\mathbf{x}). \tag{70}$$

In fact (1) is how expectation is actually defined, while (2) would need to be proven. This change of notation should stress the generality of the definitions to come. However in the case of continuous random vectors, it "simplifies" back to what we know:

$$\mathbb{E}[g(\mathbf{X})] = \int_{\mathbb{R}^d} g(\mathbf{x}) d\mu_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbb{R}^d} g(\mathbf{x}) f(\mathbf{x}) d\lambda(\mathbf{x})$$

Indeed, one is justified to write $f(\mathbf{x}) = \frac{d\mu_{\mathbf{X}}(\mathbf{x})}{d\lambda(\mathbf{x})}$, i.e. we can in a certain sense see the density f as the derivative of the probability measure $\mu_{\mathbf{X}}$ with respect to the Lebesgue measure λ. Once again expectation exists if both

$$\mathbb{E}[g(\mathbf{X})^+] = \int_\Omega \max(g(\mathbf{X}(\omega)), 0) dP(\omega) = \int_{\mathbb{R}^d} \max(g(\mathbf{x}), 0) d\mu_{\mathbf{X}}(\mathbf{x}) < \infty$$
$$\mathbb{E}[g(\mathbf{X})^-] = \int_\Omega - \min(g(\mathbf{X}(\omega)), 0) dP(\omega) = \int_{\mathbb{R}^d} - \min(g(\mathbf{x}), 0) d\mu_{\mathbf{X}}(\mathbf{x}) < \infty$$

And, as used implicitly earlier, these two integrals are finite iff

$$\mathbb{E}[|g(\mathbf{X})|] = \mathbb{E}[g(\mathbf{X})^+] + \mathbb{E}[g(\mathbf{X})^-] < +\infty,$$

which is what we usually check.

Now importantly we can define any type of Lebesgue integral also with respect to complex numbers, by just considering the real and imaginary part separately. So if $g : \mathbb{R}^d \to \mathbb{C}$, then for all $\mathbf{x} \in \mathbb{R}^d$

$$g(\mathbf{x}) = \Re(g(\mathbf{x})) + \iota \Im(g(\mathbf{x})),$$

with $\Re(g(\mathbf{x})), \Im(g(\mathbf{x})) \in \mathbb{R}$ the real and imaginary parts respectively. Then the integral of g exists iff the ones of $\Re(g(\mathbf{x})), \Im(g(\mathbf{x}))$ exist and

$$\mathbb{E}[g(\mathbf{X})] = \int_{\mathbb{R}^d} g(\mathbf{x}) d\mu_{\mathbf{X}}(\mathbf{x}) := \int_{\mathbb{R}^d} \Re(g(\mathbf{x})) d\mu_{\mathbf{X}}(\mathbf{x}) + \iota \int_{\mathbb{R}^d} \Im(g(\mathbf{x})) d\mu_{\mathbf{X}}(\mathbf{x}).$$

Note that this means in particular that the complex number $\mathbb{E}[g(\mathbf{X})]$ has $\Re(\mathbb{E}[g(\mathbf{X})]) = \mathbb{E}[\Re(g(\mathbf{X}))]$ and $\Im(\mathbb{E}[g(\mathbf{X})]) = \mathbb{E}[\Im(g(\mathbf{X}))]$.

With this definition almost all properties of the integral generalize easily to the case of complex valued functions. For instance, assume that $\mathbf{X}$ has a joint density $f_{\mathbf{X}}$ and that

$$\mathbb{E}[|g(\mathbf{X})|] = \int_{\mathbb{R}^d} |g(\mathbf{x})| f(\mathbf{x}) d\mathbf{x} < +\infty.$$

Then in particular $\mathbb{E}[|\Re(g(\mathbf{X}))|] \leq \mathbb{E}[|g(\mathbf{X})|] < +\infty$ and in the same way $\mathbb{E}[|\Im(g(\mathbf{X}))|] < +\infty$, and we can use the Fubini theorem on the real and imaginary part separately to see that

$$\int_{\mathbb{R}^d} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}} \dots \int_{\mathbb{R}} g(\mathbf{x}) f(\mathbf{x}) dx_1 \dots dx_d$$

This in fact holds true even if $\mathbf{X}$ does not have a density with the general integral notation in (70), as long as the measure of $\mathbf{X}$ is the product measure. However we only stated the Fubini Theorem in terms of the Lebesgue integral, so we stick to this case. The preceding is used in Theorem 4.9.

Now we are ready for the main definition:

Definition 4.4 (Definition 13.2 in Jacod and Protter (2004))**.** Let $\mathbf{X} : \Omega \to \mathbb{R}^d$ be a random vector. The characteristic function of $\mathbf{X}$ is the function $\phi_{\mathbf{X}} : \mathbb{R}^d \to \mathbb{C}$ defined as

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[\exp(\iota \mathbf{t}^T \mathbf{X})] = \int_{\mathbb{R}^d} \exp(\iota \mathbf{t}^T \mathbf{x}) d\mu_{\mathbf{X}}(\mathbf{x}). \tag{71}$$

We will now look at the properties of this new function and then give many examples:

Theorem 4.7. *The integral in (71) (and thus the cf)* always *exists and* $|\phi_{\mathbf{X}}(\mathbf{t})| \leq 1$ *for all* $\mathbf{t} \in \mathbb{R}^d$*, with* $\phi(\mathbf{0}) = 1$*.*

Proof. First note that for the integral in (71) to make sense, we need for any $\mathbf{t}$, the functions $\mathbf{x} \mapsto \Re(\exp(\iota \mathbf{t}^T \mathbf{x}))$ and $\mathbf{x} \mapsto \Im(\exp(\iota \mathbf{t}^T \mathbf{x}))$ to be $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R})$ measurable. But from Appendix A,

$$\Re(\exp(\iota \mathbf{t}^T \mathbf{x})) = \cos(\mathbf{t}^T \mathbf{x}), \quad \Im(\exp(\iota \mathbf{t}^T \mathbf{x})) = \sin(\mathbf{t}^T \mathbf{x}),$$

which are continuous in $\mathbf{x}$ and in particular $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R})$ measurable. Now, note that for any $\mathbf{t} \in \mathbb{R}^d$:

$$\begin{aligned}
\mathbb{E}[|\exp(\iota \mathbf{t}^T \mathbf{X})|] &= \int_{\mathbb{R}^d} |\exp(\iota \mathbf{t}^T \mathbf{x})| d\mu_{\mathbf{X}}(\mathbf{x}) \\
&= \int_{\mathbb{R}^d} 1 d\mu_{\mathbf{X}}(\mathbf{x}) \\
&= 1 < \infty,
\end{aligned}$$

since $|\exp(\iota y)| = 1$ for all $y \in \mathbb{R}$. However as mentioned above, then

$$\begin{aligned}
\mathbb{E}[|\Re(\exp(\iota \mathbf{t}^T \mathbf{X}))|] &\leq \mathbb{E}[|\exp(\iota \mathbf{t}^T \mathbf{X})|] = 1 < +\infty \\
\mathbb{E}[|\Im(\exp(\iota \mathbf{t}^T \mathbf{X}))|] &\leq \mathbb{E}[|\exp(\iota \mathbf{t}^T \mathbf{X})|] = 1 < +\infty,
\end{aligned}$$

since for any $z \in \mathbb{C}$, $|\Re(z)| \leq |z|$ and $|\Im(z)| \leq |z|$. So indeed $\mathbb{E}[\exp(\iota \mathbf{t}^T \mathbf{X})]$ exists for any $\mathbf{t} \in \mathbb{R}^d$! Then using the fact that (i) $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$ (Jensen's inequality) for any random variable X and (ii) for any $z \in \mathbb{C}$: $|z|^2 = \Re(z)^2 + \Im(z)^2$, we have:

$$\begin{aligned}
|\phi_{\mathbf{X}}(\mathbf{t})|^2 &= |\mathbb{E}[\exp(\iota \mathbf{t}^T \mathbf{X})]|^2 \\
&= \Re(\mathbb{E}[\exp(\iota \mathbf{t}^T \mathbf{X})])^2 + \Im(\mathbb{E}[\exp(\iota \mathbf{t}^T \mathbf{X})])^2 \\
&= \mathbb{E}[\Re(\exp(\iota \mathbf{t}^T \mathbf{X}))]^2 + \mathbb{E}[\Im(\exp(\iota \mathbf{t}^T \mathbf{X}))]^2 \\
&\leq \mathbb{E}[\Re(\exp(\iota \mathbf{t}^T \mathbf{X}))^2] + \mathbb{E}[\Im(\exp(\iota \mathbf{t}^T \mathbf{X}))^2] \\
&= \mathbb{E}[|\exp(\iota \mathbf{t}^T \mathbf{X})|^2] \\
&= \mathbb{E}[1^2] = 1.
\end{aligned}$$

So indeed also $|\phi_{\mathbf{X}}(\mathbf{t})| \leq 1$ for all $\mathbf{t} \in \mathbb{R}^d$. Furthermore, since $\exp(0) = 1$ it holds that $\phi_{\mathbf{X}}(\mathbf{0}) = \mathbb{E}[\exp(0)] = 1$. ■

One can also show that the function $\phi_{\mathbf{X}}(\mathbf{t})$ is *uniformly continuous* in $\mathbf{t}$ (and thus in particular continuous). However we will not need this fact here. What makes the characteristic function so powerful for our purposes is the following theorem, which we will however not proof here:

Theorem 4.8 (Adaptation of Theorem 14.1 in Jacod and Protter (2004))**.** *The cf completely characterizes $\mu_{\mathbf{X}}$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. In particular $\mu_{\mathbf{X}} = \mu_{\mathbf{Y}}$ iff $\phi_{\mathbf{X}}(\mathbf{t}) = \phi_{\mathbf{Y}}(\mathbf{t})$ for all $\mathbf{t} \in \mathbb{R}^d$.*

This has an immediate important corollary:

Theorem 4.9. *The random variables $X_1, \ldots, X_d$ are independent iff $\mathbf{X} = (X_1, \ldots, X_d)^T$ has cf*

$$\phi_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^{d} \phi_{X_i}(t_i). \tag{72}$$

Proof. We will show this in the case of $\mathbf{X}$ possessing a joint density f. However literally the same proof can also be used to show this theorem for the general case (but for that it would have been necessary to state the Fubini theorem in a more general form, for any measure instead of just the Lebesgue measure):

First assume $X_1, \ldots, X_d$ are independent, so that

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{d} f_{X_i}(x_i).$$

We start by considering the cf of $\mathbf{X}$:

$$\begin{aligned}
\phi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}[\exp(\iota \mathbf{t}^T \mathbf{X})] \\
&= \mathbb{E}\left[\exp\left(\iota\left(\sum_{i=1}^{d} t_i X_i\right)\right)\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{d} \exp\left(\iota t_i X_i\right)\right]
\end{aligned}$$

So far we did not use the independence assumption at all. Now we account for independence and can once again use Fubini's theorem on the real and imaginary parts of $\exp(\iota \mathbf{t}^T \mathbf{x})$ separately to split up the above expectation:

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{d} \exp\left(\iota t_i x_i\right)\right] &= \int_{\mathbb{R}^d} \exp(\iota \mathbf{t}^T \mathbf{X}) \prod_{i=1}^{d} f_{X_i}(x_i) d\mathbf{x} \\
&= \int_{\mathbb{R}^d} \Re(\exp(\iota \mathbf{t}^T \mathbf{X})) \prod_{i=1}^{d} f_{X_i}(x_i) d\mathbf{x} + \iota \int_{\mathbb{R}^d} \Im(\exp(\iota \mathbf{t}^T \mathbf{X})) \prod_{i=1}^{d} f_{X_i}(x_i) d\mathbf{x} \\
&= \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} \cos(\mathbf{t}^T \mathbf{x}) \prod_{i=1}^{d} f_{X_i}(x_i) dx_1 \ldots dx_n + \iota \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} \sin(\mathbf{t}^T \mathbf{x}) \prod_{i=1}^{d} f_{X_i}(x_i) dx_1 \ldots dx_n \\
&= \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} \exp(\iota \mathbf{t}^T \mathbf{X}) \prod_{i=1}^{d} f_{X_i}(x_i) dx_1 \ldots dx_n \\
&= \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} \prod_{i=1}^{d} \exp\left(\iota t_i x_i\right) f_{X_i}(x_i) dx_1 \ldots dx_n \\
&= \prod_{i=1}^{d} \int_{\mathbb{R}} \exp\left(\iota t_i x_i\right) f_{X_i}(x_i) dx_1 \ldots dx_n \\
&= \prod_{i=1}^{d} \mathbb{E}[\exp\left(\iota t_i x_i\right)] = \prod_{i=1}^{d} \phi_{X_i}(t_i).
\end{aligned}$$

where we were able to use Fubini, because

$$\begin{aligned}
\int_{\mathbb{R}^d} |\cos(\mathbf{t}^T \mathbf{x})| f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \leq 1 < \infty \\
\int_{\mathbb{R}^d} |\sin(\mathbf{t}^T \mathbf{x})| f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \leq 1 < \infty,
\end{aligned}$$

as shown above.

On the other hand assume that (72) holds true. Then by the first step, the characteristic function of $\mathbf{X}$ is equal to the one of $\mathbf{Y}$, with $\mathbf{Y}$ having the independence distribution, $\mu_{\mathbf{Y}} = \prod_{i=1}^{d} \mu_{X_i}$. By Theorem 4.8 this means $\mathbf{X} \overset{D}{=} \mathbf{Y}$, or the elements of $\mathbf{X}$ are independent. ■

Example 19. The cf of a gamma distribution can be derived by using the infinite series expansion of $\exp(itx)$:

$$\exp(\iota tx) = \sum_{n=0}^{\infty} \frac{(\iota tx)^n}{n!}$$

(in fact this is how $\exp(itx)$ or $\exp(\iota \mathbf{t}^T \mathbf{x})$ is properly defined). We will however not do this here and just spoil the fun of the derivations by stating that

$$\phi_X(t) = \mathbb{E}[\exp(\iota tX)] = (1 - \beta\iota t)^{-\alpha},$$

for $X \sim \text{Gam}(\alpha, \beta)$. With this and the above, we immediately get the following nice result: If $X_i \sim \text{Gam}(\alpha_i, \beta)$ are independent with the same scale parameter, then

$$\sum_{i=1}^{n} X_i \sim \text{Gam}(\sum_{i=1}^{n} \alpha_i, \beta).$$

Indeed we have that for $\mathbf{X} = (X_1, \ldots, X_n)^T$,

$$\phi_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^{n} \phi_{X_i}(t_i) = \prod_{i=1}^{n} (1 - \beta\iota t_i)^{-\alpha_i}$$

by independence. But then for $\mathbf{w} = (1, \ldots, 1)^T$, $\sum_{i=1}^{n} X_i = \mathbf{w}^T \mathbf{X}$ and

$$\phi_{\mathbf{w}^T\mathbf{X}}(t) = \mathbb{E}[\exp(\iota t \mathbf{w}^T \mathbf{X})] = \phi_{\mathbf{X}}(t\mathbf{w}) = \prod_{i=1}^{n} (1 - \beta\iota t w_i)^{-\alpha_i} = (1 - \beta\iota t)^{-\sum_{i=1}^{n} \alpha_i},$$

since $w_i = 1$ for all i. But this is just the cf of a gamma distribution with parameters $\sum_{i=1}^{n} \alpha_i, \beta$. By Theorem 4.8 this means that indeed the distribution of $\mathbf{w}^T\mathbf{X}$ is a gamma distribution with said parameters, proving the claim. In particular, if $Y_i \sim \chi_1^2$ for all $i = 1, \ldots, n$ (if $Y_i = Z_i^2$ with $Z_i \sim N(0, 1)$ for instance), then $\alpha_i = 1/2$ and $\beta_i = 2$ for all i and thus, $\sum_{i=1}^{n} Y_i \sim \chi_n^2$, as claimed in Section 3.2.4. ◊

In fact, linear transformations are no problem to handle for characteristic functions in general:

Theorem 4.10. *Let $\mathbf{X} : \Omega \to \mathbb{R}^d$ be random vector with cf $\phi_{\mathbf{X}}$ and $\mathbf{Y} : \Omega \to \mathbb{R}^p$ defined as*

$$\mathbf{Y} = \mathbf{a} + A\mathbf{X},$$

for some $\mathbf{a} \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p \times d}$. Then for all $\mathbf{t} \in \mathbb{R}^p$

$$\phi_{\mathbf{Y}}(\mathbf{t}) = \exp(\iota \mathbf{t}^T \mathbf{a}) \phi_{\mathbf{X}}(A^T \mathbf{t}).$$

Proof. First note, as so often, $\mathbf{Y}$ is $\mathcal{A}/\mathcal{B}(\mathbb{R}^p)$ measurable as a $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R}^p)$ measurable function of $\mathbf{X}$. So $\mathbf{Y}$ is indeed a valid random vector and its cf is given as

$$\begin{aligned}\phi_{\mathbf{Y}}(\mathbf{t}) &= \mathbb{E}[\exp(\iota \mathbf{t}^T \mathbf{Y})] \\ &= \mathbb{E}[\exp(\iota \left(\mathbf{t}^T \mathbf{a} + \mathbf{t}^T A\mathbf{X}\right))] \\ &= \exp(\iota \mathbf{t}^T \mathbf{a}) \mathbb{E}[\exp(\iota \mathbf{t}^T A\mathbf{X})] \\ &= \exp(\iota \mathbf{t}^T \mathbf{a}) \phi_{\mathbf{X}}(A^T \mathbf{t}).\end{aligned}$$

■

As usual the preceding also holds if we only have

$$\widetilde{\mathbf{Y}} \stackrel{D}{=} \mathbf{a} + A\mathbf{X},$$

since then $\widetilde{\mathbf{Y}}$ has the same distribution (and thus the same cf) as $\mathbf{Y} = \mathbf{a} + A\mathbf{X}$.

Characteristic functions allow for some of the most central results in the case of independent random variables. However they even are of great use, when there is dependence involved. To exemplify this, we will derive and work with the cf of the multivariate Gaussian distribution.

Example 20. If $\mu_X = N(0, 1)$, the characteristic function is

$$\phi_X(t) = \exp(-t^2/2). \tag{73}$$

Despite its simple form, deriving this is unfortunately somewhat hard. See for instance, Jacod and Protter (2004, p. 107) for a nice proof. However this expression forms the basis of the interesting theorems. First of all, if $\mu_X = N(a, \sigma^2)$, we immediately have that

$$\phi_X(t) = \exp(\iota t a) \exp(-(t^2\sigma^2)/2) = \exp(\iota t a - t^2\sigma^2/2) \tag{74}$$

Theorem 4.11 takes this a step further and gives a multivariate version of (74). ◊

Theorem 4.11. *The characteristic function for $\mu_{\mathbf{X}} = N(\mathbf{a}, \Sigma)$ is*

$$\phi_{\mathbf{X}}(\mathbf{t}) = \exp(\iota \mathbf{t}^T \mathbf{a} - \frac{1}{2} \mathbf{t}^T \Sigma \mathbf{t}) \tag{75}$$

In particular it is true that if $\mathbf{X} \sim N(\mathbf{a}, \Sigma)$, then for any $\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$,

$$\mathbf{w}^T \mathbf{X} \sim N(\mathbf{w}^T \mathbf{a}, \mathbf{w}^T \Sigma \mathbf{w}),$$

i.e. every linear combination is univariate Gaussian again.

Proof. Recall the construction of the multivariate Gaussian: We start by taking d independent $N(0, 1)$ r.v. $Z_1, \ldots, Z_d$ and $\mathbf{Z} = (Z_1, \ldots, Z_d)^T$. Then we defined:

$$\mathbf{X} \stackrel{D}{=} \mathbf{a} + \Sigma^{1/2} \mathbf{Z}.$$

Thus $\mathbf{Z} \sim N(\mathbf{0}, I)$ has cf

$$\phi_{\mathbf{Z}}(\mathbf{t}) = \prod_{i=1}^{d} \phi_{Z_i}(t_i) = \prod_{i=1}^{d} \exp(-t_i^2/2) = \exp\left(-\frac{1}{2}\sum_{i=1}^{d} t_i^2\right) = \exp\left(-\frac{1}{2}\mathbf{t}^T\mathbf{t}\right).$$

So that with Theorem 4.10:

$$\begin{aligned}
\phi_{\mathbf{X}}(\mathbf{t}) &= \exp(\iota \mathbf{t}^T\mathbf{a})\phi_{\mathbf{Z}}((\Sigma^{1/2})^T\mathbf{t}) \\
&= \exp(\iota \mathbf{t}^T\mathbf{a}) \exp\left(-\frac{1}{2}((\Sigma^{1/2})^T\mathbf{t})^T(\Sigma^{1/2})^T\mathbf{t}\right) \\
&= \exp\left(\iota \mathbf{t}^T\mathbf{a} - \frac{1}{2}\mathbf{t}^T\Sigma^{1/2}(\Sigma^{1/2})^T\mathbf{t}\right) \\
&= \exp\left(\iota \mathbf{t}^T\mathbf{a} - \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right).
\end{aligned}$$

Finally we can study the cf of $\mathbf{w}^T\mathbf{X}$ for some $\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$:

$$\begin{aligned}
\phi_{\mathbf{w}^T\mathbf{X}}(t) &= \mathbb{E}[\exp(\iota t \mathbf{w}^T\mathbf{X})] \\
&= \phi_{\mathbf{X}}(t\mathbf{w}) \\
&= \exp\left(\iota t \mathbf{w}^T\mathbf{a} - \frac{1}{2}t^2\mathbf{w}^T\Sigma\mathbf{w}\right).
\end{aligned}$$

This is the cf of a univariate Gaussian distribution with mean $\mathbf{w}^T\mathbf{a}$ and variance $\mathbf{w}^T\Sigma\mathbf{w}$, thus by Theorem 4.8, $\mathbf{w}^T\mathbf{X} \sim N(\mathbf{w}^T\mathbf{a}, \mathbf{w}^T\Sigma\mathbf{w})$. ■

Remark 14. Actually we defined the Gaussian distribution in a way such that $\mathbf{w} = \mathbf{0}$ would also be possible, with $\mathbf{0}^T\mathbf{X} \sim N(0, 0)$. This is a degenerate Gaussian distribution, without density and with all its mass on the point zero. However according to Definition 3.14 it would still be valid, since 0 is still a positive definite matrix. ◊

Thus we identified a kind of stability condition: The condition that if $\mathbf{X}$ has a certain multivariate distribution, then $\mathbf{w}^T\mathbf{X}$ is of that distributional class as well. This is an extremely useful property and does not hold in general. For instance if $\mathbf{X}$ has the independence distribution with $X_i \sim \text{Gam}(\alpha_i, \beta_i)$, i.e. with different scale terms β_i, $i = 1, \ldots, d$, then in fact $\mathbf{w}^T\mathbf{X}$ need not be gamma, even if $\mathbf{w} = (1, \ldots, 1)^T$. However the condition does extend to multivariate continuous mixtures:

Theorem 4.12. *Let* $\mathbf{X}$ *be a continuous multivariate mean-variance mixture, as in Definition 3.15, i.e.*

$$\mathbf{X} \overset{D}{=} \mathbf{a} + \boldsymbol{\beta} G + \sqrt{G}\Sigma^{1/2}\mathbf{Z},$$

with $\mathbf{a}, \boldsymbol{\beta} \in \mathbb{R}^d$, $\Sigma^{1/2}$ *a positive definite matrix,* $G : \Omega \to \mathbb{R}$ *a continuous random variable with nonnegative support and* $\mathbf{Z} \sim N(\mathbf{0}, I)$ *independent of* G. *Then the cf of* $\mathbf{X}$ *is given as*

$$\phi_{\mathbf{X}}(\mathbf{t}) = \exp(\iota \mathbf{t}^T\mathbf{a})\phi_G(\mathbf{t}^T\boldsymbol{\beta} + \iota\frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}). \tag{76}$$

Furthermore if $\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$*, then* $Y = \mathbf{w}^T\mathbf{X}$ *is in the same distributional class. More precisely,* Y *is a univariate mean variance mixture with parameters* $\mathbf{w}^T\mathbf{a}$*,* $\mathbf{w}^T\boldsymbol{\beta}$*,* $\mathbf{w}^T\Sigma\mathbf{w}$*, or:*

$$Y \overset{D}{=} \mathbf{w}^T\mathbf{a} + G\mathbf{w}^T\boldsymbol{\beta} + G^{1/2}\left(\mathbf{w}^T\Sigma\mathbf{w}\right)^{1/2} Z,$$

with $Z \sim N(0,1)$ *and* G, Z *independent.*

Proof. The first argument is again not completely rigorous, since we did not define conditional expectation properly. Yet the argument itself is entirely correct: Using the law of iterated expectations,

$$\begin{aligned}\phi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}[\exp(\iota\mathbf{t}^T\mathbf{X})] \\ &= \mathbb{E}[\mathbb{E}[\exp(\iota\mathbf{t}^T\mathbf{X})|G]].\end{aligned}$$

Now since the random vector $\mathbf{M} = \mathbf{X}|G \sim N(\mathbf{a} + \boldsymbol{\beta}G, G\Sigma)$, we have that the inner expectation is given as:

$$\begin{aligned}\mathbb{E}[\exp(\iota\mathbf{t}^T\mathbf{X})|G] &= \phi_{\mathbf{M}}(\mathbf{t}) \\ &= \exp(\iota\mathbf{t}^T(\mathbf{a} + \boldsymbol{\beta}G) - \frac{1}{2}\mathbf{t}^TG\Sigma\mathbf{t}),\end{aligned}$$

so that

$$\begin{aligned}\phi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}[\exp(\iota\mathbf{t}^T\mathbf{a} + \iota\mathbf{t}^T\boldsymbol{\beta}G - \frac{1}{2}\mathbf{t}^TG\Sigma\mathbf{t})] \\ &= \exp(\iota\mathbf{t}^T\mathbf{a})\mathbb{E}[\exp((\iota\mathbf{t}^T\boldsymbol{\beta} - \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t})G)] \\ &= \exp(\iota\mathbf{t}^T\mathbf{a})\mathbb{E}[\exp((\iota\mathbf{t}^T\boldsymbol{\beta} + \iota^2\frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t})G)] \\ &= \exp(\iota\mathbf{t}^T\mathbf{a})\phi_G(\mathbf{t}^T\boldsymbol{\beta} + \iota\frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}),\end{aligned}$$

as claimed.

Now let $\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. The characteristic function of $Y = \mathbf{w}^T\mathbf{X}$ is found as usual:

$$\begin{aligned}\phi_Y(t) &= \mathbb{E}[\exp(t\mathbf{w}^T\mathbf{X})] \\ &= \phi_{\mathbf{X}}(t\mathbf{w}) \\ &= \exp(\iota t\mathbf{w}^T\mathbf{a})\phi_G(t\mathbf{w}^T\boldsymbol{\beta} + \iota t^2\frac{1}{2}\mathbf{w}^T\Sigma\mathbf{w})).\end{aligned}$$

By comparison this is clearly the cf of a univariate mixture with parameters $\mathbf{w}^T\mathbf{a}$, $\mathbf{w}^T\boldsymbol{\beta}$, $\mathbf{w}^T\Sigma\mathbf{w}$. With Theorem 4.8 this also means the distribution of Y is said mixture. ■

Example 21. If $\mathbf{X} \sim N(\mathbf{a}, \boldsymbol{\Sigma})$ or $\mathbf{Y} \sim t(\nu, \mathbf{a}, \boldsymbol{\Sigma})$, then from Theorems 4.11 and 4.12 we have for any linear combination:

$$\begin{aligned}\mathbf{w}^T\mathbf{X} &\sim N(\mathbf{w}^T\mathbf{a}, \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}) \\ \mathbf{w}^T\mathbf{Y} &\sim t_\nu(\mathbf{w}^T\mathbf{a}, \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})\end{aligned}$$

for all $\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. If we for instance model logged asset returns with a multivariate Gaussian or t distribution, we immediately know the distribution of a portfolio return as well (since a portfolio is just a linear combination of assets). For portfolio optimization this is naturally a highly desirable property. We will illustrate this here with an imperfect, but still illustrative example: Let $\mathbf{Y}_{T+1}$ be the returns of some set of d assets at $T+1$. We aim to solve

$$\min_{\mathbf{w} \in \mathcal{B}_\theta} \mathbb{V}(\mathbf{w}^T \mathbf{Y}_{T+1}), \tag{77}$$

with $\mathcal{B}_\theta = \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w}^T \mathbb{I} = 1, w_i \geq 0, i \in \{1, \ldots, K\}\}$, where $\mathbb{I}$ is a $d \times 1$ vector of ones. Where is this optimization problem coming from? Variance in this context can be understood as a measure of risk and we would like to minimize this risk as far as possible with our choice of $\mathbf{w}$. In fact it would be better to minimize a far more sophisticated measure of risk, such as expected shortfall (ES):

$$\min_{\mathbf{w} \in \mathcal{B}_\theta} \mathrm{ES}_\alpha(\mathbf{w}^T \mathbf{Y}_{T+1}), \tag{78}$$

for some $\alpha \in (0, 1)$. ES can be regarded as the expected loss at some level α. See for instance McNeil et al. (2005) for details. However in case of *elliptical distributions* (which include both the Gaussian and the multivariate t), it can be demonstrated that problems (77) and (78) yield the same solutions. This is the reason we focus on the much simpler problem (77). The drawback from an illustrative point of view is that we don't really need Theorem 4.12 here, since $\mathbb{V}(\mathbf{w}^T \mathbf{Y}_{T+1}) = \mathbf{w}^T \mathbb{V}(\mathbf{Y}_{T+1}) \mathbf{w}$ is always true, no matter what the distribution of $\mathbf{Y}_{T+1}$ is. This changes however once we leave the nice world of elliptical distributions, if we work with the MVG of Section 3.4.4 for instance. In this case we need to solve problem (78) and for that, it helps greatly to know that $\mathbf{w}^T \mathbf{Y}_{T+1}$ follows a variance gamma distribution if $\mathbf{Y}_{T+1}$ is multivariate variance gamma!

Now, we assume to observe a sample of T returns of d assets, $(\mathbf{Y}_1, \ldots, \mathbf{Y}_T)$ and take

(1) $\mathbf{Y}_t \overset{iid}{\sim} N(\mathbf{a}, \mathbf{\Sigma})$ for all $t = 1, \ldots, T$

(2) $\mathbf{Y}_t \overset{iid}{\sim} t(\nu, \mathbf{a}, \mathbf{\Sigma})$ for all $t = 1, \ldots, T$.

In case (1) we estimate $(\mathbf{a}, \mathbf{\Sigma})$ as before as:

$$\hat{\mathbf{a}} = \bar{\mathbf{Y}} = \frac{1}{T} \sum_{i=1}^{T} \mathbf{Y}_t$$

$$\hat{\Sigma} = \frac{1}{T} \sum_{i=1}^{T} \left(\mathbf{Y}_t - \bar{\mathbf{Y}}\right) \left(\mathbf{Y}_t - \bar{\mathbf{Y}}\right)^T.$$

Estimating the t distribution on the other hand is more difficult. Standard maximum likelihood is not a good idea, since estimating a matrix Σ via numerical optimization is a nightmare. Instead we use the following heuristic algorithm:[23] We know from Theorem 4.12 (taking $\mathbf{w}$ to be one at index i and zero everywhere else) that $Y_{t,i} \overset{iid}{\sim} t_\nu(a_i, \sigma_i^2)$ for $i = 1, \ldots, d$. Thus in the first step

[23] A very elegant and powerful solution to this estimation problem is the so called EM algorithm, which we will however not discuss here.

we simply fit a univariate t distribution (augmented with scale term σ_i and location a_i) to each univariate sample $(Y_{t,i})_{t=1}^T$ using maximum likelihood. This will result in different ν values for each i (and to a loss in estimation efficiency). To counter this, we simply take the mean over all ν_i, i.e. we estimate

$$\hat{\nu} = \frac{1}{d}\sum_{i=1}^{d} \hat{\nu}_i,$$

where $\hat{\nu}_i$ are the estimates of ν for each component, obtained with maximum likelihood. This method gives us an estimate of ν, $\mathbf{a} = (a_1, \ldots, a_d)^T$ and the diagonal elements of Σ, $\sigma_1^2, \ldots, \sigma_d^2$. Now to estimate the off-diagonal elements of Σ, we use a so called method-of-moment approach. As we derived earlier for any $i \neq j$:

$$\mathrm{Cov}(Y_{t,i}, Y_{t,j}) = \frac{\nu}{\nu - 2}\sigma_{i,j} \text{ for all } t.$$

Now given an estimate of ν, $\hat{\nu}$, a simple idea is to just estimate $\mathrm{Cov}(Y_{t,i}, Y_{t,j})$ with the consistent estimator

$$\frac{1}{T}\sum_{t=1}^{T}(Y_{t,i} - \bar{Y}_i)(Y_{t,j} - \bar{Y}_j),$$

where $\bar{Y}_i = 1/T\sum_{t=1}^T Y_{t,i}$ and analogously with $\bar{Y}_j$. Given an estimate of the covariance, $\widehat{\mathrm{Cov}}(Y_{t,i}, Y_{t,j})$ say, we calculate

$$\hat{\sigma}_{i,j} = \widehat{\mathrm{Cov}}(Y_{t,i}, Y_{t,j})\frac{\hat{\nu} - 2}{\hat{\nu}}$$

Note that this only makes sense if $\hat{\nu} > 2$.

Thus, given some returns data $(\mathbf{Y}_t)_{t=1}^T$, we have by assumption that also $\mathbf{Y}_{T+1} \sim N(\mathbf{a}, \mathbf{\Sigma})$ in case (1) and $\mathbf{Y}_{T+1} \sim t(\nu, \mathbf{a}, \mathbf{\Sigma})$ in case (2). We can then use the estimates obtained in a first step to solve problem (77).

$\Diamond$

A Review of Complex numbers

The imaginary unit ι is defined to be "number" such that $\iota^2 = -1$. The set of all complex numbers is

$$\mathbb{C} := \{a + b\iota : a, b \in \mathbb{R}\},$$

and is closed under addition and multiplication:

$$(a + b\iota) + (c + d\iota) = (a + c) + (b + d)\iota$$
$$(a + b\iota)(c + d\iota) = (ac - bd) + (bc + ad)\iota.$$

If $z = a + b\iota$, then $\Re(z) := a$ and $\Im(z) := b$ are the real and imaginary parts of z. We can also define convergence on $\mathbb{C}$ by saying that a sequence $(z_n)_n$ in $\mathbb{C}$ converges to $z \in \mathbb{C}$ if $\Re(z_n) \to \Re(z)$ *and* $\Im(z_n) \to \Im(z)$ in $\mathbb{R}$, as n goes to infinity.

The complex conjugate of z is $\bar{z} = a - b\iota$. The product $z\bar{z} = (a+b\iota)(a-b\iota) = a^2 - b^2\iota^2 = a^2 + b^2$ is always a non-negative real number. The sum is $z + \bar{z} = (a + bi) + (a - bi) = 2a = 2Re(z)$. The absolute value of z, or its (complex) modulus, is $|z| = |a + bi| = \sqrt{z\bar{z}} = \sqrt{a^2 + b^2}$. Short calculations show that, for $z_1, z_2 \in \mathbb{C}$,

$$|z_1 z_2| = |z_1||z_2|, \quad \overline{z_1 z_2} = \bar{z}_1 \bar{z}_2, \quad |z_1 + z_2| \leq |z_1| + |z_2|.$$

The exponential function $\exp : \mathbb{C} \to \mathbb{C}$ is defined as

$$\exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!}.$$

(This is in particular true for the special case $z \in \mathbb{R}$.) As in the real case it holds that for $z_1, z_2 \in \mathbb{C}$, $\exp(z_1 + z_2) = \exp(z_1)\exp(z_2)$. Looking at the series expansion of $\sin : \mathbb{R} \to \mathbb{R}$ and $\cos : \mathbb{R} \to \mathbb{R}$, one can determine the important relation:

$$\exp(it) = \cos(t) + \iota \sin(t), \quad \forall t \in \mathbb{R}. \tag{79}$$

This is generally referred to as "Euler's formula". Using $t = \pi$, we have in particular $\exp(\iota\pi) + 1 = 0$, since $\cos(\pi) = -1$ and $\sin(\pi) = 0$. So for an arbitrary $z \in \mathbb{C}$, $z = a + \iota b$, it holds that

$$\exp(z) = \exp(a + \iota b) = \exp(a)\exp(\iota b) = \exp(a)(\cos(b) + \iota \sin(b)) = \exp(a)\cos(b) + \iota \exp(a)\sin(b).$$

So we may say $\Re(\exp(z)) = \exp(a)\cos(b)$ and $\Im(\exp(a)) = \exp(a)\sin(b)$. Thus we can study the complex conjugate:

$$\overline{\exp(z)} = \exp(a)\cos(b) - \iota \exp(a)\sin(b) = \exp(a)(\cos(b) - \iota \sin(b))$$
$$= \exp(a)(\cos(b) + \iota \sin(-b)) = \exp(a)\exp(-\iota b) = \exp(a - \iota b) = \exp(\bar{z}),$$

i.e. the complex conjugate of $\exp(z) \in \mathbb{C}$ is $\exp(\bar{z}) \in \mathbb{R}$. Very importantly we can also obtain the following equality, for all $t \in \mathbb{R}$:

$$|\exp(it)| = \sqrt{\Re(\exp(\iota t))^2 + \Im(\exp(\iota t))^2} = \sqrt{\cos(t)^2 + \sin(t)^2} = 1, \tag{80}$$

as $a = 0$ and thus $\exp(a) = 1$ here and since for any $t \in \mathbb{R}$: $\cos(t)^2 + \iota \sin(t)^2 = 1$. This is basically the reason that the characteristic function *always* exists.

References

Dudley, R. (2002). *Real Analysis and Probability.* Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Durrett, R. (2010). *Probability: Theory and Examples.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Jacod, J. and Protter, P. (2004). *Probability Essentials.* Hochschultext / Universitext. Springer.

McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools.* Princeton University Press, Princeton.

Paolella, M. S. (2006). *Fundamental Probability.* John Wiley & Sons, Chichester, West Sussex, England.

Paolella, M. S. (2007). *Intermediate Probability: A Computational Approach.* John Wiley & Sons, Chichester.

Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach.* Pearson Education, 2 edition.

van der Waart, A. W. and Wellner, J. A. (1996). *Weak convergence and Empirical Processes : With Applications to Statistics.* Springer Series in Statistics. Springer, New York.