Probability Theory

Alexander Grigoryan

WS 2010/2011, Bielefeld

Contents

Chapter 1

Introduction

Lecture 1
13.09.10

1.1 Empirical probability

It is assumed that a reader knows already elementary probability theory with coin flipping, gambler games etc. The purpose of this course is to provide a rigorous background of probability theory as a part of mathematics, and to show its close relations to other fields, especially Analysis.

Probability theory deals with random events and their probabilities. A classical example of a random event is a coin flip. The outcome of each flip may be heads or tails: H or T.

Figure 1.1: Heads and tails of a british pound

If the coin is fair then after N trials, H occurs approximately $N/2$ times, and so does T. It is natural to *believe* that if $N \to \infty$ then $\frac{\#H}{N} \to \frac{1}{2}$ so that one says that H occurs with probability $1/2$ and writes $\mathbb{P}(H) = 1/2$. In the same way $\mathbb{P}(T) = 1/2$.

If a coin is biased then $\mathbb{P}(H)$ may different from $1/2$, let

$$\mathbb{P}(H) = p \quad \text{and} \quad \mathbb{P}(T) = q := 1 - p. \tag{1.1}$$

Let us show here a curious example how the random events H and T satisfying (1.1) can be used to prove the following purely deterministic inequality:

$$(1 - p^n)^m + (1 - q^m)^n \geq 1, \tag{1.2}$$

where $0 < p, q < 1$, $p + q = 1$, and n, m are positive integers. This inequality has an algebraic proof which however is more complicated than the probabilistic argument below.

Let us flip the biased coin $N = nm$ times *independently* and write down the outcome of the trials in a $n \times m$ table putting in each cell H or T:

$$\overbrace{\left\{\begin{array}{|c|c|c|c|c|}\hline H & T & T & H & T \\ \hline T & T & H & H & H \\ \hline H & H & T & H & T \\ \hline T & H & T & T & T \\ \hline\end{array}\right.}^{m}\Bigg\} n$$

Then, using the elementary probability theory, we obtain:

$$\begin{aligned} p^n &= \mathbb{P}\{\text{a given column contains only } H\text{'}s\} \\ 1 - p^n &= \mathbb{P}\{\text{ a given column contains at least one } T\}\,. \end{aligned}$$

It follows that

$$(1 - p^n)^m = \mathbb{P}\{\text{any column contains at least one } T\ \} \tag{1.3}$$

and similarly

$$(1 - q^m)^n = \mathbb{P}\{\text{any row contains at least one } H\}\,. \tag{1.4}$$

Let us show that one of the events (1.3) and (1.4) will always take place, which would imply that the sum of their probabilities is at least 1, and hence prove (1.2). Indeed, assume that the event (1.3) does not take place, that is, some column contains only H's:

$$\begin{array}{|c|c|c|c|c|}\hline & & H & & \\ \hline & & H & & \\ \hline & & H & & \\ \hline & & H & & \\ \hline\end{array}$$

Then one easily sees that H exists in *any row* so that the event (1.4) takes place, which was to be proved.

Is this proof rigorous? It may leave impression of a rather empirical argument than a mathematical proof. The point is that we have used in this argument the *existence* of events with certain properties: firstly, H should have probability p where p a given number in $(0, 1)$ and secondly, there must be sufficiently many independent events like that. Certainly, mathematics cannot rely on the existence of biased coins (or even fair coins!). In order to make the above argument rigorous, one should have a mathematical notion of events and their probabilities.

One of the purposes of this course is to introduce such notions and, based on them, to prove the classical results of probability theory such as laws of large numbers, central limit theorems etc.

1.2 Algebras of sets

Let Ω be any set. We consider the following set theoretic operations over subsets A, B of Ω: the union $A \cup B$, the intersection $A \cap B$, the difference $A \setminus B$ and a particular case of the latter: the complement $A^c = \Omega \setminus A$.

Definition. Given a set Ω and a family $\mathcal{F}$ of its subsets, we say that $\mathcal{F}$ is *algebra* if

1. $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$
2. if $A, B \in \mathcal{F}$ then $A \cap B \in \mathcal{F}$
3. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$

Lemma 1.1 *If $\mathcal{F}$ is algebra and $A, B \in \mathcal{F}$ then $A \cup B \in \mathcal{F}$ and $A \setminus B \in \mathcal{F}$. Hence, $\mathcal{F}$ is closed under the operations $\cap, \cup, \setminus$.*

Proof. Indeed, we have

$$(A \cup B)^c = A^c \cap B^c \in \mathcal{F}$$

whence $A \cup B \in \mathcal{F}$. For the difference, we have

$$A \setminus B = A \cap B^c \in \mathcal{F}.$$

■

A simplest example of an algebra is the family 2^Ω of all subsets of Ω.

Definition. Given a set Ω and a family $\mathcal{F}$ of its subsets, we say that $\mathcal{F}$ is *semi-algebra* if

1. $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$
2. if $A, B \in \mathcal{F}$ then $A \cap B \in \mathcal{F}$
3. if $A \in \mathcal{F}$ then A^c is a finite disjoint union of sets from $\mathcal{F}$.

Clearly, any algebra is also a semi-algebra.

Example. Let $\Omega = \mathbb{R}$ and $\mathcal{F}$ is the family of all intervals in $\mathbb{R}$, that is, the sets of one of the form

$$\begin{aligned} (a,b) &= \{x \in \mathbb{R} : a < x < b\} \\ [a,b) &= \{x \in \mathbb{R} : a \le x < b\} \\ (a,b] &= \{x \in \mathbb{R} : a < x \le b\} \\ [a,b] &= \{x \in \mathbb{R} : a \le x \le b\} \end{aligned}$$

where a, b are either reals or $\pm\infty$. Clearly, $\mathcal{F}$ contains $\emptyset = (0,0)$ and $\mathbb{R} = (-\infty, +\infty)$, the intersection of any two intervals is again an interval, and the complement of any

interval is either an interval or a disjoint union of two intervals. Hence, $\mathcal{F}$ is a semi-algebra (but not algebra).

The same is true if Ω is any interval in $\mathbb{R}$ and $\mathcal{F}$ is the family of all subintervals of Ω.

Definition. Given a set Ω and a family $\mathcal{F}$ of its subsets, we say that $\mathcal{F}$ is σ-algebra if $\mathcal{F}$ is an algebra and $\mathcal{F}$ admits the following property: if $\{A_n\}$ is a countable sequence of sets from $\mathcal{F}$ then

$$\bigcap_{n=1}^{\infty} A_n \in \mathcal{F}.$$

Lemma 1.2 *If $\mathcal{F}$ is a σ-algebra and $A_n \in \mathcal{F}$ then*

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

Proof. Indeed, we have

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

whence the claim follows. ■

Hence, σ-algebra is closed under the following operations: unions and intersections of at most countable number of sets as well as subtractions (and taking complement). The difference between an algebra and a σ-algebra is that the former is closed under those operations with *finite* numbers of sets.

Example. Let $\Omega = \mathbb{N} = \{1, 2, 3, ...\}$ and $\mathcal{F}$ be the family of all subsets of Ω that are finite or their complements are finite. It is easy to see that $\mathcal{F}$ is an algebra. However, $\mathcal{F}$ is not a σ-algebra because $\mathcal{F}$ contains the singletons $\{1\}, \{3\}, \{5\}, ...$ (that is, the subset containing a single odd number) but not their union $\{1, 3, 5, ...\}$.

1.3 The notion of a measure

Let us recall the familiar from the elementary mathematics notions, which are all particular cases of a measure.

1. *Length* of intervals in $\mathbb{R}$: if I is a bounded interval with the endpoints a, b (that is, I is one of the intervals (a, b), $[a, b]$, $[a, b)$, $(a, b]$) then its length is defined by

$$\ell(I) = |b - a|.$$

The useful property of the length is the *additivity*: if an interval I is a disjoint union of a finite family $\{I_k\}_{k=1}^n$ of intervals, that is, $I = \bigsqcup_k I_k$, then

$$\ell(I) = \sum_{k=1}^{n} \ell(I_k)$$

Indeed, let $\{a_i\}_{i=0}^N$ be the set of all distinct endpoints of the intervals $I, I_1, ..., I_n$ enumerated in the increasing order. Then I has the endpoints a_0, a_N while each interval I_k has necessarily the endpoints a_i, a_{i+1} for some i (indeed, if the endpoints of I_k are a_i and a_j with $j > i+1$ then the point a_{i+1} is an interior point of I_k, which means that I_k must intersect with some other interval I_m). Conversely, any couple a_i, a_{i+1} of consecutive points are the end points of some interval I_k (indeed, the interval (a_i, a_{i+1}) must be covered by some interval I_k; since the endpoints of I_k are consecutive numbers in the sequence $\{a_j\}$, it follows that they are a_i and a_{i+1}). We conclude that

$$\ell(I) = a_N - a_0 = \sum_{i=0}^{N-1} (a_{i+1} - a_i) = \sum_{k=1}^{n} \ell(I_k).$$

(See also the proof of Theorem 2.10 below).

2. *Area* of domains in $\mathbb{R}^2$. The full notion of area can constructed only within the general measure theory, that will be partly discussed in this course. However, for rectangular domains the area is defined easily. A *rectangle* A in $\mathbb{R}^2$ is defined as the direct product of two intervals I, J from $\mathbb{R}$:

$$A = I \times J = \left\{(x, y) \in \mathbb{R}^2 : x \in I, y \in J\right\}.$$

Then set

$$\text{area}(A) = \ell(I)\,\ell(J).$$

We claim that the area is also additive: if a rectangle A is a disjoint union of a finite family of rectangles $A_1, ..., A_n$, that is, $A = \bigsqcup_k A_k$, then

$$\text{area}(A) = \sum_{k=1}^{n} \text{area}(A_k).$$

For simplicity, let us restrict the consideration to the case when all sides of all rectangles are semi-open intervals of the form $[a, b)$. Consider first a particular case, when the rectangles $A_1, ..., A_k$ form a *regular tiling* of A; that is, let $A = I \times J$ where $I = \bigsqcup_i I_i$ and $J = \bigsqcup_j J_j$, and assume that all rectangles A_k have the form $I_i \times J_j$. Then

$$\text{area}(A) = \ell(I)\,\ell(J) = \sum_i \ell(I_i) \sum_j \ell(J_j) = \sum_{i,j} \ell(I_i)\,\ell(J_j) = \sum_k \text{area}(A_k).$$

Now consider the general case when A is an arbitrary disjoint union of rectangles A_k. Let $\{x_i\}$ be the set of all X-coordinates of the endpoints of the rectangles A_k put in the increasing order, and $\{y_j\}$ be similarly the set of all the Y-coordinates, also in the increasing order. Consider the rectangles

$$B_{ij} = [x_i, x_{i+1}) \times [y_j, y_{j+1}).$$

Then the family $\{B_{ij}\}_{i,j}$ forms a regular tiling of A and, by the first case,

$$\text{area}(A) = \sum_{i,j} \text{area}(B_{ij}).$$

On the other hand, each A_k is a disjoint union of some of B_{ij}, and, moreover, those B_{ij} that are subsets of A_k, form a regular tiling of A_k, which implies that

$$\text{area}(A_k) = \sum_{B_{ij} \subset A_k} \text{area}\,(B_{ij})\,.$$

Combining the previous two lines and using the fact that each B_{ij} is a subset of exactly one set A_k, we obtain

$$\sum_k \text{area}\,(A_k) = \sum_k \sum_{B_{ij} \subset A_k} \text{area}\,(B_{ij}) = \sum_{i,j} \text{area}\,(B_{ij}) = \text{area}\,(A)\,.$$

3. *Volume* of domains in $\mathbb{R}^3$. The construction of volume is similar to that of area. Consider all *boxes* in $\mathbb{R}^3$, that is, the domains of the form $A = I \times J \times K$ where I, J, K are intervals in $\mathbb{R}$, and set

$$\text{vol}\,(A) = \ell\,(I)\,\ell\,(J)\,\ell\,(K)\,.$$

Then volume is also an additive functional, which is proved as above.

4. *Probability* provides another example of an additive functional. In probability theory, one considers a set Ω of elementary events, and certain subsets of Ω are called *events.* For each event $A \subset \Omega$, one assigns the probability, which is denoted by $\mathbb{P}\,(A)$ and which is a real number in $[0, 1]$. A reasonably defined probability must satisfy the additivity: if the event A is a disjoint union of a finite sequence of evens $A_1, ..., A_n$ then

$$\mathbb{P}\,(A) = \sum_{k=1}^{n} \mathbb{P}\,(A_k)\,.$$

The fact that A_i and A_j are disjoint, when $i \neq j$, means that the events A_i and A_j cannot occur at the same time.

The common feature of all the above example is the following. We are given a non-empty set Ω, a family $\mathcal{F}$ of its subsets (the families of intervals, rectangles, boxes, events), and a function $\mu : \mathcal{F} \to \mathbb{R}_+ := [0, +\infty)$ (length, area, volume, probability) with the following property: if $A \in \mathcal{F}$ is a disjoint union of a finite family $\{A_k\}_{k=1}^n$ of sets from $\mathcal{F}$ then

$$\mu\,(A) = \sum_{k=1}^{n} \mu\,(A_k)\,.$$

A function μ with this property is called a *finitely additive measure.* Hence, length, area, volume, probability are all finitely additive measures.

Now let us introduce an abstract notion of a measure. Fix a set Ω and a let $\mathcal{F}$ be a family of subsets of Ω. Let $\mu : \mathcal{F} \to \mathbb{R}_+$ be a non-negative function on $\mathcal{F}$.

Definition. The function μ is called *σ-additive* (or countably additive) if, for any finite or countable sequence $\{A_i\}$ of pairwise disjoint sets $A_i \in \mathcal{F}$,

$$\mu(\bigcup_i A_i) = \sum_i \mu(A), \tag{1.5}$$

provided $\bigcup_i A_i \in \mathcal{F}$. Any σ-additive function $\mu : \mathcal{F} \to \mathbb{R}_+$ is called a *measure*.

The function μ is called *finitely additive* (or a finitely additive measure) if the property (1.5) holds only for finite sequences $\{A_i\}$.

Remark. If $\mathcal{F}$ is an algebra and (1.5) holds for two disjoint sets A_1, A_2 then it holds also for any finite sequence $\{A_i\}_{i=1}^n$, which is proved by induction in n.

If $\mathcal{F}$ contains $\emptyset$ and (1.5) holds for countable sequences $\{A_i\}$ then it holds also for any finite sequences. Indeed, one first observes that $\mu(\emptyset) = 0$; then any finite sequence can be extended to a countable sequence by adding empty sets.

1.4 Probability measures

Definition. A measure μ is said to be a *probability measure* if $\Omega \in \mathcal{F}$ and $\mu(\Omega) = 1$.

Example. Let $\mathcal{F}$ be the family of all subintervals of the interval $\Omega = [0, 1]$ and $\mu(A)$ be the length of A. Then μ is a probability measure.

Lecture 2
14.09.10

Consider a class of probability measures that are called *discrete*. Let Ω be a finite or countable set and $\{p_k\}_{k\in\Omega}$ be a *stochastic sequence*, that is, p_k are non-negative reals such that $\sum_k p_k = 1$. For any subset $A \subset \Omega$ define

$$\mathbb{P}(A) = \sum_{k\in A} p_k. \tag{1.6}$$

Claim. *The identity (1.6) defines a probability measure $\mathbb{P}$ on $\mathcal{F} = 2^\Omega$. Conversely, any probability measure on 2^Ω is given by (1.6) for some stochastic sequence $\{p_k\}$.*

Proof. Let $\{A_n\}$ be a disjoint sequence subsets of Ω, finite or countable. We have

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{k\in\bigsqcup_n A_n} p_k = \sum_n \sum_{k\in A_n} p_k = \sum_n \mathbb{P}(A_n)$$

so that $\mathbb{P}$ is σ-additive. Clearly, $\mathbb{P}$ is a probability measure because

$$\mathbb{P}(\Omega) = \sum_k p_k = 1.$$

If $\mathbb{P}$ is any probability measure on Ω then define $p_k = \mathbb{P}(\{k\})$. Then (1.6) holds for all $A \subset \Omega$ by σ-additivity. In particular, for $A = \Omega$, we obtain $\sum_k p_k = 1$. ■

If the set Ω is finite and $|\Omega| = n$ then the simplest probability measure on Ω is given by the sequence $p_k = \frac{1}{n}$. The measure $\mathbb{P}$ defined in this way is called the *uniform distribution* on Ω and is denoted by $U(n)$. In this case, for any set $A \subset \Omega$, we have

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

Let $\Omega = \{0, 1, ..., n\}$. The *binomial distribution* $B(n, p)$ on Ω is the probability measure $\mathbb{P}$ on Ω that is determined by the stochastic sequence

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, ..., n,$$

where $p \in (0, 1)$ is a given parameter and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient. This sequence is stochastic because by the binomial formula

$$\sum_{k=0}^{n} p_k = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1,$$

where $q = 1 - p$.

Let $\Omega = \{0, 1, 2, ...\}$ be countable. The *Poisson distribution* $Po(\lambda)$ with parameter $\lambda > 0$ is defined by the stochastic sequence

$$p_k = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, ...$$

The *exponential* (or geometric) *distribution* with parameter $\alpha \in (0, 1)$ is defined by

$$p_k = (1-\alpha)\,\alpha^k, \quad k = 0, 1, ...$$

Definition. A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is any set, $\mathcal{F}$ is a σ-algebra of subsets of Ω and $\mathbb{P}$ is a probability measure on $\mathcal{F}$. Elements of Ω are called *elementary events*. Elements of $\mathcal{F}$ are called *events*. For any event $A \in \mathcal{F}$, $\mathbb{P}(A)$ is called its probability.

Example. Let Ω be a finite or countable set and $\mathbb{P}$ be a probability measure on $\mathcal{F} = 2^\Omega$ given by a stochastic sequence as above. Then the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, since $\mathcal{F}$ is obviously a σ-algebra. We refer to such spaces $(\Omega, \mathcal{F}, \mathbb{P})$ as *discrete* probability spaces.

Example. Let $\Omega = [0, 1]$ and $\mathbb{P}$ be the length defined on the family $\mathcal{F}$ of all subintervals of Ω. One can show that $\mathbb{P}$ is indeed a probability measure. However, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is not a probability space because $\mathcal{F}$ is not a σ-algebra (but a semi-algebra). As we will see later, the domain $\mathcal{F}$ of $\mathbb{P}$ can be extended to a σ-algebra.

The same applies to the unit square $\Omega = [0, 1]^2$ with $\mathcal{F}$ being the family of rectangles in Ω and $\mathbb{P}$ being the area. The domain $\mathcal{F}$ of $\mathbb{P}$ can be extended to a σ-algebra so that $(\Omega, \mathcal{F}, \mathbb{P})$ becomes a probability space.

Chapter 2

Construction of measures

2.1 Extension of families of subsets

Given any family $\mathcal{F}$ of subsets of Ω, there exists at least one algebra containing $\mathcal{F}$, for example, the family of *all* subsets of Ω. Note that if $\mathcal{F}_\alpha$ are algebras (where the parameter α runs over any set of indices) then

$$\bigcap_\alpha \mathcal{F}_\alpha$$

is again algebra (and the same is true for σ-algebras). Hence, by taking the intersection of all the algebras, containing $\mathcal{F}$, we obtain the minimal algebra containing $\mathcal{F}$ that will be denoted by $a(\mathcal{F})$; that is,

$$a(\mathcal{F}) = \bigcap \mathcal{A}$$

where $\mathcal{A}$ runs over all algebras of subsets of Ω, containing $\mathcal{F}$. In the same way, one defines the minimal σ-algebra $\sigma(\mathcal{F})$ containing $\mathcal{F}$:

$$\sigma(\mathcal{F}) = \bigcap \mathcal{A}$$

where now $\mathcal{A}$ runs over all σ-algebras of subsets of Ω, containing $\mathcal{F}$.

One says that the family $\mathcal{F}$ generates the algebra $a(\mathcal{F})$ and the σ-algebra $\sigma(\mathcal{F})$.

Example. (Exercise 1) If $\mathcal{F}$ consists of a single subset, say $\mathcal{F} = \{A\}$, then $a(\mathcal{F}) = \{\emptyset, \Omega, A, A^c\}$. If $\mathcal{F}$ consists of two subsets, say $\mathcal{F} = \{A, B\}$ then $a(\mathcal{F})$ consists of all possible unions of the following four disjoint sets:

$$A \cap B,\ A \setminus B,\ B \setminus A,\ (A \cup B)^c.$$

Theorem 2.1 *If $\mathcal{F}$ is semi-algebra then $a(\mathcal{F})$ consists of all finite disjoint unions of elements of $\mathcal{F}$.*

Proof. We denote disjoint union by using the sign $\bigsqcup$. Denote by $\mathcal{F}'$ the family of sets each of which is a finite disjoint union of sets from $\mathcal{F}$. We want to prove $\mathcal{F}' = a(\mathcal{F})$. Since $a(\mathcal{F})$ is an algebra containing $\mathcal{F}$, $a(\mathcal{F})$ must contain also finite unions of sets from $\mathcal{F}$ and, hence, $a(\mathcal{F}) \supset \mathcal{F}'$.

We need then only to verify $a(\mathcal{F}) \subset \mathcal{F}'$, and for that it suffices to show that $\mathcal{F}'$ is algebra. Clearly, $\emptyset$ and Ω are in $\mathcal{F}'$. Let us prove that if $A, B \in \mathcal{F}'$ then $A \cap B \in \mathcal{F}'$. By definition of $\mathcal{F}'$, A and B can be represented in the form

$$A = \bigsqcup_{i=1}^{n} A_i \quad \text{and} \quad B = \bigsqcup_{j=1}^{m} B_j$$

where A_i and B_j are in $\mathcal{F}$. Then

$$A \cap B = \left(\bigsqcup_i A_i \right) \cap \left(\bigsqcup_j B_j \right) = \bigsqcup_{i,j} (A_i \cap B_j).$$

Since $A_i \cap B_j \in \mathcal{F}$, we conclude that $A \cap B$ is a finite disjoint union of sets from $\mathcal{F}$ and, hence, $A \cap B \in \mathcal{F}'$.

We are left to verify that if $A \in \mathcal{F}'$ then $A^c \in \mathcal{F}'$. Indeed, we have

$$A^c = \left(\bigsqcup_i A_i \right)^c = \bigcap_i A_i^c.$$

Since $A_i \in \mathcal{F}$, we have that A_i^c is a finite disjoint union of sets from $\mathcal{F}$ and hence $A_i^c \in \mathcal{F}'$. As we have already proved, intersection of two sets from $\mathcal{F}'$ is again in $\mathcal{F}'$. By induction, this extends to intersection of a finite numbers of sets. Hence, $\bigcap A_i^c \in \mathcal{F}'$. ■

Example. Let $\Omega = [0, 1]$ and $\mathcal{F}$ be a family of all intervals on $[0, 1]$. As we have already seen, $\mathcal{F}$ is a semi-algebra. Hence, by Theorem 2.1, the family of all finite disjoint unions of intervals is an algebra.

A similar construction works in higher-dimensional spaces. Let Ω be a unit square $[0, 1]^2$, and $\mathcal{F}$ be a set of all rectangles in Ω (by rectangle we mean here a product $I_1 \times I_2$ where I_1 and I_2 are intervals). Then again $\mathcal{F}$ is semi-algebra (see Exercise 6) and the family of all finite disjoint unions of rectangles is an algebra.

The same applies to the n-dimensional cube $\Omega = [0, 1]^n$.

Theorem 2.2 *Let $\mathcal{F}$ be a family of subsets of Ω that contains Ω. Then $\sigma(\mathcal{F})$ consists of all sets that can be obtained from the elements of $\mathcal{F}$ by using at most countable number of operations $\cap, \cup$ and $\setminus$.*

Proof. Denote by $\mathcal{F}'$ the family of all subsets of Ω, that can be obtain from elements of $\mathcal{F}$ by using at most countable number of operations $\cap, \cup$ and $\setminus$. Clearly,

$$\mathcal{F}' \subset \sigma(\mathcal{F}).$$

On the other hand, $\mathcal{F}'$ is a σ-algebra because applying those operations on the sets from $\mathcal{F}'$, we obtain sets which can also be obtained from $\mathcal{F}$ by at most countable number of $\cap, \cup$ and $\setminus$. Since $\sigma(\mathcal{F})$ is the minimal σ-algebra containing $\mathcal{F}$, it follows that

$$\sigma(\mathcal{F}) \subset \mathcal{F}',$$

whence $\mathcal{F}' = \sigma(\mathcal{F})$. ■

Remark We should warn that the term "at most countable number of operations" is somewhat ambiguous (in contrast to "at most finite number of operations" which can be defined inductively). Its rigorous meaning requires considering *orders* on countable sets of operations (because the operations are performed in certain orders), and countable sets may be ordered in many non-isomorphic ways. For example, consider the following two orders of a countable set:

$$1, 2, 3, ...[\text{all positive integers}] \tag{2.1}$$

and

$$1, 2, 3, ..[\text{all positive integers}], \mathbf{1}, \mathbf{2}, \mathbf{3}, ...[\textbf{all bold positive integers}]. \tag{2.2}$$

They are non-isomorphic because the bold **1** in (2.2) is preceded by infinite many elements, whereas in (2.1) no elements possesses this property.

Orders of countable sets are called *transfinite numbers*. A careful description of the term "at most countable number of operations" requires the theory of transfinite numbers, in particular, the transfinite induction principle that extends the usual induction principle for integers.

Lecture 3
20.09.10

2.2 Extension of measures

Important problem in the measure theory is extension of measures from semi-algebras to σ-algebras. Let μ be a (countably additive) measure on a semi-algebra $\mathcal{F}$. Let us extend μ to $a(\mathcal{F})$ as follows. Given $A \in a(\mathcal{F})$, by Theorem 2.1, it can be represented in the form

$$A = \bigsqcup_{i=1}^{n} A_i$$

where $A_i \in \mathcal{F}$. Then define

$$\mu(A) = \sum_{i=1}^{n} \mu(A_i). \tag{2.3}$$

Theorem 2.3 *If μ is a countably additive measure on a semi-algebra $\mathcal{F}$ then its extension* (2.3) *to $a(\mathcal{F})$ is well-defined and is also countably additive. Moreover, if μ is σ-additive on $\mathcal{F}$ then μ is σ-additive on $a(\mathcal{F})$ as well.*

Proof. The first part is the contents of Exercise 7. Let us prove the second part, that is, the σ-additivity of μ on $a(\mathcal{F})$. Let $A = \bigsqcup_{l=1}^{\infty} B_l$ where $A, B_l \in a(\mathcal{F})$. We need to prove that

$$\mu(A) = \sum_{l=1}^{\infty} \mu(B_l). \tag{2.4}$$

Represent the given sets in the form $A = \bigsqcup_k A_k$ and $B_l = \bigsqcup_m B_{lm}$ where the summations in k and m are finite and the sets A_k and B_{lm} belong to $\mathcal{F}$. Set also

$$C_{klm} = A_k \cap B_{lm}$$

and observe that $C_{klm} \in \mathcal{F}$. Also, we have

$$A_k = A_k \cap A = A_k \cap \bigsqcup_{l,m} B_{lm} = \bigsqcup_{l,m} (A_k \cap B_{lm}) = \bigsqcup_{l,m} C_{klm}$$

and

$$B_{lm} = B_{lm} \cap A = B_{lm} \cap \bigsqcup_{k} A_k = \bigsqcup_{k} (A_k \cap B_{lm}) = \bigsqcup_{k} C_{klm}.$$

By the σ-additivity of μ on $\mathcal{F}$, we obtain

$$\mu(A_k) = \sum_{l,m} \mu(C_{klm})$$

and

$$\mu(B_{lm}) = \sum_{k} \mu(C_{klm}).$$

On the other hand, we have by definition of μ on $a(\mathcal{F})$ that

$$\mu(A) = \sum_{k} \mu(A_k)$$

and

$$\mu(B_l) = \sum_{m} \mu(B_{lm})$$

Combining the above lines, we obtain

$$\mu(A) = \sum_{k} \mu(A_k) = \sum_{k,l.m} \mu(C_{klm}) = \sum_{l,m} \mu(B_{lm}) = \sum_{l} \mu(B_l),$$

which finishes the proof. ■

The next step is extension from an algebra to σ-algebra. It is covered by the following deep theorem which belongs to measure theory courses.

Theorem 2.4 (Carathéodory's extension theorem) *Let μ be a σ-additive measure on an algebra $\mathcal{F}$. Then it can be uniquely extended to a σ-additive measure on $\sigma(\mathcal{F})$.*

We do not prove the existence part here. The uniqueness part will be proved below after introducing necessary tools.

Clearly, if μ has the total mass 1 then this is preserved by any extension. Hence, we obtain the following corollary.

Corollary 2.5 *If μ is a probability measure on a semi-algebra $\mathcal{F}$ then it can be uniquely extended to a probability measure on $\sigma(\mathcal{F})$.*

2.3 Equivalent definitions of σ-additivity

Let $\mathcal{F}$ be a family of subsets of a set Ω and μ be non-negative function on $\mathcal{F}$.

Definition. We say that μ is *σ-subadditive* if whenever $A \subset \bigcup_{k=1}^{n} A_k$ where A and all A_k are elements of $\mathcal{F}$ and n is either finite or infinite, then

$$\mu(A) \leq \sum_{k=1}^{n} \mu(A_k). \tag{2.5}$$

If this property is true only for finite n then μ is called finitely subadditive.

Theorem 2.6 *Let $\mathcal{F}$ be a semi-algebra and μ be a finitely additive measure on $\mathcal{F}$.*

(*a*) *μ is finitely subadditive.*

(*b*) *μ is σ-additive if and only if it is σ-subadditive.*

Proof. (*a*) By Theorem 2.3 measure μ can be extended to the algebra $a(\mathcal{F})$ so that μ is finitely additive on $a(\mathcal{F})$. Let $A \subset \bigcup_{k=1}^{n} A_k$ where $A, A_k \in \mathcal{F}$ and n is finite. Consider the sets $\{B_k\}_{k=1}^{n}$ defined by

$$B_k = (A_k \cap A) \setminus (A_1 \cup ... \cup A_{k-1}). \tag{2.6}$$

Clearly, $B_k \in a(\mathcal{F})$, $B_k \subset A_k$, the sequence $\{B_k\}$ is disjoint, and

$$A = \bigsqcup_{k=1}^{n} B_k. \tag{2.7}$$

Indeed, the inclusion $A \supset \bigsqcup_{k=1}^{n} B_k$ is obvious. To prove the opposite inclusion, consider an element $\omega \in A$ and let k be the smallest index such that $\omega \in A_k$. Then we see from (2.6) that also $\omega \in B_k$ and, hence, $\omega \in \bigsqcup_{k=1}^{\infty} B_k$, which proves (2.7). By the finite additivity of μ on $a(\mathcal{F})$, we obtain

$$\mu(A) = \sum_{k=1}^{n} \mu(B_k).$$

On the other hand, $\mu(A_k) = \mu(A_k \setminus B_k) + \mu(B_k)$ whence $\mu(B_k) \leq \mu(A_k)$ and (2.5) follows.

(*b*) The fact that σ-additivity implies σ-subadditivity is proved exactly in the same way as above by replacing everywhere a finite n by $n = \infty$.

Let us show that σ-subadditivity implies σ-additivity. Let A and $\{A_k\}_{k=1}^{\infty}$ be elements of $\mathcal{F}$ such that $A = \bigsqcup_{k=1}^{\infty} A_k$, and let us prove that

$$\mu(A) = \sum_{k=1}^{\infty} \mu(A_k).$$

The upper bound

$$\mu(A) \leq \sum_{k=1}^{\infty} \mu(A_k)$$

holds by the σ-additivity of μ. To prove the lower bound, it suffices to show that, for any positive integer n,

$$\mu(A) \geq \sum_{k=1}^{n} \mu(A_k).$$

Since the sets A and $\bigsqcup_{k=1}^{n} A_k$ are the elements of $a(\mathcal{F})$ and

$$A \supset \bigsqcup_{k=1}^{n} A_k,$$

we have

$$\mu(A) \geq \mu\left(\bigsqcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} \mu(A_k),$$

where in the last equality we have used the finite additivity of μ. ■

Definition. We say that a sequence of sets $\{A_n\}$ is *monotone increasing* if $A_n \subset A_{n+1}$ for all n, and *monotone decreasing* if $A_n \supset A_{n+1}$ for all n.

Theorem 2.7 *Let $\mathcal{F}$ be a σ-algebra and μ be a finitely additive measure on $\mathcal{F}$. The following conditions are equivalent.*

(a) *μ is σ-additive.*

(b) *For any monotone increasing sequence $\{A_n\}$ of elements of $\mathcal{F}$,*

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

(c) *For any monotone decreasing sequence $\{A_n\}$ of elements of $\mathcal{F}$,*

$$\mu\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n).$$

(d) *For any monotone decreasing sequence $\{A_n\}$ of elements of $\mathcal{F}$ such that $\bigcap_{n=1}^{\infty} A_n = \emptyset$,*

$$\lim_{n\to\infty} \mu(A_n) = 0.$$

Definition. A finitely additive measure μ on a σ-algebra $\mathcal{F}$ is called *continuous* if it satisfies one of the equivalent properties $(b), (c), (d)$.

Hence, a finitely additive measure is σ-additive if and only if it is continuous.

If $\{A_n\}$ is a monotone sequence of sets then define its limit by

$$\lim A_n = \begin{cases} \bigcup_n A_n, \text{ if } \{A_n\} \text{ is monotone increasing.} \\ \bigcap_n A_n, \text{ if } \{A_n\} \text{ is monotone decreasing.} \end{cases}$$

The the continuity property of μ can be stated also as follows: for any monotone sequence of sets $\{A_n\} \subset \mathcal{F}$,

$$\mu\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} \mu\left(A_n\right),$$

which justifies this terminology.

Proof. $(a) \Rightarrow (b)$ Denote $A = \bigcup_{n=1}^{\infty} A_n$ and observe that

$$A = A_1 \cup (A_2 \setminus A_1) \cup ... \cup (A_n \setminus A_{n-1}) \cup ... = \bigsqcup_{n=1}^{\infty} (A_n \setminus A_{n-1}),$$

where $A_0 = \emptyset$. By the σ-additivity of μ, we have

$$\mu(A) = \sum_{n=1}^{\infty} \mu(A_n \setminus A_{n-1}) = \sum_{n=1}^{\infty} (\mu(A_n) - \mu(A_{n-1})) = \lim_{n\to\infty} \mu(A_n).$$

$(b) \Rightarrow (c)$ Let $A^c = \bigcap_{n=1}^{\infty} A_n$. The sequence $\{A_n^c\}$ is monotone increasing and

$$A^c = \bigcup_n A_n^c$$

whence

$$\mu(A^c) = \lim_{n\to\infty} \mu(A_n^c).$$

It follows that

$$\mu(A) = \mu(\Omega) - \mu(A^c) = \lim_{n\to\infty} (\mu(\Omega) - \mu(A_n^c)) = \lim_{n\to\infty} \mu(A_n).$$

$(c) \Rightarrow (d)$ This is trivial because

$$\lim_{n\to\infty} \mu(A_n) = \mu\left(\bigcap_{n=1}^{\infty} A_n\right) = \mu(\emptyset) = 0.$$

$(d) \Rightarrow (a)$ Let $\{A_n\}$ be a sequence of disjoint sets from $\mathcal{F}$ and set

$$A = \bigcup_{k=1}^{\infty} A_k.$$

Consider the sets

$$B_n = A \setminus \bigcup_{k=1}^{n} A_k = \bigcup_{k=n+1}^{\infty} A_k,$$

so that $\{B_n\}$ is a decreasing sequence and $\bigcap_n B_n = \emptyset$. Then we have $\mu(B_n) \to 0$ as $n \to \infty$. On the other hand, since A is a disjoint union of B_n and the sets $A_1, ..., A_n$, we obtain by the finite additivity of μ that

$$\mu(A) = \mu(B_n) + \sum_{k=1}^{n} \mu(A_k),$$

whence it follows that

$$\mu(A) = \lim_{n\to\infty} \sum_{k=1}^{n} \mu(A_k) = \sum_{k=1}^{\infty} \mu(A_k).$$

■

2.4 Monotone class theorem

Given any family $\mathcal{F}$ of subsets of Ω and an operation $*$ on subsets of Ω, denote $\mathcal{F}^*$ the minimal extension of $\mathcal{F}$, which is closed under $*$. More precisely, consider all families of subsets containing $\mathcal{F}$ and closed under $*$. For example, the family 2^Ω of all subsets of Ω satisfies always this condition. Then $\mathcal{F}^*$ is the intersection of all such families. If $*$ is an operation over a finite number of elements then $\mathcal{F}^*$ can be obtained from $\mathcal{F}$ by applying all possible finite sequences of operations $*$.

The operations to which we apply this notion are the following:

1. Intersection "$\cap$" of two sets, that is $A, B \mapsto A \cap B$
2. Monotone difference "$-$" defined as follows: if $A \supset B$ then $A - B = A \setminus B$.
3. Monotone limit lim defined on sequences $\{A_n\}_{n=1}^\infty$ as follows: if the sequence is increasing, that is $A_{n+1} \supset A_n$ then

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

and if A_n is decreasing that is $A_{n+1} \subset A_n$ then

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

Theorem 2.8 (Monotone Class Theorem of Dynkin)

(*a*) *If $\mathcal{F}$ contains Ω and is closed under $\cap$ (for example, if $\mathcal{F}$ is a semi-algebra) then $a(\mathcal{F}) = \mathcal{F}^-$.*

(*b*) *If $\mathcal{F}$ is algebra then $\sigma(\mathcal{F}) = \mathcal{F}^{\lim}$.*

Lecture 4
21.09.10

As a consequence we see that if $\mathcal{F}$ is any family of subsets containing Ω then

$$a(\mathcal{F}) = (\mathcal{F}^\cap)^- \tag{2.8}$$

$$\sigma(\mathcal{F}) = \left((\mathcal{F}^\cap)^-\right)^{\lim}. \tag{2.9}$$

Indeed, $\mathcal{F}^\cap$ satisfies the hypotheses of Theorem 2.8, whence (2.8) and (2.9) follow by successive application of two parts of the theorem.

The most non-trivial part is (2.8). Indeed, it says that any subset that can be obtained from $\mathcal{F}$ by a finite number of operations $\cap, \cup, \setminus$, can also be obtained by first applying a finite number of $\cap$ and then applying finite number of "$-$". This is not quite obvious even for the simplest case

$$\mathcal{F} = \{\Omega, A, B\}.$$

Then (2.8) implies that the union $A \cup B$ can be obtained from the elements of $\mathcal{F}^{\cap} = \{\Omega, A, B, A \cap B\}$ by applying only monotone difference "$-$". However, Theorem 2.8 does not say how exactly one can do that. The answer in this particular case is

$$A \cup B = \Omega - ((\Omega - A) - (B - A \cap B))$$

(see Exercise 8 for further details).

Proof. (a) Assuming that $\mathcal{F}$ contains Ω and is closed under $\cap$, let us show that $\mathcal{F}^{-}$ is algebra, which will settle the claim. Indeed, as an algebra. $\mathcal{F}^{-}$ must contain $a(\mathcal{F})$. On the other hand, $a(\mathcal{F})$ is closed under "$-$" so it contains $\mathcal{F}^{-}$, whence $a(\mathcal{F}) = \mathcal{F}^{-}$.

Clearly, $\Omega \in \mathcal{F}^{-}$ and $\emptyset = \Omega - \Omega \in \mathcal{F}^{-}$. If $A \in \mathcal{F}^{-}$ then also $A^c \in \mathcal{F}^{-}$ because $A^c = \Omega - A$ is a monotone difference of sets from $\mathcal{F}^{-}$. We are left to show that $\mathcal{F}^{-}$ is closed under intersection, that is, to prove that

$$A, B \in \mathcal{F}^{-} \;\Rightarrow\; A \cap B \in \mathcal{F}^{-}. \tag{2.10}$$

Assume first that $A \in \mathcal{F}$ and define the family S of *suitable* subsets B as follows:

$$S = \left\{B \subset \Omega : A \cap B \in \mathcal{F}^{-}\right\}$$

(where A is considered as fixed). Clearly, S contains $\mathcal{F}$ because $\mathcal{F}$ is closed under intersections. Let us verify that S closed under monotone difference. Indeed, if $B_1, B_2 \in S$ and $B_1 \subset B_2$ then

$$A \cap (B_2 - B_1) = (A \cap B_2) - (A \cap B_1) \in \mathcal{F}^{-}$$

so that $B_2 - B_1 \in S$. Hence, S contains $\mathcal{F}$ and is closed under "$-$" which implies that it contains $\mathcal{F}^{-}$, that is, (2.10) holds under the additional assumption that $A \in \mathcal{F}$.

Now let us drop this assumption. Fix $B \in \mathcal{F}^{-}$ and consider a new family of suitable sets A:

$$S = \left\{A \subset \Omega : A \cap B \in \mathcal{F}^{-}\right\}.$$

As we have just proved, S contains $\mathcal{F}$ and is closed under "$-$". Therefore, S contains $\mathcal{F}^{-}$, which proves (2.10) in full generality.

(b) We have $\sigma(\mathcal{F}) \supset \mathcal{F}^{\lim}$ because $\sigma(\mathcal{F})$ is closed under countable unions and intersections.The opposite inclusion will follow if we prove that $\mathcal{F}^{\lim}$ is a σ-algebra. Let us first prove that $\mathcal{F}^{\lim}$ is an algebra. That $\Omega, \emptyset \in \mathcal{F}^{\lim}$ follows from $\Omega, \emptyset \in \mathcal{F}$.

Let us prove that

$$A \in \mathcal{F}^{\lim} \Rightarrow A^c \in \mathcal{F}^{\lim}. \tag{2.11}$$

Consider the family of suitable sets

$$S = \left\{A \in \Omega : A^c \in \mathcal{F}^{\lim}\right\}.$$

Then $S \supset \mathcal{F}$ and S is closed under lim because if $A = \lim_{n\to\infty} A_n$ and $A_n \in S$ then $A_n^c \in \mathcal{F}^{\lim}$ and

$$A^c = \lim A_n^c \in \mathcal{F}^{\lim}.$$

Hence, $S \supset \mathcal{F}^{\lim}$, which proves (2.11).

Let us prove that

$$A, B \in \mathcal{F}^{\lim} \Rightarrow A \cap B \in \mathcal{F}^{\lim}. \tag{2.12}$$

Fix first $A \in \mathcal{F}$ and consider the family of suitable sets

$$S = \left\{ B \subset \Omega : A \cap B \in \mathcal{F}^{\lim} \right\}.$$

Then $S \supset \mathcal{F}$ and let us show that S is closed under $\lim$. If $B = \lim B_n$ where $B_n \in S$, then $A \cap B_n \in \mathcal{F}^{\lim}$ and

$$A \cap B = \lim (A \cap B_n) \in \mathcal{F}^{\lim},$$

so that $B \in S$. Hence, S contains $\mathcal{F}^{\lim}$, which proves (2.12) under the additional assumption $A \in \mathcal{F}$.

Let us now drop the assumption $A \in \mathcal{F}$. For any fixed $B \in \mathcal{F}^{\lim}$ consider the following family suitable sets A:

$$S = \left\{ A \subset \Omega : A \cap B \in \mathcal{F}^{\lim} \right\}.$$

By the above argument, we have $S \supset \mathcal{F}$ and S is closed under lim, whence $S \supset \mathcal{F}^{\lim}$, which finishes the proof of (2.12).

Hence, $\mathcal{F}^{\lim}$ is an algebra, and we are left to show that $\mathcal{F}^{\lim}$ is a σ-algebra. The latter follows from the next lemma.

Lemma 2.9 *If an algebra $\mathcal{A}$ is closed under monotone limits then $\mathcal{A}$ is a σ-algebra (in other words, an algebra is a σ-algebra if and only if it is closed under monotone* lim*).*

It suffices to prove that if $\{A_n\}_{n=1}^{\infty}$ if a sequence of elements of $\mathcal{A}$, then

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

Indeed, consider the sets

$$B_n = \bigcap_{i=1}^{n} A_i$$

and observe that $B_n \in \mathcal{A}$ as a finite intersection of elements of $\mathcal{A}$. Since the sequence $\{B_n\}$ is monotone decreasing, we obtain

$$\bigcap_{n=1}^{\infty} A_n = \bigcap_{n=1}^{\infty} B_n = \lim_{n \to \infty} B_n \in \mathcal{A},$$

which was to be proved. ■

Theorem 2.8 has numerous applications. For example, it is used in the following proof.

Proof of the uniqueness in Theorem 2.4. Let μ' and μ'' be two σ-additive extensions of measure μ to $\sigma(\mathcal{F})$, and let us prove that $\mu' = \mu''$, that is, $\mu'(A) = \mu''(A)$ for all $A \in \sigma(\mathcal{F})$. Consider the following family of suitable sets

$$S = \{ A \in \sigma(\mathcal{F}) : \mu'(A) = \mu''(A) \}.$$

By hypothesis, we have $S \supset \mathcal{F}$ because for any $A \in \mathcal{F}$

$$\mu'(A) = \mu(A) = \mu''(A).$$

We need to show that $S \supset \sigma(\mathcal{F})$, and for that it suffices to verify that S is a σ-algebra. Observe that S is closed under monotone limit. Indeed, if $\{A_n\}$ is a monotone sequence from S and $A = \lim A_n$ then by Theorem 2.7

$$\mu'(A) = \lim \mu'(A_n) = \lim \mu''(A_n) = \mu''(A).$$

Then we obtain by Theorem 2.8

$$S = S^{\lim} \supset \mathcal{F}^{\lim} = \sigma(\mathcal{F}),$$

which finishes the proof. ■

2.5 Measures on $\mathbb{R}$

Let $\mathcal{F}$ be the family of all intervals on $\mathbb{R}$. Then the σ-algebra $\sigma(\mathcal{F})$ is called the *Borel σ-algebra* of $\mathbb{R}$ and is denoted by $\mathcal{B}(\mathbb{R})$. The elements of $\mathcal{B}(\mathbb{R})$ are called *Borel sets*. The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ contains all open and closed subsets of $\mathbb{R}$ (Exercise 2) as well as all possible their countable unions and intersections.

The necessity of considering the Borel sets appears naturally in measure theory, in function theory and in probability theory. Although the structure a particular Borel set may be rather complicated, we will never need to consider fancy Borel sets. What we need is the fact that all "nice" sets as intervals, open sets and closed sets can be considered as elements of this σ-algebra.

Lecture 5
27.09.10

Our purpose here is to describe all probabilities measures on $\mathcal{B}(\mathbb{R})$. For any probability measure μ on $\mathcal{B}(\mathbb{R})$, define its *distribution function* $F(x)$ by the identity

$$F(x) = \mu(-\infty, x], \quad x \in \mathbb{R}. \tag{2.13}$$

Theorem 2.10 *If $F(x)$ is the distribution function of a probability measure μ on $\mathcal{B}(\mathbb{R})$ then F possesses the following properties:*

(*i*) *F is monotone increasing;*

(*ii*) *$F(-\infty) = 0$ and $F(+\infty) = 1$ (in the sense of limits);*

(*iii*) *F is right continuous.*

Moreover, if F is a function satisfying (i)-(iii) then F is the distribution function of a unique probability measure on $\mathcal{B}(\mathbb{R})$.

Any function F satisfying (*i*)-(*iii*) is called *a distribution function*. Before the proof consider some examples of distribution functions.

Absolutely continuous measures.

Let f be a continuous or piecewise continuous function on $\mathbb{R}$ such that $f(x) \geq 0$ and

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1. \tag{2.14}$$

Define function F by

$$F(x) = \int_{-\infty}^{x} f(y)d(y). \tag{2.15}$$

Then F satisfies the conditions (i)-(iii) and, hence, is a distribution function.

Any measure μ that has a distribution function of type (2.15) is called *absolutely continuous*. The function $f(x)$ is called the *density* of μ.

By (2.13) we have for all $a \leq b$

$$\mu((a,b]) = \mu((-\infty,b]) - \mu((-\infty,a]) = F(b) - F(a) = \int_a^b f(x)\,dx.$$

In particular, $\mu(\{a\}) = 0$ whence it follows that

$$\mu((a,b]) = \mu([a,b]) = \mu((a,b)) = \mu([a,b)) = \int_a^b f(x)\,dx. \tag{2.16}$$

Consider some specific examples.

Example. Let

$$f(x) = \begin{cases} 1,\ x \in [0,1], \\ 0,\ x \notin [0,1]. \end{cases}$$

Then measure μ is supported on $[0,1]$ because $\mu(\mathbb{R} \setminus [0,1]) = 0$. For any interval $I \subset [0,1]$ with endpoints $a < b$, we have by (2.16)

$$\mu(I) = b - a.$$

This measure μ is called *Lebesgue measure* on $[0,1]$ and is denoted by λ. Clearly, λ is the unique extension of the notion of length from intervals in $[0,1]$ to the Borel σ-algebra $\mathcal{B}([0,1])$. The latter is the minimal σ-algebra that contains all intervals in $[0,1]$. It is possible to show that length cannot be extended to σ-algebra $2^{[0,1]}$ of all subsets of $[0,1]$ as a σ-additive measure.

Similarly, one can define Lebesgue measure λ on any interval $[a, a+1]$ using the density function

$$f(x) = \begin{cases} 1,\ x \in [a,a+1], \\ 0,\ x \notin [a,a+1]. \end{cases}$$

Then one extends λ to all Borel subsets of $\mathbb{R}$ by setting

$$\lambda(A) = \sum_{n=-\infty}^{+\infty} \lambda(A \cap [n,n+1]).$$

One can show that this extension of λ is a measure on $\mathcal{B}(\mathbb{R})$ with values in $[0, +\infty]$ (but not a probability measure). It is also referred to as Lebesgue measure.

Example. Consider the function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

(see Fig. 2.1).

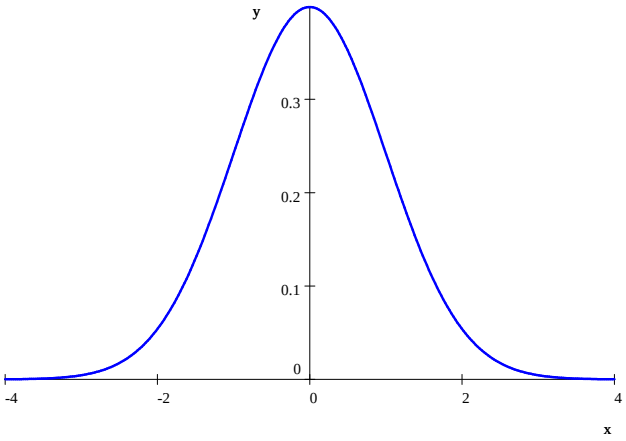


Figure 2.1: Function $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

It is known from Analysis that

$$\int_{-\infty}^{+\infty} e^{-x^2/2} = \sqrt{2\pi}$$

so that this function f satisfies (2.14). The measure μ with the density $f(x)$ is called the *Gauss measure* on $\mathbb{R}$. For any interval I with endpoints $a < b$, we have

$$\mu(I) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Discrete measures.

Fix a (finite or countable) sequence $\{x_i\}$ of distinct reals and a stochastic sequence $\{p_i\}$. As we know from Section 1.4, the stochastic sequence defines a probability measure μ on the set $\{x_i\}$ by

$$\mu(\{x_i\}) = p_i,$$

and this measure extends to a probability measure on all subsets of $\mathbb{R}$ by

$$\mu(A) = \sum_{\{i: x_i \in A\}} p_i. \tag{2.17}$$

In particular, μ is defined for all Borel sets $A \subset \mathbb{R}$. Setting here $A = (-\infty, x]$, we obtain the distribution function of this measure

$$F(x) = \mu\left((-\infty, x]\right) = \sum_{x_i \le x} p_i \,. \tag{2.18}$$

This function F jumps at x_i by p_i, and stays constant otherwise.

Any measure μ defined by (2.17) is called a *discrete measure* and the values of x_i are called the *atoms* of this measure. Obviously, μ is supported in the set $\{x_i\}$.

Singular measures.

There is a third type of measures whose distribution function is given neither by (2.15) nor by (2.18). We give an example of such a distribution function called the *Cantor function*. Outside interval $[0,1]$ the Cantor function takes the values $F(x) = 0$ if $x < 0$ and $F(x) = 1$ if $x > 1$. Inside $[0,1]$ $F(x)$ is defined as the limit of a sequence of continuous piecewise linear functions $\{F_n\}_{n=0}^{\infty}$ as follows. Set $F_0(x) = x$. If F_n is already constructed then define F_{n+1} by modifying F_n on all maximal intervals where F_n is non-constant. Namely, if $[a, \beta]$ is such an interval then set F_{n+1} to be equal to the $\frac{1}{2}(F_n(\alpha) + F_n(\beta))$ in the middle third of $[\alpha, \beta]$, and to be equal to F_n at the endpoints α and β. In the other two thirds of $[\alpha, \beta]$, define F_{n+1} by linear interpolation. Obviously, each function F_n is continuous, monotone increasing, and $F_n(0) = 0$, $F_n(1) = 1$. It is possible to prove that the sequence $\{F_n\}$ converges uniformly to a distribution function F, and the corresponding measure μ is neither absolutely continuous, nor discrete.

Let us show that the sequence $\{F_n\}$ is uniformly convergent as $n \to \infty$. Denote by $[\alpha, \beta]$ an interval, where F_n is non-constant and where the difference $F_n(\beta) - F_n(\alpha)$ is maximal possible; set

$$a_n = F_n(\beta) - F_n(\alpha) \,.$$

Then by the above construction $a_0 = 1$ and

$$a_{n+1} = \frac{1}{2}(F_n(\alpha) + F_n(\beta)) - F_n(\alpha) = \frac{1}{2} a_n,$$

whence it follows that $a_n = 2^{-n}$. On the other hand, we have

$$\max|F_n - F_{n+1}| = \frac{1}{3}F_n(\alpha) + \frac{2}{3}F_n(\beta) - \frac{1}{2}(F_n(\alpha) + F_n(\beta)) = \frac{1}{6}(F_n(\beta) - F_n(\alpha)) = \frac{1}{6}a_n < 2^{-n}.$$

It follows that for all $m > n$

$$\begin{aligned} \max|F_n - F_m| &\le \max|F_n - F_{n+1}| + \max|F_{n+1} - F_{n+2}| + ... + \max|F_{m-1} - F_m| \\ &\le 2^{-n} + 2^{-(n+1)} + ... + 2^{-(m-1)} \\ &< 2^{-(n-1)}, \end{aligned}$$

whence $\max|F_n - F_m| \to 0$ as $n, m \to \infty$. Hence, the sequence $\{F_n\}$ is Cauchy and converges uniformly to a function F on $[0,1]$ that is hence continuous and monotone increasing.

Denote by U_n the union of all intervals in $[0,1]$ where F_n is constant. By construction

$$U_0 = \emptyset,\ U_1 = \left[\frac{1}{3}, \frac{2}{3}\right],\ U_2 = \left[\frac{1}{9}, \frac{2}{9}\right] \cup \left[\frac{1}{3}, \frac{2}{3}\right] \cup \left[\frac{7}{9}, \frac{8}{9}\right], ...$$

and $U_n \subset U_{n+1}$ for all n. Set $l_n = 1 - \lambda(U_n)$, that is, l_n is the total length of the intervals where F_n is non-constant. By construction, we have

$$l_{n+1} = \frac{2}{3} l_n,$$

whence $l_n = (2/3)^n \to 0$ as $n \to \infty$. It follows that $\lambda(U_n) \to 1$ as $n \to \infty$ and, hence, the limit function $F(x)$ stays constant on the set $U = \bigcup_{n=1}^{\infty} U_n$ of the Lebesgue measure 1. Nevertheless, the function F is continuous and changes it values from 0 to 1 on the set $C = [0,1] \setminus U$ of Lebesgue measure 0. In fact, C is the Cantor set, and the corresponding probability measure μ is supported on the Cantor set. Measure μ is not discrete, because F has no jumps, and is not absolutely continuous because $\mu(C) = 1$ while $\lambda(C) = 0$.

Proof of Theorem 2.10. Let F be a distribution function of a measure μ. For all $a < b$ we have

$$F(a) = \mu(-\infty, a] \leq \mu(-\infty, b] = F(b),$$

so that F is monotone increasing. Using the continuity of the measure, we obtain, for any sequence $x_n \downarrow -\infty$,

$$F(-\infty) = \lim_{n\to\infty} \mu(-\infty, x_n] = \mu\left(\bigcap_n (-\infty, x_n]\right) = \mu(\emptyset) = 0.$$

Similarly, for any sequence $x_n \uparrow +\infty$,

$$F(+\infty) = \lim_{n\to\infty} \mu(-\infty, x_n] = \mu\left(\bigcup_n (-\infty, x_n]\right) = \mu(\mathbb{R}) = 1.$$

Finally, let us prove the right continuity of F, that is,

$$\lim_{x_n \downarrow a} F(x_n) = F(a).$$

Indeed, we have

$$\lim_{x_n \downarrow a} F(x_n) = \lim_{x_n \downarrow a} \mu(-\infty, x_n] = \mu\left(\bigcap_n (-\infty, x_n]\right) = \mu(-\infty, a] = F(a).$$

Now let us prove that for any distribution function F there exists a unique probability measure μ on $\mathcal{B}(\mathbb{R})$ such that

$$\mu(-\infty, x] = F(x). \tag{2.19}$$

Let $\mathcal{F}$ be the family of all semi-closed intervals of the form $(a, b]$ in $\mathbb{R}$ where $a, b \in [-\infty, +\infty]$. It is easy to check that $\mathcal{F}$ is a semi-algebra. If μ satisfies (2.19) then it follows that for any interval $(a, b]$

$$\mu(a, b] = \mu(-\infty, b] - \mu(-\infty, a] = F(b) - F(a)$$

so that μ is uniquely defined on $\mathcal{F}$. By the uniqueness part of Corollary 2.5, μ is uniquely defined on $\sigma(\mathcal{F}) = \mathcal{B}(\mathbb{R})$.

To prove the existence of μ, first define μ on the intervals from $\mathcal{F}$ by

$$\mu(a, b] = F(b) - F(a)$$

and check that μ is indeed a probability measure on $\mathcal{F}$. Then by the existence part of Corollary 2.5, μ extends to a probability measure on $\sigma(\mathcal{F}) = \mathcal{B}(\mathbb{R})$.

Hence, let us prove that μ is a probability measure on $\mathcal{F}$. The value $\mu(a, b]$ is non-negative by the monotonicity of F. The total mass is 1 since

$$\mu(\mathbb{R}) = F(+\infty) - F(-\infty) = 1.$$

Let us prove that μ is σ-additive on $\mathcal{F}$. By Theorem 2.6, it suffices to prove that μ is finitely additive and σ-subadditive.

Let us first show that μ is finitely additive. Let an interval $I \subset \mathcal{F}$ be a disjoint union of a finite number of intervals $I_k \subset \mathcal{F}$, $k = 1, ..., n$. Let $\{a_i\}_{i=0}^N$ be the set of all distinct endpoints of all the intervals $I, I_1, ..., I_n$ enumerated in an increasing order. Then $I = (a_0, a_N]$ while each interval I_k has necessarily the form $(a_{i-1}, a_i]$ for some i. Indeed, if $I_k = (a_j, a_i]$ with $j < i - 1$ then $a_{i-1} \in I_k$, which means that I_k must intersect with some other interval I_m. Conversely, any interval $(a_{i-1}, a_i]$ coincides with some interval I_k. Indeed, the point a_i must be covered by some interval I_k; since $I_k = (a_{j-1}, a_j]$ for some j, it follows that $a_i = a_j$ and, hence, $I_k = (a_{i-1}, a_i]$. Hence, the intervals $I_1, ..., I_n$ are in fact $(a_0, a_1], ...(a_{N-1}, a_N]$ (and $n = N$) whence it follows that

$$\sum_{i=1}^{N} \mu(I_i) = \sum_{i=1}^{N} (F(a_i) - F(a_{i-1})) = F(a_N) - F(a_0) = \mu(I).$$

We are left to show that μ is σ-subadditive. Assume that

$$(a, b] \subset \bigcup_{k=1}^{\infty} (a_k, b_k]$$

and prove that

$$\mu(a, b] \leq \sum_{k=1}^{\infty} \mu(a_k, b_k]. \tag{2.20}$$

If $b = +\infty$ then it suffices to prove this inequality for any finite value of b, so we assume in the sequel that b is finite. Replacing b_k by $\min\{b, b_k\}$, we can assume that all b_k are also finite.

Fix some $\varepsilon > 0$ and using the right continuity of F choose $a' \in (a, b)$ such that $F(a') < F(a) + \varepsilon$, whence

$$\mu(a, b] < \mu(a', b] + \varepsilon.$$

Similarly, choose $b'_k > b_k$ so that

$$F(b'_k) < F(b) + \varepsilon/2^k,$$

whence

$$\mu(a_k, b'_k] < \mu(a_k, b_k] + \varepsilon/2^k.$$

We have then

$$[a', b] \subset (a, b] \subset \bigcup_{k=1}^{\infty} (a_k, b_k] \subset \bigcup_{k=1}^{\infty} (a_k, b'_k),$$

that is, the bounded closed interval $[a', b]$ is covered by a sequence $\{(a_k, b'_k)\}$ of open intervals. By the Borel-Lebesgue lemma, there is a finite subsequence of such intervals that also covers $[a', b]$, say

$$[a', b] \subset \bigcup_{i=1}^{N} \left(a_{k_i}, b'_{k_i}\right)$$

for some finite N. It follows that also

$$(a', b] \subset \bigcup_{i=1}^{N} (a_{k_i}, b'_{k_i}]$$

By Theorem 2.6, the finitely additive measure μ is finitely subadditive. It follows that

$$\mu(a', b] \leq \sum_{i=1}^{N} \mu(a_{k_i}, b'_{k_i}] \leq \sum_{k=1}^{\infty} \mu(a_k, b_k],$$

whence

$$\mu(a, b] \leq \varepsilon + \mu(a', b] \leq \varepsilon + \sum_{k=1}^{\infty} \mu(a_k, b'_k] \leq \varepsilon + \sum_{k=1}^{\infty} \left(\mu(a_k, b_k] + \frac{\varepsilon}{2^k}\right) = 2\varepsilon + \sum_{k=1}^{\infty} \mu(a_k, b_k].$$

Since $\varepsilon > 0$ is arbitrary, we obtain (2.20) by letting $\varepsilon \to 0$. ■

Chapter 3

Probability spaces

Lecture 6
28.09.10

3.1 Events

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we can assign the exact meaning to the notion of (a random) event. Recall that an *elementary event* is any element ω from Ω. An elementary event can be identified with the outcome of a particular series of trials. An *event* is any element from $\mathcal{F}$, that is a subset A of Ω, that belongs to the σ-algebra $\mathcal{F}$. The probability of the event A is the value $\mathbb{P}(A)$. The elements of $\mathcal{F}$ are also called $\mathbb{P}$-measurable subsets of Ω to emphasize the fact that the value $\mathbb{P}(A)$ is defined only for $A \in \mathcal{F}$.

The fact that an event A occurs in a given series of trials ω means that $\omega \in A$; that is

$$A \text{ occurs} \Leftrightarrow \omega \in A.$$

The logical operations on events correspond to the set theoretic operations on sets as follows:

$$\begin{aligned} (A \text{ and } B) &= A \cap B \\ (A \text{ or } B) &= A \cup B \\ (\text{not } A) &= A^c. \end{aligned}$$

Indeed, for example, we have

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

Example. Consider a series of n trials of coin flipping. In this case each elementary event ω is a sequence of letters H and T of length n, and Ω is the set of all such sequences, that is,

$$\Omega = \left\{\omega = \{\omega_i\}_{i=1}^n : \omega_i \in \{H, T\} \text{ for all } i = 1, ..., n\right\}.$$

Set $\mathcal{F} = 2^\Omega$. For example, the event "H occurs exactly k times" (where $k = 1, ..., n$ is given) is the set

$$A_k = \{\omega \in \Omega : \omega_i = H \text{ for exactly } k \text{ values of } i\}. \tag{3.1}$$

For example, if $n = 3$ then

$$A_2 = \{HHT, HTH, THH\}.$$

One can introduce a probability measure on Ω in different ways. For example, consider a uniform distribution on Ω (which corresponds to independent tossing of a fair coin, as we will see below). Since $|\Omega| = 2^n$, we obtain that for any $\omega \in \Omega$,

$$\mathbb{P}(\{\omega\}) = 2^{-n}.$$

Let us compute the probability of the above event (3.1). For that, we only need to evaluate the number of sequences $\omega \in \Omega$ where H occurs exactly k times. It is known from combinatorics that $|A_k| = \binom{n}{k}$. It follows that

$$\mathbb{P}(A_k) = \binom{n}{k} 2^{-n}.$$

Consider the event "the outcome of the k-th trial is T". Denote it by B_k, that is

$$B_k = \{\omega \in \Omega : \omega_k = T\}.$$

For example, if $n = 3$ then

$$B_2 = \{HTH, HTT, TTH, TTT\}.$$

Since in a sequence $\{\omega_1, ..., \omega_k, ..., \omega_n\} \in B_k$ every element except for ω_k can take two values and ω_k takes only one value, we obtain $|B_k| = 2^{n-1}$. Hence,

$$\mathbb{P}(B_k) = 2^{n-1}/2^n = \frac{1}{2}.$$

3.2 Conditional probability

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and B be an event with a positive probability. For any event A define the *conditional probability* of A with respect to B by

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

that is also referred to as "the probability of A given B".

Theorem 3.1 (*a*) *Let $\mathbb{P}(B) > 0$. Then the function $A \mapsto \mathbb{P}(A|B)$ is a probability measure on $\mathcal{F}$. Hence, $(\Omega, \mathcal{F}, \mathbb{P}(\cdot|B))$ is a probability space.*

(*b*) (Bayes' formula) *If A and B events with positive probability, then*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{3.2}$$

Furthermore, if $\{A_i\}$ *is a sequence of events that form a partition of* Ω*, that is, if* $\Omega = \bigsqcup_i A_i$*, then*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\,\mathbb{P}(A_j)}, \tag{3.3}$$

provided all the probabilities $\mathbb{P}(A_j)$ *and* $\mathbb{P}(B)$ *are positive.*

Proof. (a) Let us show that the conditional probability is σ-additive. If $\{A_k\}$ is a disjoint sequence of events then

$$\mathbb{P}\left(\bigcup A_k|B\right) = \frac{\mathbb{P}\left(\left(\bigcup A_k\right)\cap B\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\bigcup_k (A_k\cap B)\right)}{\mathbb{P}(B)} = \sum_k \frac{\mathbb{P}(A_k\cap B)}{\mathbb{P}(B)} = \sum_k \mathbb{P}(A_k|B).$$

The conditional probability is a probability measure since

$$\mathbb{P}(\Omega|B) = \frac{\mathbb{P}(\Omega\cap B)}{\mathbb{P}(B)} = 1.$$

(b) We have

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(A)}\frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \mathbb{P}(B|A)\,\frac{\mathbb{P}(A)}{\mathbb{P}(B)}, \tag{3.4}$$

which proves (3.2). If $\{A_i\}$ is a partition of Ω then

$$\mathbb{P}(B) = \sum_j \mathbb{P}(B\cap A_j) = \sum_j \mathbb{P}(B|A_j)\,\mathbb{P}_j(A). \tag{3.5}$$

Using (3.4) and (3.5), we obtain

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\,\mathbb{P}(A_j)}.$$

■

An example of a partition is a pair $\{A, A^c\}$. For this partition (3.3) becomes

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B|A)\,\mathbb{P}(A) + \mathbb{P}(B|A^c)\,\mathbb{P}(A^c)}. \tag{3.6}$$

The identities (3.2), (3.3), (3.6) are referred to as *Bayes' formulas*. As we have seen, its proof is very simple, but these formulas find numerous of application in applied probability. Let us consider an example.

Example. A treasure chest has been buried in one of n islands of an archipelago.

Someone tries to find the chest by digging on the islands. Due to environmental conditions, the probability to find the chest on the k-th island, given that the chest is buried on this island, is $p_k \in (0,1)$. The chest is buried equally likely in any of the islands. After digging on the first island the result of the search was negative. What is the probability that the chest is buried on k-th island?

An elementary event in this problem is any point ω on the archipelago (where the chest could be buried). The even A_k that the chest is buried on the k-th island is the set of all points ω on the k-th island. It is given that the events $A_1, ..., A_n$ form a partition of the whole space Ω and $\mathbb{P}(A_k) = \frac{1}{n}$ for all $k = 1, ..., n$. The event B that the chest was found consists of all points ω where one actually digs. It is given that

$$\mathbb{P}(B|A_k) = p_k.$$

The event N that the chest was not found on the first island is equal to

$$N = (B \cap A_1)^c.$$

The question is to evaluate $\mathbb{P}(A_k|N)$. By Bayes' formula

$$\mathbb{P}(A_k|N) = \frac{\mathbb{P}(N|A_k)\,\mathbb{P}(A_k)}{\sum_{j=1}^{n} \mathbb{P}(N|A_j)\,\mathbb{P}(A_j)}.$$

Let us evaluate all terms here. We have

$$\mathbb{P}(N|A_1) = 1 - \mathbb{P}(B \cap A_1|A_1) = 1 - \mathbb{P}(B|A_1) = 1 - p_1,$$

while

$$\mathbb{P}(N|_1 A_k) = 1 - \mathbb{P}(B \cap A_1|A_k) = 1, \quad k > 1.$$

It follows that

$$\sum_{j=1}^{n} \mathbb{P}(N|A_j)\,\mathbb{P}(A_j) = (1-p_1)\frac{1}{n} + \frac{n-1}{n} = \frac{n-p_1}{n}.$$

Therefore,

$$\mathbb{P}(A_1|N) = \frac{\mathbb{P}(N|A_1)\,\mathbb{P}(A_1)}{\frac{n-p_1}{n}} = \frac{(1-p_1)\frac{1}{n}}{\frac{n-p_1}{n}} = \frac{1-p_1}{n-p_1}$$

and for $k > 1$

$$\mathbb{P}(A_k|N) = \frac{\mathbb{P}(N|A_k)\,\mathbb{P}(A_k)}{\frac{n-p_1}{n}} = \frac{1/n}{\frac{n-p_1}{n}} = \frac{1}{n-p_1}.$$

For example, if $n = 4$ and $p_1 = 0.5$ then $\mathbb{P}(A_k|N) = \frac{2}{7}$ for any $k = 2, 3, 4$.

Let us extend this question as follows. Suppose that after digging on the islands $1, 2, ..., m$ the result is still negative. Let us evaluate the probability that the chest is on k-th island. Denote by N_m the event that the chest was not found on the islands $1, 2, ..., m$. The event N_m^c that the chest was found on one of the islands $1, 2, ..., m$ is given by

$$N_m^c = B \cap (A_1 \cup A_2 \cup ... \cup A_m).$$

The conditional probability $\mathbb{P}(N_m|A_k)$ can be evaluated as follows: for $k \leq m$

$$\mathbb{P}(N_m|A_k) = 1 - \mathbb{P}(N_m^c|A_k) = 1 - \mathbb{P}(B \cap (A_1 \cup A_2 \cup ... \cup A_m)|A_k) = 1 - \mathbb{P}(B|A_k) = 1 - p_k,$$

and for $k > m$

$$\mathbb{P}(N_m|A_k) = 1 - \mathbb{P}(B \cap (A_1 \cup A_2 \cup ... \cup A_m)|A_k) = 1.$$

It follows that

$$\sum_{j=1}^{n} \mathbb{P}(N_m|A_j)\mathbb{P}(A_j) = \frac{1-p_1}{n} + ... + \frac{1-p_m}{n} + \frac{n-m}{n} = \frac{n-p_1-...-p_m}{n}.$$

By Bayes' formula we obtain: for $k \leq m$

$$\mathbb{P}(A_k|N_m) = \frac{\mathbb{P}(N_m|A_k)\mathbb{P}(A_k)}{\frac{n-p_1-...-p_m}{n}} = \frac{(1-p_k)\frac{1}{n}}{\frac{n-p_1-...-p_m}{n}} = \frac{1-p_k}{n-p_1-...-p_m}$$

and for $k > m$

$$\mathbb{P}(A_k|N_m) = \frac{\mathbb{P}(N_m|A_k)\mathbb{P}(A_k)}{\frac{n-p_1-...-p_m}{n}} = \frac{\frac{1}{n}}{\frac{n-p_1-...-p_m}{n}} = \frac{1}{n-p_1-...-p_m}.$$

In particular, the probability that the chest is buried in one of the $n-m$ remaining islands given that the search on the first m islands was unsuccessful is equal to

$$\frac{n-m}{n-p_1-...-p_m}.$$

For example, if $n=4$, $m=2$ and $p=0.5$ then the above probability is equal to $\frac{2}{3}$.

3.3 Product of probability spaces

3.3.1 Product of discrete probability spaces

Let $(\Omega', \mathcal{F}', \mathbb{P}')$ and $(\Omega'', \mathcal{F}'', \mathbb{P}'')$ be two discrete probability spaces. Consider the *product space* $(\Omega, \mathcal{F}, \mathbb{P})$ that is defined as follows:

$$\Omega = \Omega' \times \Omega'' = \{(k,l) : k \in \Omega' \text{ and } l \in \Omega''\},$$

$\mathcal{F} = 2^{\Omega}$, and $\mathbb{P}$ is defined by the stochastic sequence

$$p_{(k,l)} = p'_k p''_l,$$

where p'_k and p''_l are the stochastic sequences of $\mathbb{P}'$ and $\mathbb{P}''$, respectively. The sequence $p_{(k,l)}$ is stochastic because

$$\sum_{k,l} p_{(k,l)} = \sum_{k,l} p'_k p''_l = \sum_k p'_k \sum_l p''_l = 1.$$

Hence, the product space $(\Omega, \mathcal{F}, \mathbb{P})$ is a discrete probability space.

Claim. *For all $A \subset \Omega'$ and $B \subset \Omega''$, we have*

$$\mathbb{P}(A \times B) = \mathbb{P}'(A)\mathbb{P}''(B). \tag{3.7}$$

Proof. We have

$$\mathbb{P}(A \times B) = \sum_{(k,l)\in A\times B} p_{(k.l)} = \sum_{k\in A, l\in B} p'_k p''_l = \sum_{k\in A} p'_k \sum_{l\in B} p''_l = \mathbb{P}'(A)\,\mathbb{P}''(B).$$

■

The above probability measure $\mathbb{P}$ is called the product of $\mathbb{P}'$ and $\mathbb{P}''$ and is denoted by $\mathbb{P}' \times \mathbb{P}''$.

Example. Let $\mathbb{P}' \sim U(n)$ and $\mathbb{P}'' \sim U(m)$. Then $\mathbb{P}' \times \mathbb{P}'' \sim U(nm)$, because $p_{(k,l)} = \frac{1}{n}\frac{1}{m} = \frac{1}{nm}$.

By induction one can consider the product of finitely many discrete probability spaces: if $\left\{\left(\Omega^{(i)}, \mathcal{F}^{(i)}, \mathbb{P}^{(i)}\right)\right\}_{i=1}^{n}$ is a finite sequence of discrete probability spaces then their product is the discrete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where

$$\Omega = \Omega^{(1)} \times \Omega^{(2)} \times ... \times \Omega^{(n)} = \left\{(k_1, ..., k_n) : k_i \in \Omega^{(i)} \text{ for all } i = 1, ..., n\right\},$$

$\mathcal{F} = 2^{\Omega}$, and $\mathbb{P}$ is given by the stochastic sequence

$$p_{(k_1,...,k_n)} = p^{(1)}_{k_1} p^{(2)}_{k_2} ... p^{(n)}_{k_n},$$

where $p^{(i)}_k$ is the stochastic sequence of $\mathbb{P}^{(i)}$. If $A_i \subset \Omega^{(i)}$ then it follows from (3.7) that

$$\mathbb{P}(A_1 \times ... \times A_n) = \mathbb{P}^{(1)}(A_1) ... \mathbb{P}^{(n)}(A_n).$$

3.3.2 Product of general probability spaces

Let $(\Omega', \mathcal{F}', \mathbb{P}')$ and $(\Omega'', \mathcal{F}'', \mathbb{P}'')$ be two arbitrary probability spaces. We would like to define a probability space on the product set $\Omega = \Omega' \times \Omega''$. Consider the family of subsets of Ω

$$\mathcal{F}' \times \mathcal{F}'' = \left\{A \times B : A \in \mathcal{F}', B \in \mathcal{F}''\right\},$$

that is by Exercise 6 a semi-algebra (but not necessarily an algebra). Define function $\mathbb{P}$ on $\mathcal{F}' \times \mathcal{F}''$ by

$$\mathbb{P}(A \times B) = \mathbb{P}'(A)\,\mathbb{P}''(B). \tag{3.8}$$

Theorem 3.2 *Function $\mathbb{P}$ is a probability measure on $\mathcal{F}' \times \mathcal{F}''$.*

This theorem is hard and is proved in measure theory courses. The difficult part is to prove the σ-additivity of $\mathbb{P}$ (for the finite additivity see Exercise 13**)**. That $\mathbb{P}(\Omega) = 1$ is trivial since

$$\mathbb{P}(\Omega) = \mathbb{P}(\Omega' \times \Omega'') = \mathbb{P}'(\Omega')\,\mathbb{P}''(\Omega'') = 1.$$

Lecture 7
04.10.10

The probability measure $\mathbb{P}$ is called the product of $\mathbb{P}'$ and $\mathbb{P}''$ and is denoted by $\mathbb{P}' \times \mathbb{P}''$.

Corollary 3.3 *The probability measure* $\mathbb{P}$ *defined on* $\mathcal{F}' \times \mathcal{F}''$ *by* (3.8) *uniquely extends to the* σ*-algebra* $\mathcal{F} = \sigma(\mathcal{F}' \times \mathcal{F}'')$.

Proof. Indeed, this is a direct consequence of Theorem 3.2 and Corollary 2.5. ■

Definition. The probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega = \Omega' \times \Omega''$, $\mathcal{F} = \sigma(\mathcal{F}' \times \mathcal{F}'')$, and $\mathbb{P} = \mathbb{P}' \times \mathbb{P}''$ is called the *product* of the probability spaces $(\Omega', \mathcal{F}', \mathbb{P}')$ and $(\Omega'', \mathcal{F}'', \mathbb{P}'')$.

The product of discrete probability spaces is a particular case of this construction. Of course, by induction the notion of the product extends to any finite number of probability spaces.

Example. Consider the probability space $([0,1], \mathcal{B}, \lambda)$, where $\mathcal{B}$ is the Borel σ-algebra of the unit interval $[0,1]$ and λ is the Lebesgue measure. The product of n copies of this space yields a σ-algebra $\mathcal{B}_n$ on the unit cube $[0,1]^n$ and a probability measure λ_n on $\mathcal{B}_n$, which is called the n-dimensional Lebesgue measure. More precisely, one first obtains a semi-algebra $\mathcal{F}_n$ that consists of all boxes $I_1 \times ... \times I_n$ where I_k are subintervals of $[0,1]$. One defines λ_n on $\mathcal{F}_n$ by

$$\lambda_n(I_1 \times ... \times I_n) = \lambda(I_1) ... \lambda(I_n) = \ell(I_1) ... \ell(I_n),$$

that is, $\lambda_n(I_1 \times ... \times I_n)$ is the n-dimensional volume of the box. Then by Theorem 2.4 measure λ_n extends from the semi-algebra $\mathcal{F}_n$ to the minimal σ-algebra $\sigma(\mathcal{F}_n)$. The latter is called the Borel σ-algebra of the unit cube $[0,1]^n$ and is denoted by $\mathcal{B}([0,1]^n)$. One can show that it is the minimal σ-algebra containing all open and closed subsets of $[0,1]^n$ (see Exercise 2). The elements of $\mathcal{B}([0,1]^n)$ are called *Borel subsets* of $[0,1]^n$.

Define the Borel σ-algebra $\mathcal{B}(\mathbb{R}^n)$ as the minimal σ-algebra containing all boxes in $\mathbb{R}^n$, or equivalently, all open subsets of $\mathbb{R}^n$. By splitting the space $\mathbb{R}^n$ into a countable union of unit cubes, one extends the Lebesgue measure λ_n from $\mathcal{B}([0,1]^n)$ to $\mathcal{B}(\mathbb{R}^n)$, although the latter is allowed to take ∞ values.

3.4 Independent events

Independence is one of the most important notions of Probability Theory, that distinguishes it from Measure Theory.

3.4.1 Definition and examples

Definition. Two events A, B in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are called *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) P(B). \tag{3.9}$$

The motivation for the identity is as follows. It is natural to say that A is independent of B, if

$$\mathbb{P}(A|B) = \mathbb{P}(A), \tag{3.10}$$

that is, event A occurs with the same probability regardless of whether B is given or not. Using the definition

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

we obtain that (3.10) is equivalent to (3.9). However, (3.9) has advantage because it is explicitly symmetric in A, B and does not require that $\mathbb{P}(B) \neq 0$. Hence, one takes (3.9) as the definition of the independence of A and B. The above argument shows that (3.10) is equivalent to the independence provided $\mathbb{P}(B) \neq 0$.

Example. If A and B are two events and $\mathbb{P}(B) = 0$ or 1 then A and B are independent. Indeed, if $\mathbb{P}(B) = 0$ then also $\mathbb{P}(A \cap B) = 0$ whence the identity (3.9) follows. If $\mathbb{P}(B) = 1$ then $\mathbb{P}(B^c) = 0$ and

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c \cup B^c) = 1 - \mathbb{P}(A^c) = \mathbb{P}(A) = \mathbb{P}(A)\,\mathbb{P}(B).$$

In particular, A and $\emptyset$ are always independent as well as A and Ω.

Example. (0-1 law) Suppose that the events A and A are independent. We claim that $\mathbb{P}(A) = 0$ or 1, which follows from

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2.$$

Example. Let $\Omega = \{k \in \mathbb{Z}\colon 0 \leq k \leq 99\}$ and let $\mathbb{P}$ be the uniform distribution on Ω. We consider each integer in Ω as a two-digit decimal number and define the following two events:

$$\begin{aligned} A &= \{k \in \Omega : \text{the first digit of } k \text{ is } 1\} \\ B &= \{k \in \Omega\colon \text{the second digit of } k \text{ is } 2\}. \end{aligned}$$

The number of integers in Ω with the first digit 1 is 10 so that $\mathbb{P}(A) = \frac{10}{100} = \frac{1}{10}$. The number of integers in Ω with the second digit 2 is also 10, so that $\mathbb{P}(B) = \frac{1}{10}$. The only element in $A \cap B$ is 12, so that

$$\mathbb{P}(A \cap B) = \frac{1}{100} = \mathbb{P}(A)\,\mathbb{P}(B).$$

Hence, A and B are independent.

Example. Let $\Omega = [0,1]^2$ be the unit square and $\mathbb{P}$ be the area (=two dimensional Lebesgue measure) defined on Borel σ-algebra $\mathcal{F}$. Let I and J be two subintervals of $[0,1]$ and consider the rectangles

$$A = I \times [0,1], \quad B = [0,1] \times J$$

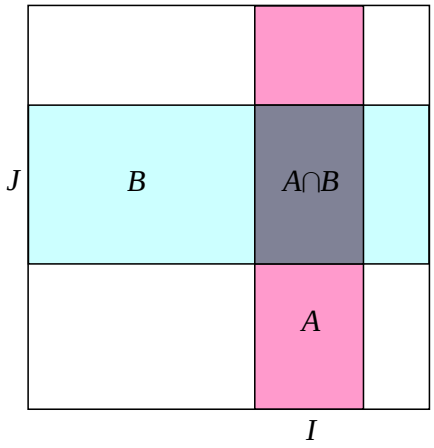


Figure 3.1: Independent events A and B

(see Fig. 3.1). We claim that A and B are independent.

Indeed, we have $\mathbb{P}(A) = \ell(I)$, $\mathbb{P}(B) = \ell(J)$, whence

$$\mathbb{P}(A \cap B) = \mathbb{P}(I \times J) = \ell(I)\,\ell(J) = \mathbb{P}(A)\,\mathbb{P}(B).$$

This example shows how to construct two independent events with prescribed probabilities.

Definition. Let $\{A_i\}$ be an indexed family of events, where the index i varies in an arbitrary set. The family $\{A_i\}$ is called *independent* (or the events in this family are called independent) if, for any finite sequence of distinct indices $i_1, ..., i_m$,

$$\mathbb{P}(A_{i_1} \cap ... \cap A_{i_m}) = \mathbb{P}(A_{i_1}) ....\mathbb{P}(A_{i_m}).$$

For example, three events A, B, C are independent if the following identities are satisfied:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B), \quad \mathbb{P}(A \cap C) = \mathbb{P}(A)\,\mathbb{P}(C), \quad \mathbb{P}(B \cap C) = \mathbb{P}(B)\,\mathbb{P}(C)$$

and

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C). \tag{3.11}$$

In other words, three events A, B, C are independent if they are pairwise independent and in addition satisfy (3.11) (see Exercises 21, 22, 26 for various examples).

Many examples of independent events can be constructed by using products of probability spaces. Let $\left\{\left(\Omega^{(i)}, \mathcal{F}^{(i)}, \mathbb{P}^{(i)}\right)\right\}_{i=1}^{n}$ be a finite sequence of probability spaces and let $(\Omega, \mathcal{F}, \mathbb{P})$ be their product. In particular,

$$\Omega = \Omega^{(1)} \times ... \times \Omega^{(n)} = \left\{(\omega_1, ..., \omega_n) : \omega_i \in \Omega^{(i)} \text{ for all } i = 1, ..., n\right\}$$

and

$$\mathbb{P}(A_1 \times ... \times A_n) = \mathbb{P}^{(1)}(A_1) ...\mathbb{P}^{(n)}(A_n) \tag{3.12}$$

for all $A_i \in \mathcal{F}^{(i)}$.

Theorem 3.4 *Chose one event A_i in each $\mathcal{F}_i$ and consider the following cylindrical events in $\mathcal{F}$ for all $i = 1, ..., n$:*

$$\begin{aligned} C_i &= \{(\omega_1, ..., \omega_n) \in \Omega : \omega_i \in A_i\} \\ &= \Omega^{(1)} \times ... \times \Omega^{(i-1)} \times A_i \times \Omega^{(i+1)} \times ... \times \Omega^{(n)}. \end{aligned} \tag{3.13}$$

Then the sequence $\{C_i\}_{i=1}^n$ is independent and

$$\mathbb{P}(C_i) = \mathbb{P}^{(i)}(A_i). \tag{3.14}$$

Proof. The identity (3.14) follows immediately from (3.12). To prove the independence, we need to verify that, for any $1 \leq m \leq n$ and for any sequence of indices $1 \leq i_1 < i_2 < ... < i_m \leq n$,

$$\mathbb{P}(C_{i_1} \cap C_{i_2} \cap ... \cap C_{i_m}) = \mathbb{P}(C_{i_1}) \mathbb{P}(C_{i_2}) ... \mathbb{P}(C_{i_m}).$$

For simplicity of notation, take $i_1 = 1, i_2 = 2, ..., i_m = m$ (the same argument applies to a general sequence). Then we have

$$\begin{aligned} C_1 \cap C_2 \cap ... \cap C_m &= \{(\omega_1, ..., \omega_n) \in \Omega : \omega_1 \in A_1, \omega_2 \in A_2, ..., \omega_m \in A_m\} \\ &= A_1 \times A_2 \times ... \times A_m \times \Omega^{(m+1)} \times ... \times \Omega^{(n)}. \end{aligned}$$

It follows from (3.7) that

$$\begin{aligned} \mathbb{P}(C_1 \cap C_2 \cap ... \cap C_m) &= \mathbb{P}^{(1)}(A_1) \mathbb{P}^{(2)}(A_2) ... \mathbb{P}^{(m)}(A_m) \mathbb{P}^{(m+1)}\left(\Omega^{(m+1)}\right) ... \mathbb{P}^{(n)}\left(\Omega^{(n)}\right) \\ &= \mathbb{P}^{(1)}(A_1) \mathbb{P}^{(2)}(A_2) ... \mathbb{P}^{(m)}(A_m) \end{aligned}$$

Finally using (3.14) we obtain

$$\mathbb{P}(C_1 \cap C_2 \cap ... \cap C_m) = \mathbb{P}(C_1) \mathbb{P}(C_2) ... \mathbb{P}(C_m).$$

■

Example. Let Ω_n be the set $\{k \in \mathbb{Z} : 0 \leq k \leq 10^n - 1\}$ where n is a positive integer, and let $\mathbb{P}_n$ be the uniform distribution on Ω_n, so that $\left(\Omega_n, 2^{\Omega_n}, \mathbb{P}_n\right)$ is a discrete probability space. Each integer $k \in \Omega_n$ can be regarded as an n-digit decimal (by adding zeros in front if necessary). Choose a sequence $\{d_i\}_{i=1}^n$ of decimal digits, that is, $d_i = 0, 1, ..., 9$ and consider the events

$$C_i = \{k \in \Omega \text{: the } i\text{-th decimal digit of } k \text{ is } d_i\},$$

where $i = 1, ..., n$. We claim that the events $C_1, ..., C_n$ are independent. Observe that the space $\left(\Omega_n, 2^{\Omega_n}, \mathbb{P}_n\right)$ is the product of n copies of $\left(\Omega_1, 2^{\Omega_1}, \mathbb{P}_1\right)$ where $\Omega_1 = \{0, 1, ..., 9\}$ and $\mathbb{P}_1$ is the uniform distribution on Ω_1 (because the product of uniform distributions is again a uniform distribution). The event C_i has the form (3.13) with $A_i = \{d_i\}$ so that the sequence $C_1, ..., C_n$ is independent by Theorem 3.4.

Example. Theorem 3.4 allows to construct an arbitrarily long sequences of independent events with prescribed probabilities. Indeed, suppose we would like

to construct a sequence of n independent events $C_1, ..., C_n$ such that $\mathbb{P}(C_i) = p_i$ where $p_1, ..., p_n$ are given values from $[0, 1]$. Then we first chose n probability spaces $\left\{\left(\Omega^{(i)}, \mathcal{F}^{(i)}, \mathbb{P}^{(i)}\right)\right\}_{i=1}^n$ and events $A_i \in \mathcal{F}_i$ such that $\mathbb{P}^{(i)}(A_i) = p_i$. Then by Theorem 3.4 we obtain in the product space a sequence $\{C_i\}_{i=1}^n$ of independent events, having the probabilities p_i respectively.

Of course, one still has to show the existence of the initial events A_i with the required property. For example, one can choose $\Omega^{(i)}$ to consist of two elements, say $\{1, 2\}$, and $\mathbb{P}^{(i)}$ to be a discrete probability measure defined by the stochastic sequence $\{p_i, 1 - p_i\}$. Then the event $A_i = \{1\}$ has $\mathbb{P}^{(i)}$-probability p_i. Alternatively, take $\Omega^{(i)}$ to be the unit interval $[0, 1]$ and $\mathbb{P}^{(i)}$ to be the Lebesgue measure on the Borel σ-algebra. Then the event $A_i = [0, p_i]$ has the probability p_i.

3.4.2 Operations with independent events I

It is natural to expect that independence is preserved by certain operations on events. For example, let A, B, C, D be independent events, and try to understand why the following couples of events are independent:

1. $A \cap B$ and $C \cap D$
2. $A \cup B$ and $C \cup D$
3. D and $E = (A \cap B) \cup (C \setminus A)$

It is easy to show that $A \cap B$ and $C \cap D$ are independent. Indeed, we have

$$\mathbb{P}((A \cap B) \cap (C \cap D)) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)\mathbb{P}(D) = \mathbb{P}(A \cap B)\mathbb{P}(C \cap D).$$

It is less obvious why $A \cup B$ and $C \cup D$ are independent. This will follow from the following statement.

Lemma 3.5 *Let $\mathcal{A}$ be a sequence of independent events. Suppose that a sequence $\mathcal{A}'$ is obtained from $\mathcal{A}$ by one (or finite number) of the following procedures:*

1. *Adding to $\mathcal{A}$ the event $\emptyset$ or Ω.*
2. *Two events $A, B \in \mathcal{A}$ are replaced by $A \cap B$, and the rest is the same.*
3. *An event $A \in \mathcal{A}$ is replaced by A^c, and the rest is the same.*
4. *Two events $A, B \in \mathcal{A}$ are replaced by $A \cup B$, and the rest is the same.*
5. *Two events $A, B \in \mathcal{A}$ are replaced by $A \setminus B$, and the rest is the same.*

Then the sequence $\mathcal{A}'$ is independent.

As a consequence, we see that if A, B, C, D are independent then $A \cup B$ and $C \cup D$ are independent. However, Lemma 3.5 is not yet enough to prove the independence of D and E as above, because A is involved twice in the formula defining E. A general theorem will be proved below that covers all such cases.

Lecture 8
05.10.10

Proof. Each of the above procedures adds to $\mathcal{A}$ a new event A' and removes from $\mathcal{A}$ some of the events. In order to prove that $\mathcal{A}'$ is independent, it suffices to show that, for any events $A_1, A_2, ..., A_k$ with distinct indices which remain in $\mathcal{A}$,

$$\mathbb{P}(A' \cap A_1 \cap ... \cap A_k) = \mathbb{P}(A')\mathbb{P}(A_1)...\mathbb{P}(A_k). \tag{3.15}$$

Case 1. $A' = \emptyset$ or Ω. Both sides of (3.15) vanish if $A' = \emptyset$. If $A' = \Omega$ then it can be removed from both sides of (3.15), so (3.15) follows from the independence of $A_1, A_2, ..., A_k$..

Case 2. $A' = A \cap B$. We have

$$\begin{aligned} \mathbb{P}((A \cap B) \cap A_1 \cap A_2 \cap ... \cap A_k) &= \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(A_1)\mathbb{P}(A_2)...\mathbb{P}(A_k) \\ &= \mathbb{P}(A \cap B)\mathbb{P}(A_1)\mathbb{P}(A_2)...\mathbb{P}(A_k). \end{aligned}$$

Case 3. $A' = A^c$. Then

$$\begin{aligned} \mathbb{P}(A^c \cap A_1 \cap A_2 \cap ... \cap A_k) &= \mathbb{P}(A_1 \cap A_2 \cap ... \cap A_k) - \mathbb{P}(A \cap A_1 \cap A_2 \cap ... \cap A_k) \\ &= \mathbb{P}(A_1)\mathbb{P}(A_2)...\mathbb{P}(A_k) - \mathbb{P}(A)\mathbb{P}(A_1)\mathbb{P}(A_2)...\mathbb{P}(A_k) \\ &= \mathbb{P}(A^c)\mathbb{P}(A_1)\mathbb{P}(A_2)...\mathbb{P}(A_k) \end{aligned}$$

Case 4. $A' = A \cup B$. Suffice to note that by the identity

$$(A \cup B) = (A^c \cap B^c)^c$$

this case amounts to the previous two.

Case 5. $A' = A \setminus B$. Use the identity

$$A \setminus B = A \cap B^c$$

and the previous cases. ■

Lemma 3.5 allows to complete the justification of the argument in Introduction for the proof of the inequality

$$(1 - p^n)^m + (1 - q^m)^n \geq 1, \tag{3.16}$$

where $p \in [0, 1]$ and $q = 1 - p$. Indeed, we need for the proof nm independent events each having the given probability p. These events can be constructed as was explained above. Denote them by A_{ij} where $i = 1, 2, ..., n$ and $j = 1, 2, ..., m$. Define a random $n \times m$ matrix (M_{ij}) where i is the row index and j is the column index, as follows:

$$M_{ij} = 1 \text{ if } A_{ij} \text{ occurs, and } M_{ij} = 0 \text{ otherwise.}$$

More precisely, M_{ij} is a function of ω such that

$$M_{ij}(\omega) = \begin{cases} 1, & \omega \in A_{ij}, \\ 0, & \omega \notin A_{ij}. \end{cases}$$

Denote by C_j the event that the j-th column contains only 1, that is,

$$C_j = \{M_{ij} = 1 \text{ for all } i = 1, ..., n\} = \{A_{ij} \text{ occurs for all } i = 1, ..., n\}.$$

Since $C_j = \bigcap_{i=1}^n A_{ij}$ and $\{A_{ij}\}$ are independent, we obtain

$$\mathbb{P}(C_j) = \prod_{i=1}^n \mathbb{P}(A_{ij}) = p^n.$$

By Lemma 3.5, events $\{C_j\}$ are independent so $\left\{C_j^c\right\}$ are also independent, whence

$$\mathbb{P}(\bigcap_{j=1}^m C_j^c) = (1 - p^n)^m.$$

Similarly, considering the events R_i that the i-th row contains only 0, we obtain that

$$\mathbb{P}(\bigcap_{i=1}^n R_i^c) = (1 - q^m)^n.$$

The desired inequality (3.16) follows then from the subadditivity of probability and the following identity

$$\left(\bigcap_j C_j^c\right) \bigcup \left(\bigcap_i R_i^c\right) = \Omega.$$

To prove this identity, observe that it is equivalent to

$$\left(\bigcup_j C_j\right) \bigcap \left(\bigcup_i R_i\right) = \emptyset.$$

The left hand side here is the event

$$\{\text{some column contains only 1}\} \cap \{\text{some row contains only 0}\},$$

that can never occur, because the matrix entry at the intersection of the said row and column must be simultaneously 0 and 1.

3.4.3 Operations with independent events II

Let us generalize the notion of independence as follows.

Definition. Let $\{\mathcal{A}_i\}$ be a sequence of families of events. We say that the sequence $\{\mathcal{A}_i\}$ is independent if, for all choices of $A_i \in \mathcal{A}_i$ the sequence $\{A_i\}$ is independent.

Example. As follows from Lemma 3.5, the sequence of events $\{A_1, A_2, ...\}$ is independent if and only if the following sequence of families is independent:

$$\left\{\begin{pmatrix} A_1 \\ A_1^c \end{pmatrix}, \begin{pmatrix} A_2 \\ A_2^c \end{pmatrix}, ...\right\}.$$

Theorem 3.6 *Let each family $\mathcal{A}_i$ be closed under $\cap$. If the sequence $\{\mathcal{A}_i\}$ is independent then the sequence of the algebras $\{a(\mathcal{A}_i)\}$ is also independent. Moreover, the sequence of σ-algebras $\{\sigma(\mathcal{A}_i)\}$ is independent, too.*

Proof. By adding Ω to each of the families $\mathcal{A}_i$ we do not change the independence of the sequence $\{\mathcal{A}_i\}$, so we can assume $\Omega \in \mathcal{A}_i$.

In order to check the independence of the sequence $\{a(\mathcal{A}_i)\}$, we need to test finite sequences of events chosen from those families. Hence, it suffice to restrict consideration to the case when this sequence is finite, say $i = 1, 2, ..., n$. Also, it suffices to prove that the independence of $\{\mathcal{A}_1, ..., \mathcal{A}_n\}$ implies that the sequence $\{a(\mathcal{A}_1), \mathcal{A}_2, ..., \mathcal{A}_n\}$ is independent. If we know that then we can by induction replace $\mathcal{A}_2$ by $a(\mathcal{A}_2)$ etc.

To show the independence of $\{a(\mathcal{A}_1), \mathcal{A}_2, ..., \mathcal{A}_n\}$, it suffices to verify that, for arbitrary events $A_1 \in a(\mathcal{A}_1)$, $A_2 \in \mathcal{A}_2, ..., A_n \in \mathcal{A}_n$, the following identity holds:

$$\mathbb{P}(A_1 \cap A_2 \cap ... \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2)...\mathbb{P}(A_n). \tag{3.17}$$

Indeed, by the definition of the independent events, one needs to check this property also for subsequences $\{A_{i_k}\}$ but this amounts to the full sequence $\{A_i\}$ by choosing the missing events to be Ω.

Denote for simplicity $A = A_1$ and $B = A_2 \cap A_3 \cap ... \cap A_n$. Since $\{A_2, A_3, ..., A_n\}$ are independent, (3.17) is equivalent to

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{3.18}$$

In other words, we are left to prove that A and B are independent, for any $A \in a(\mathcal{A}_1)$ and B being an intersection of events from $\mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n$.

Fix such an event B and consider the family S of suitable events A that satisfy (3.18), that is,

$$S = \{A \in a(\mathcal{A}_1) : \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\}.$$

We need to show that $S = a(\mathcal{A}_1)$. Observe that all events from $\mathcal{A}_1$ are suitable, that is, $S \supset \mathcal{A}_1$. Let us prove that S is closed under the monotone difference "$-$". Indeed, if $A, A' \in S$ and $A \supset A'$ then

$$\begin{aligned} \mathbb{P}((A - A') \cap B) &= \mathbb{P}(A \cap B) - \mathbb{P}(A' \cap B) \\ &= \mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(A')\mathbb{P}(B) \\ &= \mathbb{P}(A - A')\mathbb{P}(B). \end{aligned}$$

It follows that

$$S = S^{-} \supset \mathcal{A}_1^{-} = a(\mathcal{A}_1)$$

where the last identity holds by Theorem 2.8, whence $S = a(\mathcal{A}_1)$ follows.

The independence of σ-algebras $\{\sigma(\mathcal{A}_i)\}$ is proved in the same way. It amounts to veryfiyng that, for any $B \in \mathcal{A}_2 \cap ... \cap \mathcal{A}_n$, the family of suitable events

$$S = \{A \in \sigma(\mathcal{A}_1) : \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\}$$

coincides with $\sigma(\mathcal{A}_1)$. By the first part, we have $S \supset a(\mathcal{A}_1)$. Let us show that S is closed under monotone limits. Indeed, if $A = \lim_{k\to\infty} A^{(k)}$ where$\left\{A^{(k)}\right\}$ is a monotone sequence of events from S then $A \cap B = \lim_{k\to\infty} A^{(k)} \cap B$ whence by the continuity of $\mathbb{P}$

$$\mathbb{P}(A \cap B) = \lim_{k\to\infty} \mathbb{P}\left(A^{(k)} \cap B\right) = \lim_{k\to\infty} \mathbb{P}\left(A^{(k)}\right) \mathbb{P}(B) = \mathbb{P}(A)\,\mathbb{P}(B)$$

so that $A \in S$. It follows that

$$S = S^{\lim} \supset a(\mathcal{A}_1)^{\lim} = \sigma(\mathcal{A}_1),$$

which finishes the proof. ■

The next statement illustrates how one applies Theorem 3.6.

Corollary 3.7 *Suppose that $\{A_{ij}\}$ is a sequence of independent events parametrized by two indices i and j. Denote by $\mathcal{A}_j$ the family of all events $\{A_{ij}\}$ with a fixed j and arbitrary i. Then the sequence of σ-algebras $\{\sigma(\mathcal{A}_j)\}$ is independent.*

For example, if the sequence $\{A_{ij}\}$ is represented by a matrix with a row index i and a column index j

$$\begin{pmatrix} \mathcal{A}_1 & \mathcal{A}_2 & \dots & \dots & \mathcal{A}_j & \dots \\ A_{11} & A_{12} & \dots & \dots & \dots & \dots \\ A_{21} & A_{22} & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ A_{i1} & A_{i2} & \dots & \dots & A_{ij} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

then $\mathcal{A}_j$ consists of all the events in the column j, and the claim is that the algebras generated by different columns, are independent. Clearly, the same applies to the rows.

Proof. Observe that the extended families $\left\{\mathcal{A}_j^{\cap}\right\}$ are also independent. Indeed, each event $B_j \in \mathcal{A}_j^{\cap}$ is an intersection of a finite number of events from $\mathcal{A}_j$, that is, has the form

$$B_j = A_{i_1 j} \cap A_{i_2 j} \cap ... \cap A_{i_k j}\,.$$

Hence, the sequence $\{B_j\}$ can be obtained by replacing in the double sequence $\{A_{ij}\}$ some elements by their intersections (and throwing away the rest), and the independence of $\{B_j\}$ follows by Lemma 3.5 from the independence of $\{A_{ij}\}$ across all i and j.

The sequence $\left\{\mathcal{A}_j^{\cap}\right\}$ satisfies the hypotheses of Theorem 3.6, whence it follows that the sequence $\{\sigma(\mathcal{A}_j)\}$ is independent. ■

Example. Let apply Corollary 3.7 to the aforementioned example of the events D and $E = (A \cap B) \cup (C \setminus A)$ assuming that A, B, C, D are independent. Indeed, in the matrix

$$\begin{pmatrix} A & D \\ B & \Omega \\ C & \Omega \end{pmatrix}$$

all events are independent. Therefore, the algebras $a(A, B, C)$ and $a(D, \Omega, \Omega)$ are independent. Since $E \in a(A, B, C)$, we conclude that D and E are independent.

Example. (Exercise 17) If three events A, B, C are independent then also C and D are independent for any $D \in a(A, B)$. Indeed, consider the matrix

$$\begin{pmatrix} A & C \\ B & \Omega \end{pmatrix}$$

where all entries are independent. By Corollary 3.7, the algebras $a(A, B)$ and $a(C, \Omega)$ are independent, whence the claim follows. Note that in Exercise 17 one is asked to prove the same *directly*, without using Corollary 3.7.

Chapter 4

Lebesgue integration

4.1 Null sets and complete measures

Let μ be a measure on a σ-algebra $\mathcal{F}$ of subsets of Ω. Recall that all elements of $\mathcal{F}$ are also called (μ- or $\mathcal{F}$-) measurable subsets of Ω.

Definition. We say that a set $N \subset \Omega$ is a *null* set if $N \subset M$ for some measurable set M with $\mu(M) = 0$.

Note that the set N itself may be not measurable so that $\mu(N)$ is not defined. It is difficult to construct an explicit example of such sets but the existence can be proved as follows. Consider the probability space $([0,1], \mathcal{B}, \lambda)$, where $\mathcal{B} = \mathcal{B}[0,1]$ is the Borel σ-algebra on $[0,1]$ and λ is the Lebesgue measure. It is possible to prove that $|\mathcal{B}| = |\mathbb{R}|$ where $|\cdot|$ denotes the cardinality of a set. Denote by C the Cantor set on $[0,1]$ so that $\lambda(C) = 0$ and $|C| = |\mathbb{R}|$. Hence, any subset of C is a null set, and the family of subsets of C has the cardinality

$$\left|2^C\right| = \left|2^{\mathbb{R}}\right| > |\mathbb{R}| = |\mathcal{B}|.$$

Hence, there are subsets of C that are not Borel sets, while they are nevertheless null sets.

For many applications it is desirable that the null sets are measurable. For example, if $\mathbb{P}(A) = 0$ for some event A, then it would be natural to conclude that $\mathbb{P}(B) = 0$ for any event $B \subset A$, while strictly speaking this is not always the case. To avoid such abnormal situations, we introduce the following definition.

Definition. A measure μ is said to be *complete* if $\mu(A) = 0$ implies that all subsets of A are μ-measurable. In other words, μ is complete if all null sets are measurable.

Of course, any discrete probability space is always complete. However, the Lebesgue measure is not complete on the Borel σ-algebra.

Lecture 9
11.10.10

Theorem 4.1 *Let μ be a measure on a σ-algebra $\mathcal{F}$ of subsets of Ω. Denote by $\mathcal{N}$ the family of all null sets of μ and by $\mathcal{F}'$ the family of all subsets A of Ω that satisfy the condition*

$$A \triangle B \in \mathcal{N} \quad \text{for some } B \in \mathcal{F}. \tag{4.1}$$

Then $\mathcal{F}'$ is a σ-algebra. Furthermore, for any $A \in \mathcal{F}'$ define $\mu'(A)$ by

$$\mu'(A) = \mu(B),$$

where B is a set from (4.1)*. Then μ' is a complete measure on $\mathcal{F}'$.*

Measure μ' is called the *completion* of measure μ, and the σ-algebra $\mathcal{F}'$ is called the completion of $\mathcal{F}$.

For example, the completion of the Borel σ-algebra (on intervals or on $\mathbb{R}$) with measure λ is a larger σ-algebra that is called *Lebesgue* σ-algebra. The elements of the Lebesgue σ-algebra are called *Lebesgue measurable* sets. The extension of the Lebesgue measure λ from the Borel σ-algebra to the Lebesgue σ-algebra is also called the Lebesgue measure and is denoted also by λ. The same applies to the Lebesgue measure λ_n in $\mathbb{R}^n$.

Proof of Theorem 4.1. The proof consists of a series of claims. Let us first prove some properties of symmetric difference. Recall that $\triangle$ is defined by

$$A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B).$$

Then condition $\omega \in A \triangle B$ means that ω belongs to exactly one of the sets A, B.

Claim 0. (*a*) *For any sets A, B,*

$$A^c \triangle B^c = A \triangle B.$$

(*b*) *For any sequences $\{A_k\}$ and $\{B_k\}$, set $A = \bigcup_k A_k$ and $B = \bigcup_k B_k$. Then*

$$A \triangle B \subset \bigcup_k (A_k \triangle B_k). \tag{4.2}$$

(*c*) *For any sets A, B_1, B_2*

$$B_1 \triangle B_2 \subset (A \triangle B_1) \cup (A \triangle B_2).$$

To prove (*a*), observe that the condition $\omega \in A^c \triangle B^c$ means that ω belongs to exactly one of the sets A^c, B^c, that is, ω belongs to exactly one of the sets A, B, which is equivalent to $\omega \in A \triangle B$.

To prove (*b*), assume that $\omega \in A$ and $\omega \notin B$, and show that $\omega \in \bigcup_k (A_k \triangle B_k)$. By the definition of union, ω belongs to one of the sets A_k and to none of B_k. Hence, $\omega \in A_k \triangle B_k$ for some k, whence the claim follows. The case $\omega \notin A$ and $\omega \in B$ is treated similarly.

To prove (*c*), assume that $\omega \in B_1$ and $\omega \notin B_2$, and show that $\omega \in U := (A \triangle B_1) \cup (A \triangle B_2)$. If $\omega \in A$ then $\omega \in A \triangle B_2$ and, hence, $\omega \in U$. If $\omega \notin A$ then $\omega \in A \triangle B_1$ whence $\omega \in U$. The case $\omega \notin B_1$, $\omega \in B_2$ is treated in the same way.

Claim 1. *$\mathcal{F}'$ is a σ-algebra and $\mathcal{F}' \supset \mathcal{F}$.*

Every set $A \subset \mathcal{F}$ belongs also to $\mathcal{F}'$ because $A \triangle A = \emptyset$ is a null set. Hence, $\mathcal{F}' \supset \mathcal{F}$. It follows that $\emptyset, \Omega \in \mathcal{F}'$.

If $A \in \mathcal{F}'$ then, using the set B from (4.1) and Claim 0, we obtain

$$A^c \triangle B^c = A \triangle B \in \mathcal{N}.$$

Since $B^c \in \mathcal{F}$, it follows that $A^c \in \mathcal{F}'$.

Let $\{A_k\}$ be a finite or countable sequence of sets from $\mathcal{F}'$. Let us prove that $A := \bigcup_k A_k \in \mathcal{F}'$. For any A_k there exists a set $B_k \in \mathcal{F}$ that satisfies condition (4.1), that is,

$$A_k \triangle B_k \subset C_k \tag{4.3}$$

where $\mu(C_k) = 0$. Setting $B = \bigcup_k B_k$, we obtain by Claim 0

$$A \triangle B \subset \bigcup_k (A_k \triangle B_k) \subset \bigcup_k C_k =: C. \tag{4.4}$$

Since $B \in \mathcal{F}$ and $\mu(C) = 0$, we conclude that $A \in \mathcal{F}'$.

CLAIM 2. *μ' is a measure.*

First of all, μ' is well-defined in the following sense: if the condition (4.1) is satisfied for two different sets B, say for $B = B_1$ and $B = B_2$ then $\mu(B_1) = \mu(B_2)$ so that $\mu'(A)$ does not depend on a particular choice of B. Indeed, we have by Claim 0

$$B_1 \triangle B_2 \subset (A \triangle B_1) \cup (A \triangle B_2) \in \mathcal{N},$$

whence it follows that $\mu(B_1 \triangle B_2) = 0$ and

$$\mu(B_1) = \mu(B_1 \cap B_2) = \mu(B_2).$$

Let us prove that μ' is σ-additive. Let $\{A_k\}$ be a disjoint sequence of sets from $\mathcal{F}'$ and let $\{B_k\}$ and $\{C_k\}$ be as in the previous Claim. Set as above

$$A = \bigcup_k A_k, \quad B = \bigcup_k B_k, \quad C = \bigcup_k C_k.$$

It follows from (4.4) that

$$\mu'(A) = \mu(B) = \mu\left(\bigcup_k B_k\right) = \mu\left(\bigcup_k (B_k \setminus C_k)\right), \tag{4.5}$$

where the last equality holds because the set

$$\left(\bigcup_k B_k\right) \setminus \left(\bigcup_k (B_k \setminus C_k)\right) \subset \bigcup_k C_k = C$$

has measure 0. Observe also that the condition (4.3) implies that

$$B_k \setminus C_k \subset A_k.$$

It follows that all the sets $B_k \setminus C_k$ are disjoint, whence by the σ-additivity of μ

$$\mu\left(\bigcup_k (B_k \setminus C_k)\right) = \sum_k \mu(B_k \setminus C_k) = \sum_k \mu(B_k) = \sum_k \mu'(A_k). \tag{4.6}$$

Comparing (4.5) and (4.6) we obtain that μ' is σ-additive.

CLAIM 3. *Measure μ' is complete.*

Let us first show that $\mu'(A) = 0$ if and only if $A \in \mathcal{N}$ (which will imply the completeness of μ' since any subset of a null set is a null set). Indeed, if $\mu'(A) = 0$ then there is a set $B \in \mathcal{F}$ such that $\mu(B) = 0$ and $A \triangle B \in \mathcal{N}$. The latter means that $A \triangle B \subset M$ for some set $M \in \mathcal{F}$ with $\mu(M) = 0$. It follows that $A \subset B \cup M$. Since $\mu(B \cup M) = 0$, we conclude that A is a null set, which finishes the proof. ■

4.2 Measurable functions

Let Ω be an arbitrary non-empty set and $\mathcal{F}$ be a σ-algebra of subsets of Ω. Recall that a set $A \subset \Omega$ is called measurable (or $\mathcal{F}$-measurable) if $A \in \mathcal{F}$.

Definition. We say that a function $f : \Omega \to \mathbb{R}$ is *$\mathcal{F}$-measurable* (or simply *measurable* if the choice of $\mathcal{F}$ is obvious) if, for any $c \in \mathbb{R}$, the set $\{\omega \in \Omega : f(\omega) \leq c\}$ is $\mathcal{F}$-measurable, that is, belongs to $\mathcal{F}$.

Of course, the measurability of sets and functions depends on the choice of the σ-algebra $\mathcal{F}$. For example, in $\mathbb{R}^n$ we distinguish *Borel (measurable)* sets and functions, that are measurable with respect to the Borel σ-algebra, and Lebesgue measurable sets and functions, that are measurable with respect to the Lebesgue σ-algebra.

The measurability of a function f can be also restated as follows. Since

$$\{f \leq c\} = f^{-1}(-\infty, c],$$

we can say that a function f is measurable if, for any $c \in \mathbb{R}$, the set $f^{-1}(-\infty, c]$ is measurable.

Example. For any subset $A \subset \Omega$, its *indicator function* $\mathbf{1}_A$ is defined by

$$\mathbf{1}_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

Then set A is measurable if and only if function $\mathbf{1}_A$ is measurable. Indeed, for any real c, we have

$$\{\mathbf{1}_A \leq c\} = \begin{cases} \emptyset, & c < 0, \\ A^c, & 0 \leq c < 1, \\ \Omega, & c \geq 1. \end{cases}$$

The sets $\emptyset$ and Ω are always measurable, and A^c is measurable if and only if A is measurable, whence the claim follows.

Example. Any continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is Borel. Indeed, the set $f^{-1}(-\infty, c]$ is a closed subset of $\mathbb{R}^n$ as the preimage of a closed set $(-\infty, c]$ in $\mathbb{R}$. Since closed sets are Borel, we conclude that f is Borel.

Example. Any monotone function $f : \mathbb{R} \to \mathbb{R}$ is Borel. Indeed, the monotonicity implies that the set $\{f \leq c\}$ an interval, that is a Borel set.

Definition. A mapping $f : \Omega \to \mathbb{R}^n$ is called *measurable* (or $\mathcal{F}$-measurable) if, for all $c_1, c_2, ..., c_n \in \mathbb{R}$, the set

$$\{\omega \in \Omega : f_1(\omega) \leq c_1, f_2(\omega) \leq c_2, ..., f_n(\omega) \leq c_n\}$$

is measurable. Here f_k is the k-th component of f.

In other words, consider an infinite box in $\mathbb{R}^n$ of the form:

$$B = (-\infty, c_1] \times (-\infty, c_2] \times ... \times (-\infty, c_n], \tag{4.7}$$

and call it a *special box*. Then a mapping $f : \Omega \to \mathbb{R}^n$ is measurable if, for any special box B, the preimage $f^{-1}(B)$ is a $\mathcal{F}$-measurable subset of Ω.

Theorem 4.2 *A mapping $f : \Omega \to \mathbb{R}^n$ is $\mathcal{F}$-measurable if and only if, for any Borel set $A \subset \mathbb{R}^n$, the preimage $f^{-1}(A)$ is an $\mathcal{F}$-measurable set.*

Proof. If $f^{-1}(A)$ is measurable for any Borel set $A \subset \mathbb{R}^n$ then $f^{-1}(B)$ is measurable for special boxes B and, hence, the mapping f is measurable.

To prove the opposite implication, denote by$\mathcal{A}$ the family of all sets $A \subset \mathbb{R}^n$ such that $f^{-1}(A)$ is measurable. By hypothesis, $\mathcal{A}$ contains all special boxes. Let us prove that $\mathcal{A}$ is a σ-algebra (this follows also from Exercise 4 since in the notation of that exercise $\mathcal{A} = f(\mathcal{F})$). If $A \in \mathcal{A}$ then $f^{-1}(A)$ is measurable whence

$$f^{-1}(A^c) = \left(f^{-1}(A)\right)^c \in \mathcal{F},$$

whence $A^c \in \mathcal{A}$. Also, if $\{A_k\}$ is a finite or countable sequence of sets from $\mathcal{A}$ then $f^{-1}(A_k)$ is measurable for all k whence

$$f^{-1}\left(\bigcap_k A_k\right) = \bigcap_k f^{-1}(A_k) \in \mathcal{F},$$

which implies that $\bigcap A_k \in \mathcal{A}$. Finally, $\mathcal{A}$ contains $\mathbb{R}$ because $\mathbb{R}$ is the countable union of the intervals $(-\infty, n]$ where $n \in \mathbb{N}$, and $\mathcal{A}$ contains $\emptyset = \mathbb{R}^c$.

Hence, $\mathcal{A}$ is a σ-algebra containing all special boxes. It remains to show that $\mathcal{A}$ contains all the boxes in $\mathbb{R}^n$, which will imply that $\mathcal{A}$ contains all Borel sets in $\mathbb{R}^n$. In fact, it suffices to show that any box in $\mathbb{R}^n$ can be obtained from special boxes by a countable sequence of set-theoretic operations.

Assume first $n = 1$ and consider different types of intervals. If $A = (-\infty, a]$ then $A \in \mathcal{A}$ by hypothesis.

Let $A = (a, b]$ where $a < b$. Then $A = (-\infty, b] \setminus (-\infty, a]$, which proves that A belongs to $\mathcal{A}$ as the difference of two special intervals.

Let $A = (a, b)$ where $a < b$ and $a, b \in \overline{\mathbb{R}}$. Consider a strictly increasing sequence $\{b_k\}_{k=1}^{\infty}$ such that $b_k \uparrow b$ as $k \to \infty$. Then the intervals $(a, b_k]$ belong to $\mathcal{A}$ by the previous argument, and the obvious identity

$$A = (a, b) = \bigcup_{k=1}^{\infty} (a, b_k]$$

implies that $A \in \mathcal{A}$.

Let $A = [a, b)$. Consider a strictly increasing sequence $\{a_k\}_{k=1}^{\infty}$ such that $a_k \uparrow a$ as $k \to \infty$. Then $(a_k, b) \in \mathcal{A}$ by the previous argument, and the set

$$A = [a, b) = \bigcap_{k=1}^{\infty} (a_k, b)$$

is also in $\mathcal{A}$.

Finally, let $A = [a, b]$. Observing that $A = \mathbb{R} \setminus (-\infty, a) \setminus (b, +\infty)$ where all the terms belong to $\mathcal{A}$, we conclude that $A \in \mathcal{A}$.

Consider now the general case $n > 1$. We are given that $\mathcal{A}$ contains all boxes of the form

$$B = I_1 \times I_2 \times ... \times I_n$$

where I_k are special intervals of the form $(-\infty, c]$, and we need to prove that $\mathcal{A}$ contains all boxes of this form with arbitrary intervals I_k. If I_1 is an arbitrary

interval and $I_2, ..., I_n$ are special intervals then one shows that $B \in$A using the same argument as in the case $n = 1$ since I_1 can be obtained from the special intervals by a countable sequence of set-theoretic operations, and the same sequence of operations can be applied to the product $I_1 \times I_2 \times ... \times I_n$. Now let I_1 and I_2 be arbitrary intervals and $I_3, ..., I_n$ be special. We know that if I_2 is special then $B \in \mathcal{A}$. Obtaining an arbitrary interval I_2 from special intervals by a countable sequence of operations, we obtain that $B \in \mathcal{A}$ also for arbitrary I_1 and I_2. Continuing the same way, we obtain that $B \in \mathcal{A}$ if I_1, I_2, I_3 are arbitrary intervals while $I_4, ..., I_n$ are special, etc. Finally, we allow all the intervals $I_1, ..., I_n$ to be arbitrary. ■

Example. If $f : \Omega \to \mathbb{R}$ is a measurable function then the set

$$\{\omega \in \Omega : f(\omega) \text{ is irrational}\}$$

is measurable, because this set coincides with $f^{-1}(\mathbb{Q}^c)$, and $\mathbb{Q}^c$ is Borel since $\mathbb{Q}$ is Borel as a countable set.

Theorem 4.3 *Let $f_1, ..., f_n$ be measurable functions from Ω to $\mathbb{R}$ and let Φ be a Borel function from $\mathbb{R}^n$ to $\mathbb{R}$. Then the function*

$$F = \Phi(f_1, ..., f_n) : \Omega \to \mathbb{R}$$

is measurable.

In other words, the composition of a Borel function with measurable functions is measurable. Note that Borel functions cannot be replaced here by Lebesgue measurable functions.

Example. It follows from Theorem 4.3 that if f_1 and f_2 are two measurable functions on Ω then their sum $f_1 + f_2$ is also measurable. Indeed, consider the function $\Phi(x_1, x_2) = x_1 + x_2$ in $\mathbb{R}^2$, which is continuous and, hence, is Borel. Then $f_1 + f_2 = \Phi(f_1, f_2)$ and this function is measurable by Theorem 4.3. A direct proof by definition may be difficult: the fact that the set $\{f_1 + f_2 \leq c\}$ is measurable, is not immediately clear how to reduce this set to the measurable sets $\{f_1 \leq a\}$ and $\{f_1 \leq b\}$.

In the same way, the functions $f_1 f_2$, f_1/f_2 (provided $f_2 \neq 0$) are measurable. Also, the functions $\max(f_1, f_2)$ and $\min(f_1, f_2)$ are measurable, etc.

Proof of Theorem 4.3. Consider the mapping $f : \Omega \to \mathbb{R}^n$ whose components are f_k. This mapping is measurable because for any $c \in \mathbb{R}^n$, the set

$$\{\omega \in \Omega : f_1(\omega) \leq c_1, ..., f_n(\omega) \leq c_n\} = \{f_1(\omega) \leq c_1\} \cap \{f_2(\omega) \leq c_2\} \cap ... \cap \{f_n(\omega) \leq c_n\}$$

is measurable as the intersection of measurable sets. Let us show that $F^{-1}(I)$ is a measurable set for any special interval I, which will prove that F is measurable. Indeed, since $F(\omega) = \Phi(f(\omega))$, we obtain that

$$\begin{aligned} F^{-1}(I) &= \{\omega \in \Omega : F(\omega) \in I\} \\ &= \{\omega \in \Omega : \Phi(f(\omega)) \in I\} \\ &= \left\{\omega \in \Omega : f(\omega) \in \Phi^{-1}(I)\right\} \\ &= f^{-1}\left(\Phi^{-1}(I)\right). \end{aligned}$$

Since $\Phi^{-1}(I)$ is a Borel set, we obtain by Theorem 4.2 that $f^{-1}(\Phi^{-1}(I))$ is measurable, which proves that $F^{-1}(I)$ is measurable. ■

Example. If $A_1, A_2, ..., A_n$ is a finite sequence of measurable sets then the function

$$f = c_1\mathbf{1}_{A_1} + c_2\mathbf{1}_{A_2} + ... + c_n\mathbf{1}_{A_n}$$

is measurable for any choice of real constants c_i. Indeed, each of the function $\mathbf{1}_{A_i}$ is measurable, whence the claim following upon application of Theorem 4.3 with the function

$$\Phi(x) = c_1x_1 + ... + c_nx_n.$$

4.3 Sequences of measurable functions

As before, let Ω be an arbitrary set and $\mathcal{F}$ be a σ-algebra on Ω.

Definition. We say that a sequence $\{f_n\}_{n=1}^{\infty}$ of functions on Ω converges to a function f on Ω *pointwise* and write $f_n \to f$ if $f_n(\omega) \to f(\omega)$ as $n \to \infty$ for any $\omega \in \Omega$.

Theorem 4.4 *Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of measurable functions that converges pointwise to a function f. Then f is measurable, too.*

Proof. Fix some real c. Using the definition of a limit and the hypothesis that $f_n(\omega) \to f(\omega)$ as $n \to \infty$, we obtain that the inequality $f(\omega) \leq c$ is equivalent to the following condition:

$$\forall \varepsilon > 0 \ \exists N \in \mathbb{N} \ \forall n \geq N \quad f_n(\omega) \leq c + \varepsilon.$$

This can be written in the form of set-theoretic inclusion as follows:

$$\{f \leq c\} = \bigcap_{\varepsilon > 0} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{f_n \leq c + \varepsilon\}.$$

The values of ε can be restricted to rationals so that the set $\{f \leq c\}$ is obtained from the measurable sets $\{f_n \leq c + \varepsilon\}$ by countable unions and intersections. Therefore, the set $\{f \leq c\}$ is measurable, which finishes the proof. ■

Corollary 4.5 *Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of Lebesgue (or Borel) measurable functions on $\mathbb{R}^n$ that converges pointwise to a function f. Then f is Lebesgue (resp., Borel) measurable as well.*

4.4 The Lebesgue integral

Let Ω be an arbitrary set, $\mathcal{F}$ be a σ-algebra on Ω and μ be a complete measure on $\mathcal{F}$. We will define the notion of the integral $\int_{\Omega} f d\mu$ for an appropriate class of functions f.

4.4.1 Simple functions

Definition. A function $f : \Omega \to \mathbb{R}$ is called *simple* if it is measurable and the set of its values is at most countable.

Let $\{a_k\}$ be the sequence of all distinct values of a simple function f. Observe that the sets

$$A_k = \{\omega \in \Omega : f(\omega) = a_k\} \tag{4.8}$$

are measurable and form a partition of Ω, that is,

$$\Omega = \bigsqcup_k A_k. \tag{4.9}$$

Clearly, we have the identity

$$f = \sum_k a_k \mathbf{1}_{A_k} \tag{4.10}$$

for all $\omega \in \Omega$. Indeed, for each $\omega \in \Omega$, there is exactly one value of k such that $\omega \in A_k$ so that the series in the right hand side of (4.10) amounts to a single term a_k, that is exactly $f(\omega)$.

Conversely, any partition $\{A_k\}$ of Ω with measurable sets and any sequence of distinct reals $\{a_k\}$ determine by (4.10) a function $f(\omega)$ that satisfies also (4.8), which means that all simple functions have the form (4.10).

Definition. For any non-negative simple function $f : \Omega \to \mathbb{R}$ define the *Lebesgue integral* $\int_\Omega f d\mu$ by

$$\int_\Omega f d\mu := \sum_k a_k \mu(A_k). \tag{4.11}$$

Lecture 10
12.10.10

Note that the value on the right hand side of (4.11) is always defined as the sum of a non-negative series, and can be either a non-negative real number or infinity, that is, the integral $\int_\Omega f\, d\mu$ takes values in $[0, +\infty]$.

Note also that in order to be able to define $\mu(A_k)$, the sets A_k must be measurable, which is equivalent to the measurability of f.

If f a signed simple function then one still can define $\int_\Omega f d\mu$ by (4.11) assuming in addition that the series in the right hand side of (4.11) absolutely converges. However, we do not use this case.

The expression $\int_\Omega f d\mu$ has the full title "the integral of f over Ω against measure μ". The notation $\int_\Omega f d\mu$ should be understood as a whole, since we do not define what $d\mu$ means. This notation is traditionally used and has certain advantages.

Remark. In probability theory one uses a different notation for the integral. If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space then one writes

$$\mathbb{E}f = \int_\Omega f d\mathbb{P}$$

and refers to $\mathbb{E}f$ as the *expectation* of f. If the space $(\Omega, \mathcal{F}, \mathbb{P})$ is discrete then every function f on Ω is simple, which means that $\mathbb{E}f$ is then defined for all non-negative functions f on Ω.

Example. If $f \equiv a\mathbf{1}_A$ for some measurable set A then the decomposition (4.10) becomes

$$f = a\mathbf{1}_A + 0\mathbf{1}_{A^c} ,$$

whence it follows that

$$\int_\Omega f d\mu = a\mu(A) .$$

We will extend the notion of integral to more general measurable functions. However, first we prove some properties of the integral of simple functions.

Lemma 4.6 *Let $\Omega = \bigsqcup_k B_k$ where $\{B_k\}$ is a finite or countable sequence of measurable sets. Define a function f by*

$$f = \sum_k b_k 1_{B_k}$$

where $\{b_k\}$ is a sequence of non-negative reals, not necessarily distinct (note that $f = b_k$ on B_k). Then

$$\int_\Omega f d\mu = \sum_k b_k \mu(B_k) .$$

Proof. Let $\{a_j\}$ be the sequence of all distinct values in $\{b_k\}$, that is, $\{a_j\}$ is the sequence of all distinct values of f. Set

$$A_j = \{f = a_j\} .$$

Then

$$A_j = \bigsqcup_{\{k:b_k=a_j\}} B_k$$

and

$$\mu(A_j) = \sum_{\{k:b_k=a_j\}} \mu(B_k) ,$$

whence

$$\int_\Omega f d\mu = \sum_j a_j \mu(A_j) = \sum_j a_j \sum_{\{k:b_k=a_j\}} \mu(B_k) = \sum_j \sum_{\{k:b_k=a_j\}} b_k \mu(B_k) = \sum_k b_k \mu(B_k) .$$

■

Lemma 4.7 *For all non-negative simple functions f, g, the following is true.*

(a) For any positive constant c,

$$\int_\Omega cf d\mu = c \int_\Omega f d\mu$$

(b)

$$\int_\Omega (f+g) d\mu = \int_\Omega f d\mu + \int_\Omega g d\mu.$$

(*c*) *If* $f \leq g$ *then*

$$\int_\Omega f d\mu \leq \int_\Omega g d\mu.$$

(*d*)

$$\int_\Omega \min(f, n)\, d\mu \to \int_\Omega f\, d\mu \quad as\ n \to \infty. \tag{4.12}$$

Proof. (*a*) Let

$$f = \sum_k a_k 1_{A_k} \tag{4.13}$$

where $\bigsqcup_k A_k = \Omega$. Then

$$cf = \sum_k c a_k 1_{A_k}$$

whence

$$\int_\Omega cf d\mu = \sum_k c a_k \mu(A_k) = c \sum_k a_k \mu(A_k).$$

(*b*) Let f be as in (4.13) and $g = \sum_j b_j 1_{B_j}$ where $\bigsqcup_j B_j = \Omega$. Then

$$\Omega = \bigsqcup_{k,j} (A_k \cap B_j)$$

and on the set $A_k \cap B_j$ we have $f = a_k$ and $g = b_j$ so that $f + g = a_k + b_j$. Hence, $f + g$ is a non-negative simple function,

$$f + g = \sum_{k,j} (a_k + b_j)\, \mathbf{1}_{A_k \cap B_j},$$

whence by Lemma 4.6

$$\int_\Omega (f + g)\, d\mu = \sum_{k,j} (a_k + b_j)\, \mu(A_k \cap B_j).$$

Applying the same formula to functions f and g, we obtain

$$\int_\Omega f d\mu = \sum_{k,j} a_k \mu(A_k \cap B_j)$$

and

$$\int_\Omega g d\mu = \sum_{k,j} b_j \mu(A_k \cap B_j),$$

whence the claim follows.

(*c*) Clearly, $g - f$ is a non-negative simple functions so that by (*b*)

$$\int_\Omega g d\mu = \int_\Omega (g - f)\, d\mu + \int_\Omega f d\mu \geq \int_\Omega f d\mu.$$

(d) Let f be as in (4.13). Then we have

$$\min(f,n) = \sum_{\{k:a_k\le n\}} a_k \mathbf{1}_{A_k} + \sum_{\{k:a_k>n\}} n\mathbf{1}_{A_k} \ge \sum_{\{k:a_k\le n\}} a_k \mathbf{1}_{A_k},$$

whence by Lemma 4.6

$$\sum_{\{k:a_k\le n\}} a_k \mu(A_k) \le \int_\Omega \min(f,n)\, d\mu \le \int_\Omega f d\mu = \sum_k a_k \mu(A_k). \tag{4.14}$$

We are left to observe that, when $n\to\infty$, the series in the left hand side of (4.14) converges to the full series $\sum_k a_k\mu(A_k)$ (because for any index k the condition $a_k\le n$ is fulfilled for large enough n), whence (4.12) follows. ■

4.4.2 Non-negative measurable functions

Definition. Let $\{f_n\}_{n=1}^\infty$ be a sequence of real valued functions on Ω. We say that $\{f_n\}$ converges to a function f *uniformly* and write $f_n \rightrightarrows f$ if

$$\sup_\Omega |f_n - f| \to 0 \text{ as } n\to\infty.$$

Of course, the uniform convergence implies the pointwise convergence.

Definition. Let f be any non-negative measurable function on Ω. The Lebesgue integral of f is defined by

$$\int_\Omega f d\mu = \lim_{n\to\infty} \int_\Omega f_n d\mu$$

where $\{f_n\}$ is any sequence of non-negative simple functions such that $f_n \rightrightarrows f$ on Ω.

To justify this definition, we prove the following statement.

Lemma 4.8 *For any non-negative measurable functions f, there is a sequence of non-negative simple functions $\{f_n\}$ such that $f_n \rightrightarrows f$ on Ω. Moreover, for any such sequence the limit*

$$\lim_{n\to\infty} \int_\Omega f_n d\mu$$

exists with value in $[0,+\infty]$ and does not depend on the choice of the sequence $\{f_n\}$.

Proof. Fix the index $n\in\mathbb{N}$ and, for any non-negative integer k, consider the set

$$A_{k,n} = \left\{\omega\in\Omega : \frac{k}{n} \le f(\omega) < \frac{k+1}{n}\right\}.$$

Clearly, $\Omega = \bigsqcup_{k=0}^\infty A_{k,n}$. Define function f_n by

$$f_n = \sum_k \frac{k}{n} 1_{A_{k,n}},$$

that is, $f_n = \frac{k}{n}$ on $A_{k,n}$. Then f_n is a non-negative simple function and, on the set $A_{k,n}$, we have

$$0 \leq f - f_n < \frac{k+1}{n} - \frac{k}{n} = \frac{1}{n}$$

so that

$$\sup_{\Omega} |f - f_n| \leq \frac{1}{n}.$$

It follows that $f_n \rightrightarrows f$ on Ω, which proves the existence of such sequences.

Let now $\{f_n\}$ be any sequence of non-negative simple functions such that $f_n \rightrightarrows f$. Let us show that

$$\lim_{n\to\infty} \int_{\Omega} f_n d\mu \tag{4.15}$$

exists. The condition $f_n \rightrightarrows f$ on Ω implies that

$$\sup_{\Omega} |f_n - f_m| \to 0 \text{ as } n, m \to \infty.$$

In particular, $\sup |f_n - f_m|$ is finite provided n, m large enough. Since

$$f_m \leq f_n + \sup_{\Omega} |f_n - f_m|,$$

we obtain by Lemma 4.7

$$\int_{\Omega} f_m d\mu \leq \int_{\Omega} f_n d\mu + \sup_{\Omega} |f_n - f_m| \, \mu(\Omega). \tag{4.16}$$

If

$$\int_{\Omega} f_m d\mu < \infty$$

for all large enough m, then it follows from (4.16) and the analogous inequality with switched n, m that

$$\left| \int_{\Omega} f_m d\mu - \int_{\Omega} f_n d\mu \right| \leq \sup_{\Omega} |f_n - f_m| \, \mu(\Omega).$$

Therefore, the numerical sequence

$$\left\{ \int_{\Omega} f_n d\mu \right\}$$

is Cauchy and, hence, has a limit.

If

$$\int_{\Omega} f_m d\mu = +\infty$$

for some m, then (4.16) implies that $\int_{\Omega} f_n d\mu = +\infty$ for all large enough n, whence it follows that

$$\lim_{n\to\infty} \int_{\Omega} f_n d\mu = +\infty.$$

Hence, in the both cases the limit (4.15) exists.

Let now $\{f_n\}$ and $\{g_n\}$ be two sequences of non-negative simple functions such that $f_n \rightrightarrows f$ and $g_n \rightrightarrows f$. Let us show that

$$\lim_{n\to\infty} \int_\Omega f_n d\mu = \lim_{n\to\infty} \int_\Omega g_n d\mu. \tag{4.17}$$

Indeed, consider a mixed sequence $\{f_1, g_1, f_2, g_2, ...\}$. Obviously, this sequence converges uniformly to f. Hence, by the previous part of the proof, the sequence of integrals

$$\int_\Omega f_1 d\mu, \ \int_\Omega g_1 d\mu, \ \int_\Omega f_2 d\mu, \ \int_\Omega g_2 d\mu, ...$$

converges, which implies (4.17). ■

Hence, if f is a non-negative measurable function then the integral $\int_\Omega f d\mu$ is well-defined and takes value in $[0, +\infty]$.

Lecture 11
18.10.10

Example. Let $\Omega = [a, b]$ where $a < b$ and let λ be the Lebesgue measure on $[a, b]$. Let f be a non-negative continuous function on $[a, b]$. Then f is measurable so that the Lebesgue integral $\int_{[a,b]} f d\lambda$ is defined. Let us show that it coincides with the Riemann integral $\int_a^b f(x)\, dx$. Let $p = \{x_i\}_{i=0}^n$ be a partition of $[a, b]$ that is,

$$a = x_0 < x_1 < x_2 < ... < x_n = b$$

and let $\xi = \{\xi_i\}_{i=1}^n$ be a sequence of tags such that $\xi_i \in [x_{i-1}, x_i]$. The Riemann integral sum is defined by

$$S_*(f, p, \xi) = \sum_{i=1}^n f(\xi_i)(x_i - x_{i-1}),$$

and the Riemann integral is defined by

$$\int_a^b f(x)\, dx = \lim_{m(p)\to 0} S(f, p, \xi) \tag{4.18}$$

where $m(p) = \max_i |x_i - x_{i-1}|$ is the mesh of the partition, provided the above limit exists. It is known that the latter is the case for any continuous function f on $[a, b]$.

Consider now a simple function $F_{p,\xi}$ defined by

$$F_{p,\xi} = \sum_{i=1}^n f(\xi_i) \mathbf{1}_{(x_{i-1}, x_i]}.$$

Then we have

$$\int_{[a,b]} F_{p,\xi} d\lambda = \sum_{i=1}^n f(\xi_i) \lambda(x_{i-1}, x_i] = S(f, p, \xi).$$

On the other hand, by the uniform continuity of function f, we have $F_{p,\xi} \rightrightarrows f$ as $m(p) \to 0$, which implies by the definition of the Lebesgue integral that

$$\int_{[a,b]} f d\lambda = \lim_{m(p)\to 0} \int_{[a,b]} F_{p,\xi} d\lambda = \lim_{m(p)\to 0} S(f,p,\xi).$$

Comparing with (4.18) we obtain the identity

$$\int_{[a,b]} f \, d\lambda = \int_a^b f(x)\, dx.$$

Example. Consider on $[0,1]$ the Dirichlet function

$$f(x) = \begin{cases} 1, & x \in \mathbb{Q}, \\ 0, & x \notin \mathbb{Q}. \end{cases}$$

This function is not Riemann integrable because if one chooses the tags ξ_i to be rational then $S(f,p,\xi) = 1$ while for irrational ξ_i we have $S(f,p,\xi) = 0$ so that the limit (4.18) does not exist. However, the function f is non-negative and simple since it can be represented in the form $f = 1_A$ where $A = \mathbb{Q} \cap [0,1]$ is a Borel set. Therefore, the Lebesgue integral $\int_{[0,1]} f \, d\lambda$ is defined. Moreover, since A is a countable set, we have $\lambda(A) = 0$ and, hence, $\int_{[0,1]} f \, d\lambda = 0$.

The following statement extends Lemma 4.7 to non-negative measurable functions.

Theorem 4.9 *For all non-negative measurable functions f, g, the following is true.*
(a) *For any positive constant c,*

$$\int_\Omega cf d\mu = c \int_\Omega f d\mu.$$

(b)

$$\int_\Omega (f+g)\, d\mu = \int_\Omega f d\mu + \int_\Omega g d\mu.$$

(c) *If $f \le g$ then*

$$\int_\Omega f d\mu \le \int_\Omega g d\mu.$$

(d)

$$\int_\Omega \min(f,n)\, d\mu \to \int_\Omega f d\mu \quad \text{as } n \to \infty. \tag{4.19}$$

Proof. (a) By Lemma 4.8, there is a sequence $\{f_k\}$ of non-negative simple functions such that $f_k \rightrightarrows f$ on Ω. Then $cf_k \rightrightarrows cf$ and by Lemma 4.7,

$$\int_\Omega cf d\mu = \lim_{k\to\infty} \int_\Omega cf_k d\mu = \lim_{k\to\infty} c \int_\Omega f_k d\mu = c \int_\Omega f d\mu.$$

(b) Using sequences $\{f_k\}$ and $\{g_k\}$ of non-negative simple functions that converge uniformly to f and g, respectively, we obtain that

$$f_k + g_k \rightrightarrows f + g.$$

Therefore, by Lemma 4.7,

$$\int_\Omega (f+g)\, d\mu = \lim_{k\to\infty} \int_\Omega (f_k + g_k)\, d\mu = \lim_{k\to\infty} \left(\int_\Omega f_k d\mu + \int_\Omega g_k d\mu \right) = \int_\Omega f d\mu + \int_\Omega g d\mu.$$

(c) Clearly, $g - f$ is a non-negative simple functions so that by (b)

$$\int_\Omega g d\mu = \int_\Omega (g-f)\, d\mu + \int_\Omega f d\mu \geq \int_\Omega f d\mu.$$

(d) Let $\{f_k\}$ be a sequence of non-negative simple functions such that $f_k \rightrightarrows f$ on Ω. By construction of Lemma 4.8, we can assume that $f_k \leq f$.

Using (c) and Lemma 4.7, we obtain

$$\lim_{n\to\infty} \int_\Omega \min(f, n)\ d\mu \geq \lim_{n\to\infty} \int_\Omega \min(f_k, n)\ d\mu = \int_\Omega f_k d\mu.$$

Passing to the limit as $k \to \infty$, we obtain

$$\lim_{n\to\infty} \int_\Omega \min(f, n)\ d\mu \geq \int_\Omega f\, d\mu.$$

Since the opposite inequality is satisfied by (c), we obtain (4.19). ■

4.4.3 Integrable functions

To define the integral of a signed measurable function f on Ω, let us introduce the notation

$$f_+(\omega) = \begin{cases} f(\omega), & \text{if } f(\omega) \geq 0 \\ 0, & \text{if } f(\omega) < 0 \end{cases} \quad \text{and} \quad f_-(\omega) = \begin{cases} 0, & \text{if } f(\omega) \geq 0 \\ -f(\omega), & \text{if } f(\omega) < 0 \end{cases}.$$

The function f_+ is called the *positive part* of f and f_- is called the *negative part* of f. Note that f_+ and f_- are non-negative functions,

$$f = f_+ - f_- \quad \text{and} \quad |f| = f_+ + f_-.$$

It follows that

$$f_+ = \frac{|f| + f}{2} \text{ and } f_- = \frac{|f| - f}{2}.$$

In particular, if f is measurable then both functions f_+ and f_- are measurable.

Definition. A measurable function f on Ω is called (Lebesgue) *integrable* if

$$\int_\Omega f_+ d\mu < \infty \text{ and } \int_\Omega f_- d\mu < \infty.$$

For any integrable function, define its Lebesgue integral by

$$\int_\Omega f d\mu := \int_\Omega f_+ d\mu - \int_\Omega f_- d\mu.$$

Note that the integral $\int_\Omega f d\mu$ takes values in $(-\infty, +\infty)$.

In particular, if $f \geq 0$ then $f_+ = f$, $f_- = 0$ and f is integrable if and only if

$$\int_\Omega f d\mu < \infty.$$

Lemma 4.10 *Let f be a measurable function on Ω.*

(*a*) *The following conditions are equivalent:*

(*i*) *f is integrable.*

(*ii*) *f_+ and f_- are integrable,*

(*iii*) *$|f|$ is integrable.*

(*b*) *If f is integrable then*

$$\left| \int_\Omega f d\mu \right| \leq \int_\Omega |f| \, d\mu.$$

Proof. (*a*) The equivalence (*i*) $\Leftrightarrow$ (*ii*) holds by definition. Since $|f| = f_+ + f_-$, we have by Theorem 4.9

$$\int_\Omega |f| \, d\mu = \int_\Omega f_+ d\mu + \int_\Omega f_- d\mu.$$

It follows that

$$\int_\Omega |f| \, d\mu < \infty \;\Leftrightarrow\; \int_\Omega f_+ d\mu < \infty \;\text{ and }\; \int_\Omega f_- d\mu < \infty,$$

that is, (*ii*) $\Leftrightarrow$ (*iii*).

(*b*) We have

$$\left| \int_\Omega f d\mu \right| = \left| \int_\Omega f_+ d\mu - \int_\Omega f_- d\mu \right| \leq \int_\Omega f_+ d\mu + \int_\Omega f_- d\mu = \int_\Omega |f| \, d\mu.$$

■

Example. Let us show that if f is a continuous function on an interval $[a, b]$ and λ is the Lebesgue measure on $[a, b]$ then f is Lebesgue integrable. Indeed, f_+ and f_- are non-negative and continuous so that they are Lebesgue integrable by the above Example. Hence, f is also Lebesgue integrable. Moreover, we have

$$\int_{[a,b]} f \, d\mu = \int_{[a,b]} f_+ \, d\mu - \int_{[a,b]} f_- \, d\mu = \int_a^b f_+ \, d\mu - \int_a^b f_- \, d\mu = \int_a^b f \, d\mu$$

so that the Riemann and Lebesgue integrals of f coincide.

Theorem 4.11 *Let f, g be integrable functions on Ω.*

(a) For any real c, function cf is also integrable and

$$\int_\Omega cf d\mu = c\int_\Omega f d\mu.$$

(b) Function $f+g$ is integrable and

$$\int_\Omega (f+g)\, d\mu = \int_\Omega f d\mu + \int_\Omega g d\mu. \tag{4.20}$$

(c) If $f \leq g$ then

$$\int_\Omega f d\mu \leq \int_\Omega g d\mu.$$

(d)

$$(\inf f)\, \mu(\Omega) \leq \int_\Omega f\, d\mu \leq (\sup f)\, \mu(\Omega).$$

Proof. (a) If $c = 0$ then there is nothing to prove. Let $c > 0$. Then $(cf)_+ = cf_+$ and $(cf)_- = cf_-$ whence by Theorem 4.9

$$\int_\Omega cf d\mu = \int_\Omega cf_+ d\mu - \int_\Omega cf_- d\mu = c\int_\Omega f_+ d\mu - c\int_\Omega f_- d\mu = c\int_\Omega f d\mu.$$

If $c < 0$ then $(cf)_+ = |c|\, f_-$ and $(cf)_- = |c|\, f_+$ whence

$$\int_\Omega cf d\mu = \int_\Omega |c|\, f_- d\mu - \int_\Omega |c|\, f_+ d\mu = -|c| \int_\Omega f d\mu = c\int_\Omega f d\mu.$$

(b) Note that $(f+g)_+$ is not necessarily equal to $f_+ + g_+$ so that the previous simple argument does not work here. Using the triangle inequality

$$|f+g| \leq |f| + |g|,$$

we obtain

$$\int_\Omega |f+g|\, d\mu \leq \int_\Omega |f|\, d\mu + \int_\Omega |g|\, d\mu < \infty,$$

which implies that the function $f+g$ is integrable.

To prove (4.20), observe that

$$f_+ + g_+ - f_- - g_- = f + g = (f+g)_+ - (f+g)_-$$

whence

$$f_+ + g_+ + (f+g)_- = (f+g)_+ + f_- + g_-.$$

Since all the functions in the last identity are non-negative and measurable, we obtain by Theorem 4.9

$$\int_\Omega f_+ d\mu + \int_\Omega g_+ d\mu + \int_\Omega (f+g)_-\, d\mu = \int_\Omega (f+g)_+\, d\mu + \int_\Omega f_- d\mu + \int_\Omega g_- d\mu.$$

It follows that

$$\begin{aligned}\int_\Omega (f+g)\,d\mu &= \int_\Omega (f+g)_+\,d\mu - \int_\Omega (f+g)_-\,d\mu \\ &= \int_\Omega f_+ d\mu + \int_\Omega g_+ d\mu - \int_\Omega f_- d\mu - \int_\Omega g_- d\mu \\ &= \int_\Omega f d\mu + \int_\Omega g d\mu.\end{aligned}$$

(c) Using the identity $g = (g-f)+f$ and that $g - f \geq 0$, we obtain by part (b)

$$\int_\Omega g\,d\mu = \int_\Omega (g-f)\ d\mu + \int_\Omega f\,d\mu \geq \int_\Omega f\,d\mu.$$

(c) Consider a constant function $g \equiv \sup f$ so that $f \leq g$ on Ω. Since g is a constant, we have

$$\int_\Omega f\,d\mu \leq \int_\Omega g d\mu = (\sup f)\,\mu(\Omega).$$

In the same way one proves the lower bound. ■

4.5 Relation "almost everywhere"

Definition. We say that two measurable functions f, g are equal *almost everywhere* and write $f = g$ a.e. if

$$\mu\{f \neq g\} = 0.$$

In the same way we say that inequality $f \leq g$ is true almost everywhere and write $f \leq g$ a.e. if

$$\mu\{f > g\} = 0.$$

The following statement shows the connection of integration to the notion $f = g$ a.e.

Theorem 4.12 *Let f be a non-negative measurable function. Then $\int_\Omega f\,d\mu = 0$ is equivalent to $f = 0$* a.e.

Proof. Let $f = 0$ a.e..We have by definition

$$\int_\Omega f d\mu = \lim_{n\to\infty} \int_\Omega f_n d\mu,$$

where f_n is a simple function defined by

$$f_n = \sum_{k=0}^{\infty} \frac{k}{n} \mathbf{1}_{A_{k,n}},$$

where

$$A_{k,n} = \left\{ \frac{k}{n} \leq f < \frac{k+1}{n} \right\}.$$

The set $A_{k,n}$ has measure 0 if $k > 0$ whence it follows that

$$\int_\Omega f_n d\mu = \sum_{k=0}^{\infty} \frac{k}{n} \mu\left(A_{k,n}\right) = 0.$$

Hence, also $\int_\Omega f d\mu = 0$.

Now assume that $f = 0$ a.e. is not true, that is,

$$\mu\left\{f > 0\right\} > 0.$$

Since

$$\left\{f > 0\right\} = \bigcup_{\varepsilon > 0} \left\{f > \varepsilon\right\},$$

where ε can be assumed rational, there exists $\varepsilon > 0$ such that

$$\mu\left\{f > \varepsilon\right\} > 0.$$

Since $f \geq g := \varepsilon \mathbf{1}_{\{f > \varepsilon\}}$, it follows that

$$\int_\Omega f d\mu \geq \int_\Omega g d\mu = \varepsilon \mu\left(f > \varepsilon\right) > 0,$$

which finishes the proof. ■

Corollary 4.13 *If f is a measurable function, g is integrable, and $f = g$* a.e. *then f is also integrable and*

$$\int_\Omega f \, d\mu = \int_\Omega g \, d\mu. \tag{4.21}$$

Proof. Since $f - g = 0$ a.e. it follows that also $|f - g| = 0$ a.e.whence by Theorem 4.12 $\int_\Omega |f - g| \, d\mu = 0$. By Lemma 4.10, $f - g$ is integrable and $\int_\Omega (f - g) \, d\mu = 0$. Finally, by Theorem 4.11, function $f = g + (f - g)$ is integrable and (4.21) holds. ■

Hence, the functions that are equal almost everywhere, are indistinguishable for the Lebesgue integral.

Corollary 4.14 *If f, g are integrable and $f \leq g$* a.e. *then*

$$\int_\Omega f \, d\mu \leq \int_\Omega g \, d\mu.$$

Proof. Since $(g - f)_- = 0$ a.e., we obtain using Theorem 4.12

$$\begin{aligned} \int_\Omega g \, d\mu &= \int_\Omega f \, d\mu + \int_\Omega (g - f) \, d\mu \\ &= \int_\Omega f + \int_\Omega (g - f)_+ \, d\mu - \int_\Omega (g - f)_- \, d\mu \geq \int_\Omega f \, d\mu. \end{aligned}$$

■

Chapter 5

Random variables

Lecture 12
19.10.10

Let us fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition. A function $X : \Omega \to \mathbb{R}$ is called a *random variable* if X is $\mathcal{F}$-measurable.

5.1 The distribution of a random variable

The events have probabilities $\mathbb{P}(A)$. The analogue of that for random variables is the notion of *a distribution.* For any Borel set $A \subset \mathbb{R}$, the set

$$X^{-1}(A) = \{\omega : X(\omega) \in A\}$$

is measurable by Theorem 4.2. Hence, this set is an event, and we denote is shortly by $\{X \in A\}$. Define the quantity

$$P_X(A) := \mathbb{P}(X \in A).$$

Theorem 5.1 (*a*) *For any random variable X, P_X is a probability measure on $\mathcal{B}(\mathbb{R})$.*

(*b*) *Conversely, if μ is any probability measure on $\mathcal{B}(\mathbb{R})$, then there exists a probability space and a random variable X on it such that $P_X = \mu$.*

Proof. (*a*) By definition, we have $P_X(A) = \mathbb{P}(X^{-1}(A))$. Since X^{-1} preserves all set-theoretic operations, by this formula the probability measure $\mathbb{P}$ on $\mathcal{F}$ induces a probability measure on $\mathcal{B}(\mathbb{R})$ (see Exercise 4). Note also that

$$P_X(\mathbb{R}) = \mathbb{P}(X^{-1}(\mathbb{R})) = \mathbb{P}(\Omega) = 1.$$

(*b*) Consider the probability space

$$(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{R}, \mathcal{B}, \mu)$$

and the random variable on it

$$X(\omega) = \omega.$$

Then

$$P_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(\omega : X(\omega) \in A) = \mu(A).$$

■

Definition. The measure P_X is called *the distribution* of X.

Recall that by Theorem 2.10, any probability measure μ on $\mathcal{B}(\mathbb{R})$ is characterized by its distribution function

$$F(x) = \mu(-\infty, x].$$

Definition. For any random variable X define its *distribution function* $F_X(x)$ as the distribution function of measure P_X, that is,

$$F_X(x) = P_X(-\infty, x] = \mathbb{P}(X \leq x).$$

Recall that there is a class of absolutely continuous distributions on $\mathcal{B}(\mathbb{R})$ that are given by their densities. After having learned the Lebesgue integration, we can give the following definition.

Definition. We say that a distribution function $F(x)$ (and the corresponding measure μ) is absolutely continuous if there is a non-negative measurable function $f(x)$ on $\mathbb{R}$ such that, for all $x \in \mathbb{R}$,

$$F(x) = \int_{(-\infty, x]} f d\lambda, \tag{5.1}$$

where λ is the Lebesgue measure on $\mathbb{R}$. The function f is called the *density* of μ.

If function f is continuous (or piecewise continuous) then the Lebesgue integral in (5.1) coincides with the Riemann integral $\int_{-\infty}^{x} f(x)\, dx$.

Here are some well-known examples of distributions on $\mathcal{B}(\mathbb{R})$ that are defined by their densities (some of them we have seen already).

1. A uniform distribution $\mathcal{U}(a, b)$ on $[a, b]$:

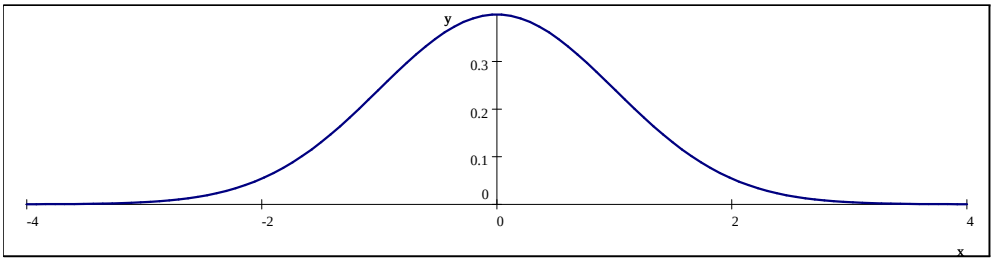
$$f(x) = \frac{1}{b-a} \text{ on } [a, b], \text{ and 0 otherwise.}$$

(It is obvious that $\int_{-\infty}^{+\infty} f(x)\, dx = 1$).

2. A normal distribution: $\mathcal{N}(0, 1)$

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

Here is the plot of this density function:



More generally, the normal distribution $\mathcal{N}(a,b)$ with parameters $a \in \mathbb{R}$ and $b > 0$ is defined by the density

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

Note that the identity

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

follows from

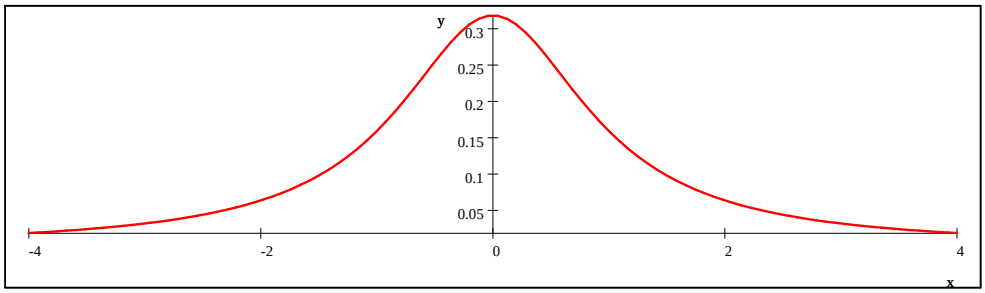
$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy = 1$$

by the change $y = \frac{x-a}{\sqrt{b}}$.

3. A Cauchy distribution

$$f(x) = \frac{1}{\pi(x^2+1)}.$$

Here is the plot of this function:



More generally, a Cauchy distribution with parameter a is given by the density

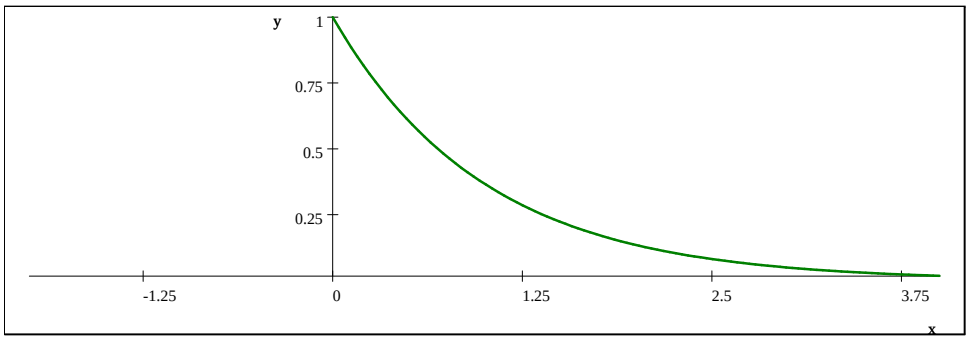
$$f(x) = \frac{a}{\pi(x^2+a^2)}.$$

Observe that $\int_{-\infty}^{+\infty} \frac{a}{\pi(x^2+a^2)} dx = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{dy}{y^2+1} = \frac{1}{\pi} [\arctan y]_{-\infty}^{+\infty} = 1$.

4. Exponential distribution with parameter $a > 0$:

$$f(x) = \begin{cases} ae^{-ax}, & x > 0, \\ 0, & x \leq 0. \end{cases}.$$

The case $a = 1$ is plotted here:



5. Gamma distribution with parameters $a, b > 0$:

$$f(x) = \begin{cases} c_{a,b} x^{a-1} \exp(-x/b), & x > 0, \\ 0, & x \leq 0 \end{cases},$$

where

$$c_{a,b} = \frac{1}{\Gamma(a) b^a}.$$

Observe that the gamma function Γ is defined by

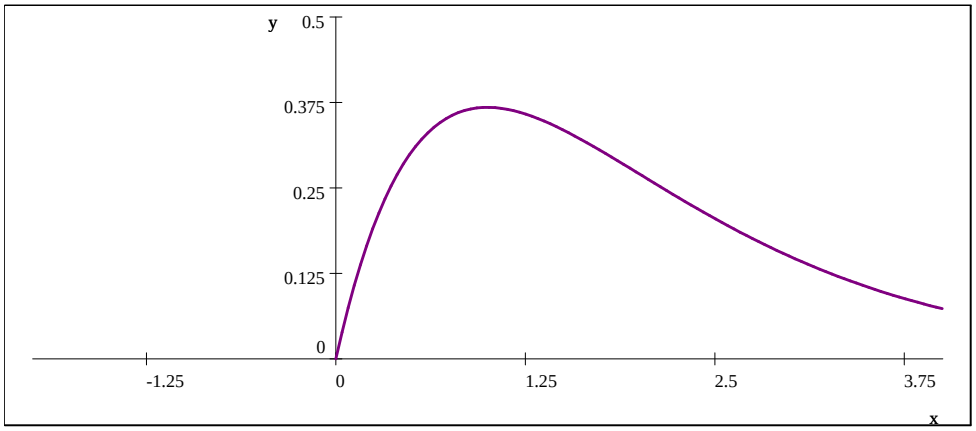
$$\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x) dx.$$

Hence, it follows from definition that

$$\int_{-\infty}^{\infty} f(x)\, dx = \frac{1}{\Gamma(a)\, b^a} \int_0^\infty x^{a-1} \exp(-x/b) dx = \frac{1}{\Gamma(a)} \int_0^\infty y^{a-1} \exp(-y)\, dy = 1,$$

so that f is indeed a density function. Note that $\Gamma(a) = (a-1)!$ if a is a positive integer. This follows from $\Gamma(1) = 1$ and from the identity $\Gamma(a+1) = a\Gamma(a)$ that holds for all $a > 0$.

For the case $a = 2$ and $b = 1$, we have $c_{a,b} = 1$ and $f(x) = xe^{-x}$, which is plotted below.



5.2 Absolutely continuous measures

Let μ be a measure on a σ-algebra $\mathcal{F}$ on Ω. For any non-negative measurable or integrable function f on Ω and for any set $A \in \mathcal{F}$ define the integral of f over A by

$$\int_A f d\mu = \int_\Omega f \mathbf{1}_A d\mu.$$

If f is a fixed non-negative integrable function then $\int_A f d\mu$ can be regarded as a non-negative function of A. By one of the Exercises, this function is a measure on $\mathcal{F}$; that is, if $A = \bigsqcup_{n=1}^{\infty} A_n$ then

$$\int_A f d\mu = \sum_{n=1}^{\infty} \int_{A_n} f d\mu.$$

Another useful observation is that if $\mu(A) = 0$ then $\int_A f d\mu = 0$, which follows from Theorem 4.12.

For what follows we will use the integration against the Lebesgue measure λ on $\mathcal{B}(\mathbb{R})$. For any bounded interval $I \subset \mathbb{R}$, measure λ is finite on $\mathcal{B}(I)$ and, hence, one can use the general definition of the Lebesgue integration. Hence, for any non-negative Borel function f on $\mathbb{R}$, the following integral

$$\int_I f d\lambda$$

is defined. Then set

$$\int_{\mathbb{R}} f d\lambda := \sum_{n=-\infty}^{+\infty} \int_{(n,n+1]} f d\lambda.$$

For any Borel set $A \subset \mathbb{R}$, set

$$\int_A f d\lambda := \int_{\mathbb{R}} f \mathbf{1}_A d\lambda = \sum_{n=-\infty}^{+\infty} \int_{(n,n+1]} f \mathbf{1}_{A \cap (n,n+1]} d\lambda.$$

Theorem 5.2 *Let μ be an absolutely continuous probability measure on $\mathcal{B}(\mathbb{R})$ with the density f. Then, for any Borel set $A \subset \mathbb{R}$*

$$\mu(A) = \int_A f d\lambda. \tag{5.2}$$

Moreover, for any non-negative Borel function g on $\mathbb{R}$,

$$\int_{\mathbb{R}} g d\mu = \int_{\mathbb{R}} g f d\lambda. \tag{5.3}$$

One writes therefore $d\mu = f d\lambda$.

Proof. For any $A \in \mathcal{B}(\mathbb{R})$ denote by $\nu(A)$ the right hand side of (5.2), that is,

$$\nu(A) = \int_{\mathbb{R}} f \mathbf{1}_A d\lambda.$$

We need to prove that $\nu(A) = \mu(A)$.

If $A = (-\infty, x]$ then by Definition of the density

$$\nu(-\infty, x] = \int_{(-\infty,x]} f d\lambda = F(x) = \mu(-\infty, x].$$

Denote by $\mathcal{F}$ the semi-algebra of all intervals of the form $(a, b]$. For any such interval we have

$$\nu(a, b] = F(b) - F(a) = \mu(a, b],$$

that is, $\nu = \mu$ on $\mathcal{F}$.

Let us now prove $\nu(A) = \mu(A)$ for all Borel sets A that are contained in some bounded interval I. Indeed, ν is a measure on $\mathcal{B}(I)$ as was remarked above, and so is μ. Since ν and μ coincide on all subintervals $(a, b] \subset I$, it follows by the uniqueness part of the Carathéodory extension theorem, that ν and μ coincide on $\mathcal{B}(I)$. Finally, for any $A \in \mathcal{B}(\mathbb{R})$ we have

$$\begin{aligned} \nu(A) &= \int_{\mathbb{R}} f\mathbf{1}_A d\lambda = \sum_n \int_{(n,n+1]} f\mathbf{1}_A d\lambda \\ &= \sum_n \nu(A \cap (n, n+1]) \\ &= \sum_n \mu(A \cap (n, n+1]) \\ &= \mu(A). \end{aligned}$$

■

5.3 Expectation and variance

Definition. If X is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ then its expectation is defined by

$$\mathbb{E}X = \int_\Omega X d\mathbb{P}, \tag{5.4}$$

provided the Lebesgue integral in the right hand side is defined.

Recall that there are two cases when the Lebesgue integral is defined. If $X \geq 0$ then $\mathbb{E}X$ is always defined and takes values in $[0, +\infty]$. If X is signed then then $\mathbb{E}X$ is defined by

$$\mathbb{E}X = \mathbb{E}X_+ - \mathbb{E}X_-,$$

provided X is integrable, that is, when both $\mathbb{E}X_+$, $\mathbb{E}X_-$ are finite, which is equivalent to

$$\mathbb{E}|X| < \infty.$$

The quantity $\mathbb{E}|X|$ is called the *first moment* of X. Similarly, the quantity $\mathbb{E}|X|^k$ is called the k-th moment of X.

By simple properties of Lebesgue integration (Theorems 4.9, 4.11) and a probability measure, we have the following properties of $\mathbb{E}$ (for integrable or non-negative random variables, whichever is appropriate):

1. $\mathbb{E}(cX) = c\mathbb{E}X$
2. $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$.
3. $\mathbb{E}\mathbf{1}_A = \mathbb{P}(A)$, in particular, $\mathbb{E}1 = 1$.

4. If $X \le Y$ then $\mathbb{E}X \le \mathbb{E}Y$; in particular $\inf X \le \mathbb{E}X \le \sup X$ and $|\mathbb{E}X| \le \mathbb{E}|X|$.

5. $\mathbb{E}\min(X, n) \to \mathbb{E}X$ as $n \to +\infty$

Definition. Given a random variable X, its *variance* is defined by

$$\operatorname{var} X = \mathbb{E}\left((X - \mathbb{E}X)^2\right), \tag{5.5}$$

assuming that $\mathbb{E}|X| < \infty$.

Clearly, the variance measures the quadratic mean deviation of X from its mean value $\mathbb{E}X$. Another useful expression for variance is the following:

Lemma 5.3 *If* $\mathbb{E}|X| < \infty$ *then*

$$\operatorname{var} X = \mathbb{E}\left(X^2\right) - (\mathbb{E}X)^2. \tag{5.6}$$

Proof. Indeed, from (5.5), we obtain

$$\operatorname{var} X = \mathbb{E}X^2 - 2\mathbb{E}(X\mathbb{E}X) + (\mathbb{E}X)^2 = \mathbb{E}X^2 - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

■

Corollary 5.4 *If* $\mathbb{E}|X| < \infty$ *then*

$$(\mathbb{E}X)^2 \le \mathbb{E}(X^2). \tag{5.7}$$

Proof. Indeed, (5.5) implies var $X \ge 0$ so that (5.7) follows from (5.6). ■
Alternatively, (5.7) follows from a more general inequality.

Theorem 5.5 (Cauchy-Schwarz inequality) *For any two random variables* X *and* Y*,*

$$(\mathbb{E}|XY|)^2 \le \mathbb{E}(X^2)\mathbb{E}(Y^2). \tag{5.8}$$

Proof. Without loss of generality, we may assume $X \ge 0$ and $Y \ge 0$. If one of the expectations $\mathbb{E}(X^2)$, $\mathbb{E}(Y^2)$ is infinite then there is nothing to prove. If both are finite then also

$$\mathbb{E}(XY) < \infty$$

which follows from the elementary inequality

$$XY \le \frac{1}{2}(X^2 + Y^2).$$

For any real t, consider the identity

$$X^2 + 2tXY + t^2Y^2 = (X + tY)^2.$$

Since the right hand here is non-negative and the left hand side has finite expectation, we obtain

$$\mathbb{E}X^2 + 2t\mathbb{E}(XY) + t^2\mathbb{E}Y^2 \geq 0.$$

The left hand side here can be regarded as a quadratic polynomial in t, which is non-negative for all real t. Hence, it discriminant is non-positive, which is exactly (5.8). ■

If we take in (5.8) $Y = 1$, we obtain

$$(\mathbb{E}\,|X|)^2 \leq \mathbb{E}(X^2) \tag{5.9}$$

which implies (5.7).

Remark. It follows from (5.6) and (5.9) that X has a finite variance if and only if X^2 is integrable (that is, X is square integrable). Indeed, if X has a finite variance then by definition X is integrable, and (5.6) implies $\mathbb{E}X^2 < \infty$. Conversely, if X is square integrable then by (5.9) X is integrable and by (5.6) the variance is finite.

The definition of expectation is convenient to prove the properties of $\mathbb{E}$ but no good for computation. For the latter, there is the following theorem.

Theorem 5.6 *For any random variable X, we have*

$$\mathbb{E}X = \int_{\mathbb{R}} x\, dP_X(x), \tag{5.10}$$

assuming that either $\mathbb{E}\,|X| < \infty$ or $\int_{-\infty}^{+\infty} |x|\ dP_X(x) < \infty$. Also, denoting $m = \mathbb{E}X$, we have

$$\text{var } X = \int_{\mathbb{R}} (x-m)^2 dP_X = \int_{\mathbb{R}} x^2 dP_X - \left(\int_{\mathbb{R}} x dP_X\right)^2. \tag{5.11}$$

Hence, to evaluate the expected value and the variance of a random variable, it suffices to know its distribution. If P_X has density f_X then using Theorem 5.2 we obtain

$$\mathbb{E}X = \int_{\mathbb{R}} x f_X(x)\, d\lambda(x).$$

If f is continuous or piecewise continuous then the Lebesgue integral amounts to the Riemann integral:

$$\mathbb{E}X = \int_{-\infty}^{+\infty} x f_X(x)\, dx. \tag{5.12}$$

Let P_X be a discrete distribution with the stochastic sequence $\{p_i\}$ and with atoms $\{x_i\}$. In other words, X takes only values x_i and $\mathbb{P}(X = x_i) = p_i$. Then X is a simple function on Ω, and (5.4) yields

$$\mathbb{E}X = \sum_{i=0}^{\infty} x_i p_i, \tag{5.13}$$

which can also be seen from (5.10).

Example. We claim that if $X \sim \mathcal{N}(a, b)$ then $\mathbb{E}X = a$ and $\operatorname{var} X = b$. Indeed, since

$$f_X(x) = \frac{1}{\sqrt{2\pi b}} \exp\left(-\frac{(x-a)^2}{2b}\right),$$

we have by (5.12)

$$\begin{aligned}
\mathbb{E}X &= \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2\pi b}} \exp\left(-\frac{(x-a)^2}{2b}\right) dx \\
&= \int_{-\infty}^{+\infty} \frac{y+a}{\sqrt{2\pi b}} \exp\left(-\frac{y^2}{2b}\right) dy \\
&= \int_{-\infty}^{+\infty} \frac{y}{\sqrt{2\pi b}} \exp\left(-\frac{y^2}{2b}\right) dy + a \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi b}} \exp\left(-\frac{y^2}{2b}\right) dy \\
&= 0 + a = a.
\end{aligned}$$

Then by (5.11)

$$\begin{aligned}
\operatorname{var} X &= \int_{-\infty}^{+\infty} \frac{(x-a)^2}{\sqrt{2\pi b}} \exp\left(-\frac{(x-a)^2}{2b}\right) dx \\
&= \int_{-\infty}^{+\infty} \frac{y^2}{\sqrt{2\pi b}} \exp\left(-\frac{y^2}{2b}\right) dy.
\end{aligned}$$

Denote $t = 1/b$ and recall the following identity

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-t\frac{y^2}{2}\right) dy = t^{-1/2}. \tag{5.14}$$

Differentiating it in t, we obtain

$$\int_{-\infty}^{+\infty} \frac{y^2}{2\sqrt{2\pi}} \exp\left(-t\frac{y^2}{2}\right) dy = \frac{1}{2} t^{-3/2}.$$

Substituting $t = 1/b$ and dividing by $\frac{1}{2} b^{1/2}$ we obtain

$$\int_{-\infty}^{+\infty} \frac{y^2}{\sqrt{2\pi b}} \exp\left(-\frac{y^2}{2b}\right) dy = b,$$

whence $\operatorname{var} X = b$ follows.

Example. We claim that if $X \sim Po(\lambda)$ (that is, X has the Poisson distribution with parameter $\lambda > 0$) then $\mathbb{E}X = \operatorname{var} X = \lambda$. Since P_X is given by the stochastic sequence $\left\{\frac{\lambda^i}{i!} e^{-\lambda}\right\}_{i=0}^{\infty}$ with atoms $x_i = i$, we obtain (5.13)

$$\mathbb{E}X = \sum_{i=0}^{\infty} i \frac{\lambda^i}{i!} e^{-\lambda} = \lambda \sum_{i=1}^{\infty} \frac{(\lambda-1)^i}{(i-1)!} e^{-\lambda} = \lambda.$$

Similarly, we have

$$\mathbb{E}\left(X\left(X-1\right)\right)=\sum_{i=0}^{\infty}i\left(i-1\right)\frac{\lambda^{i}}{i!}e^{-\lambda}=\lambda^{2}\sum_{i=2}^{\infty}\frac{\lambda^{i-2}}{(i-2)!}e^{-\lambda}=\lambda^{2}$$

and

$$\operatorname{var}X=\mathbb{E}X^{2}-\left(\mathbb{E}X\right)^{2}=\mathbb{E}\left(X\left(X-1\right)\right)+\mathbb{E}X-\mathbb{E}X^{2}=\lambda^{2}+\lambda-\lambda^{2}=\lambda.$$

Theorem 5.6 follows from a more general one.

Theorem 5.7 *Let X be a random variable and g be a non-negative Borel function on $\mathbb{R}$. Then $g(X)$ is also a random variable and*

$$\boxed{\mathbb{E}(g(X))=\int_{\mathbb{R}}gdP_{X}.} \tag{5.15}$$

Proof. Note that the function $g\left(X\right)$ is $\mathcal{F}$-measurable by Theorem 4.3. Consider first a particular case when g is a simple function on $\mathbb{R}$, say

$$g=\sum_{k=1}^{\infty}a_{k}\mathbf{1}_{A_{k}},$$

where A_k are Borel subsets of $\mathbb{R}$ that form a partition of $\mathbb{R}$ and $a_k\geq 0$. By definition, we have

$$\int_{\mathbb{R}}gdP_{X}=\sum_{k=1}^{\infty}a_{k}P_{X}\left(A_{k}\right).$$

Observe that the function

$$g\left(X\right)=\sum_{k=1}^{\infty}a_{k}\mathbf{1}_{A_{k}}\left(X\right)=\sum_{k=1}^{\infty}a_{k}\mathbf{1}_{\{X\in A_{k}\}}$$

is a simple function on Ω, whence it follows that

$$\mathbb{E}g(X)=\int_{\Omega}g\left(X\right)d\mathbb{P}=\sum_{k=1}^{\infty}a_{k}\mathbb{P}\left(X\in A_{k}\right).$$

Since

$$\mathbb{P}\left(X\in A_{k}\right)=P_{X}\left(A_{k}\right)$$

by the definition of distribution measure P_X, we obtain (5.15) for simple functions.

Let g be any non-negative Borel function. Then by Lemma 4.8 there is a sequence $\{g_n\}$ of simple non-negative Borel functions on $\mathbb{R}$ such that $g_n\rightrightarrows g$ as $n\to\infty$. Then $g_n\left(X\right)$ is a non-negative simple function on Ω and $g_n\left(X\right)\rightrightarrows g\left(X\right)$ as $n\to\infty$. By the previous part of the proof we have

$$\mathbb{E}g_{n}\left(X\right)=\int_{\mathbb{R}}g_{n}dP_{X}.$$

Passing to the limit as $n\to\infty$ and using again Lemma 4.8, we obtain the same identity for g, that is, (5.15). ■

Corollary 5.8 *Let X be a random variable and g be a Borel function on $\mathbb{R}$. Then the identity* (5.15) *is satisfied provided either*

$$\mathbb{E}\left|g\left(X\right)\right| < \infty \quad \text{or} \quad \int_{\mathbb{R}} |g| \, dP_X < \infty. \tag{5.16}$$

Proof. Applying Theorem 5.7 to $|g|$ we obtain that *both* conditions (5.16) are satisfied. Applying (5.15) for g_+ and g_- we obtain the same identity for g. ■

Finally, Theorem 5.6 follows upon application of (5.15) with functions $g\left(x\right) = x$, $g\left(x\right) = \left(x - m\right)^2$, $g\left(x\right) = x^2$.

Lecture 13
25.10.10

5.4 Random vectors and joint distributions

Let us generalize the notion of a random variable as follows.

Definition. A mapping $X : \Omega \to \mathbb{R}^n$ is called a *random vector* (or a vector-valued random variable) if X is $\mathcal{F}$-measurable.

By Theorem 4.2, X is a random vector if, for any Borel set $A \in \mathcal{B}(\mathbb{R}^n)$, we have $X^{-1}(A) \in \mathcal{F}$.

The relation between the notions of random variables and random vectors is given by the following statement.

Lemma 5.9 (*a*) *If $X_1, X_2, ..., X_n$ are random variables on Ω then the vector-valued function*

$$X = (X_1, X_2, ..., X_n) : \Omega \to \mathbb{R}^n \tag{5.17}$$

is a random vector.

(*b*) *Conversely, if X is a random vector then all its components are random variables.*

Proof. (*a*) By the definition of measurability, we need to check that, for all $c_1, ..., c_n$, the set

$$C = \{\omega \in \Omega : X\left(\omega\right) \in (-\infty, c_1] \times (-\infty, c_2] \times ... \times (-\infty, c_n]\}$$

is $\mathcal{F}$-measurable. Since

$$\begin{aligned} C &= \{X_1 \leq c_1, X_2 \leq c_2, ..., X_n \leq c_n\} \\ &= \bigcap_{k=1}^{n} \{X_k \leq c_k\} \end{aligned}$$

and each set $\{X_k \leq c_k\}$ is $\mathcal{F}$-measurable, we obtain that C is also $\mathcal{F}$-measurable.

(*b*) Let us show that X_1 is measurable. We have

$$\begin{aligned} \{X_1 \leq c\} &= \{X_1 \in (-\infty, c], X_2 \in (-\infty, +\infty), ..., X_n \in (-\infty, +\infty)\} \\ &= \{X \in (-\infty, c] \times (-\infty, +\infty) \times ... \times (-\infty, +\infty)\} \end{aligned}$$

and the latter set is $\mathcal{F}$-measurable by the measurability of X. ■

Hence, considering several random variables is equivalent to considering a random vector.

Similarly to the one-dimensional case, we introduce a distribution measure P_X on $\mathcal{B}(\mathbb{R}^n)$ by

$$P_X(A) = \mathbb{P}\{X \in A\}.$$

In particular, a distribution P_X of a random vector (5.17) can be regarded as a *joint distribution* of the random variables $X_1, X_2, ..., X_n$. Sometimes it is convenient to use the notation

$$P_X = P_{X_1 X_2 ... X_n}.$$

Theorem 5.10 (*a*) *If X is a random vector then P_X is a probability measure on $\mathcal{B}(\mathbb{R}^n)$.*

(*b*) *Conversely, if μ is any probability measure on $\mathcal{B}(\mathbb{R}^n)$ then there exists a random vector X such that $P_X = \mu$.*

The proof is similar to the proof of Theorem 5.1. The probability space in part (*b*) is $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$ and $\mathbb{P} = \mu$.

As in the one-dimensional case, we say that a measure μ on $\mathcal{B}(\mathbb{R}^n)$ is *absolutely continuous* with respect to the Lebesgue measure λ_n if there is a non-negative Borel function f such that, for all Borel sets A,

$$\mu(A) = \int_A f d\lambda_n.$$

The function f is called the *density* of μ with respect to λ_n, and one writes $d\mu = f d\lambda_n$.

If the distribution P_X of a random vector X has the density function then it is also referred to as the density of X and is normally denoted by f_X. If $X = (X_1, ..., X_n)$ then the density of X is called the joint density of $X_1, ..., X_n$ is is denoted by $f_{X_1 ... X_n}$.

Similarly to Theorems 5.2 and 5.7, we have the following statement.

Theorem 5.11 *If $X : \Omega \to \mathbb{R}^n$ is a random vector and g is a Borel function $g : \mathbb{R}^n \to \mathbb{R}$, then $g(X)$ is a random variable and*

$$\boxed{\mathbb{E}g(X) = \int_{\mathbb{R}^n} g dP_X}, \tag{5.18}$$

provided g is either non-negative or integrable with respect to P_X.

If in addition X has the density f_X then

$$\boxed{\mathbb{E}g(X) = \int_{\mathbb{R}^n} g f_X d\lambda_n}. \tag{5.19}$$

Before we can use these formulas, let us state some properties of integration with respect to the Lebesgue measure. In the next theorem, we identify $\mathbb{R}^n$ with the product $\mathbb{R}^m \times \mathbb{R}^{n-m}$ and represent any point $x \in \mathbb{R}^n$ in the form $x = (x', x'')$ where $x' = (x_1, ..., x_m)$ and $x'' = (x_{m+1}, ..., x_n)$.

Theorem 5.12 (Fubini's theorem)

(*a*) *Let f be a non-negative measurable (or integrable) function on $\mathbb{R}^n$. Then the following identity is true for any integer $1 \leq m < n$:*

$$\int_{\mathbb{R}^n} f d\lambda_n = \int_{\mathbb{R}^{m-m}} \left(\int_{\mathbb{R}^m} f(x', x'') \, d\lambda_m(x') \right) d\lambda_{n-m}(x'').$$

(*b*) *Let $(\Omega', \mathcal{F}', \mu')$ and $(\Omega'', \mathcal{F}'', \mu'')$ be two probability spaces and consider their product $(\Omega, \mathcal{F}, \mu)$ where $\Omega = \Omega' \times \Omega''$, $\mathcal{F} = \sigma(\mathcal{F}' \times \mathcal{F}'')$ and $\mu = \mu' \times \mu''$. Then for any non-negative $\mathcal{F}$-measurable (or integrable) function f on Ω,*

$$\int_{\Omega} f d\mu = \int_{\Omega''} \left(\int_{\Omega'} f(x', x'') \, d\mu'(x') \right) d\mu''(x'').$$

Both statements include the claims that the functions under integration in the right hand side are measurable (respectively, integrable). In fact, the function

$$x' \mapsto f(x', x'')$$

is measurable for almost all x'', and the function

$$x'' \mapsto \int_{\mathbb{R}^m} f(x', x'') \, d\mu'(x')$$

is also measurable.

The proof of Theorem 5.12 requires some advanced tools from measure theory and will not be given here.

For the next theorem we need the following notion. A mapping $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ is called a diffeomorphism if it is a bijection and both Φ and the inverse mapping Φ^{-1} are continuously differentiable. Recall also that if Φ is continuously differentiable then its total derivative Φ' at any point $x \in \mathbb{R}^n$ coincides with the Jacobi matrix, that is,

$$\Phi'(x) = \left(\frac{\partial \Phi_i}{\partial x_j}(x) \right)_{i,j=1}^{n}.$$

Theorem 5.13 (Change of variables in Lebesgue integral) *Let $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ be a diffeomorphism. Then for any non-negative measurable (or integrable) function $f : \mathbb{R}^n \to \mathbb{R}$, the function $f \circ \Phi : \mathbb{R}^n \to \mathbb{R}$ is measurable (resp. integrable) and*

$$\int_{\mathbb{R}^n} f \, d\lambda_n = \int_{\mathbb{R}^n} (f \circ \Phi) \left| \det \Phi' \right| \, d\lambda_n. \tag{5.20}$$

Using the notation $y = \Phi(x)$, we can rewrite (5.20) as follows:

$$\int_{\mathbb{R}^n} f(y)\, d\lambda_n(y) = \int_{\mathbb{R}^n} f(y(x)) \left|\det y'(x)\right| d\lambda_n(x).$$

For example, in the case $n = 1$ we obtain

$$\int_{\mathbb{R}} f(y)\, d\lambda(y) = \int_{\mathbb{R}} f(y(x)) \left|y'\right| dx. \tag{5.21}$$

Recall that for the Riemann integral one has the following formula for the change of variables

$$\int_{y(-\infty)}^{y(+\infty)} f(y)\, dy = \int_{-\infty}^{+\infty} f(y(x))\, y' dx. \tag{5.22}$$

Assuming that $y = y(x)$ is a diffeomorphism of $\mathbb{R}$ it is easy to see that (5.22) amounts to (5.21). Indeed, if y is an increasing diffeomorphism then $y' \geq 0$ and $y(\infty) = +\infty$, $y(-\infty) = -\infty$. Hence, both integrals in (5.22) are identical with the corresponding integrals in (5.21). If y is a decreasing diffeomorphism then $y' \leq 0$ and $y(\infty) = -\infty$, $y(-\infty) = +\infty$. Hence, after changing the signs of the both sides of (5.22) we obtain (5.21).

Let us return to random variables.

Corollary 5.14 *If the random variables $X_1, ..., X_n$ have the joint density function f then the variables $X_1, ..., X_m$ where $m < n$ have the joint density function*

$$f_{X_1 ... X_m}(x_1, ..., x_m) = \int_{\mathbb{R}^{n-m}} f(x_1, x_2, ..., x_n)\, d\lambda_{n-m}(x_{m+1}, ..., x_n). \tag{5.23}$$

In particular, X_1 has the density function

$$f_{X_1}(x) = \int_{\mathbb{R}^{n-1}} f(x, x_2, ..., x_n)\, d\lambda_{n-1}(x_2, ..., x_n). \tag{5.24}$$

Similar formulas hold for other components.

Proof. Fix a Borel set $A \in \mathcal{B}(\mathbb{R}^m)$ and consider function

$$g(x_1, ..., x_m, ..., x_n) = \mathbf{1}_{\{(x_1, ..., x_m) \in A\}}.$$

For this function we have

$$\mathbb{E}g(X_1, ..., X_n) = \mathbb{P}(X_1, ..., X_m \in A) = P_{X_1 ... X_m}(A).$$

On the other hand, since

$$\{(x_1, ..., x_m) \in A\} = \left\{(x_1, ..., x_m, x_{m+1}, ..., x_n) \in A \times \mathbb{R}^{n-m}\right\}$$

we have

$$g = \mathbf{1}_{A \times \mathbb{R}^{n-m}}.$$

Applying (5.19) to compute the above expectation we obtain

$$\begin{aligned} P_{X_1...X_m}(A) &= \int_{\mathbb{R}^n} \mathbf{1}_{A\times\mathbb{R}^{n-m}}\,(x_1,...,x_n)\, f(x_1,x_2,...,x_n) d\lambda_n\,(x_1,...,x_n) \\ &= \int_{A\times\mathbb{R}^{n-m}} f\,(x_1,...,x_n)\, d\lambda_n\,(x_1,...,x_n) \\ &= \int_A \left(\int_{\mathbb{R}^{n-mn-1}} f\,(x_1,...,x_n)\, d\lambda_{n-m}\,(x_{m+1},...,x_n) \right) d\lambda_m\,(x_1,...,x_m)\,. \end{aligned}$$

Hence, the function in the brackets (as a function of $x_1,...x_m$) is the joint density of $X_1,...X_m$, which proves (5.23). ■

Lecture 14
26.10.10

Example. Assume that random variables X, Y have the joint density function

$$f(x,y) = \frac{1}{2\pi} \exp\left(-\frac{x^2+y^2}{2}\right), \tag{5.25}$$

which is called the *2-dimensional normal distribution*. Note that

$$\begin{aligned} \int_{\mathbb{R}^2} f\,(x,y)\,dxdy &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \frac{1}{2\pi} \exp\left(-\frac{x^2+y^2}{2}\right) dy \right) dx \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \left(\exp\left(-\frac{x^2}{2}\right) \int_{\mathbb{R}} \exp\left(-\frac{y^2}{2}\right) dy \right) dx \\ &= \frac{\sqrt{2\pi}}{2\pi} \int_{\mathbb{R}} \exp\left(-\frac{x^2}{2}\right) dx \\ &= \frac{2\pi}{2\pi} = 1 \end{aligned}$$

so that indeed f is a density function. Then the random variable random vector X has the density

$$f_X(x) = \int_{\mathbb{R}} f(x,y)dy = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

which is the one-dimensional normal distribution.

Example. Let random variables X, Y have the joint density function

$$f\,(x,y) = \begin{cases} \frac{1}{2}(x+y)\exp\left(-x-y\right), & x,y \geq 0, \\ 0, \text{ otherwise.} \end{cases}$$

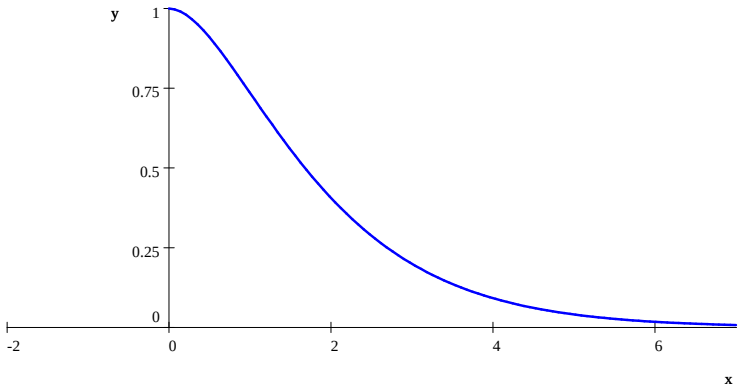
This is indeed a density function since

$$\begin{aligned} \int_{\mathbb{R}^2} f\,(x,y)\,dxdy &= \frac{1}{2}\int_0^\infty \left(\int_0^\infty (x+y)e^{-x-y} dy \right) dx \\ &= \frac{1}{2}\int_0^\infty \left(\int_0^\infty xe^{-x}e^{-y} dy \right) dx \\ &\quad + \frac{1}{2}\int_0^\infty \left(\int_0^\infty e^{-x}ye^{-y} dy \right) dx \\ &= \frac{1}{2}\int_0^\infty \left(xe^{-x} + e^{-x} \right) dx \\ &= 1. \end{aligned}$$

This calculation also shows that

$$f_X(x) = \int_{\mathbb{R}} f(x,y)dy = \frac{1}{2}\left(xe^{-x} + e^{-x}\right).$$

This function is plotted below:



Theorem 5.15 *Let $X : \Omega \to \mathbb{R}^n$ be a random vector with the density function f_X. Let $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ be a diffeomorphism. Then the random vector $Y = \Phi(X)$ has the following density function*

$$f_Y = f_X \circ \Phi^{-1} \left|\det\left(\Phi^{-1}\right)'\right|.$$

Proof. We have for any Borel set $A \subset \mathbb{R}^n$

$$\begin{aligned}
\mathbb{P}(\Phi(X) \in A) &= \mathbb{P}\left(X \in \Phi^{-1}(A)\right) \\
&= \int_{\Phi^{-1}(A)} f_X(x)\, d\lambda_n(x) \\
&= \int_{\mathbb{R}^n} \mathbf{1}_{\Phi^{-1}(A)} f_X(x)\, d\lambda_n(x).
\end{aligned}$$

Substituting $x = \Phi^{-1}(y)$ we obtain by Theorem 5.13 that the above integral is equal to

$$\begin{aligned}
&\int_{\mathbb{R}^n} \mathbf{1}_A f_X\left(\Phi^{-1}(y)\right) \left|\det\left(\Phi^{-1}(y)\right)'\right| d\lambda_n(y) \\
= &\int_A f_X\left(\Phi^{-1}(y)\right) \left|\det\left(\Phi^{-1}(y)\right)'\right| d\lambda_n(y)
\end{aligned}$$

whence the claim follows. ■

Example. If X has the density f_X then, for any non-zero real c, the random variable $Y = cX$ has the density

$$f_Y(x) = f_X\left(\frac{x}{c}\right)|c|^{-n}.$$

Indeed, using

$$\Phi(x) = cx$$

and noticing that $\Phi^{-1}(y) = \frac{1}{c}y$ and $\det(\Phi^{-1})' = \frac{1}{c^n}$, we obtain the claim.

In particular, if $X \sim \mathcal{N}(a, b)$, that is,

$$f_X(x) = \frac{1}{\sqrt{2\pi b}} \exp\left(-\frac{(x-a)^2}{2b}\right),$$

then

$$\begin{aligned} f_Y(x) &= \frac{1}{\sqrt{2\pi b}} \frac{1}{|c|} \exp\left(-\frac{(x/c-a)^2}{2b}\right) \\ &= \frac{1}{\sqrt{2\pi b c^2}} \exp\left(-\frac{(x-ca)^2}{2c^2 b}\right) \end{aligned}$$

so that $cX \sim \mathcal{N}(ca, c^2 b)$.

Example. Given random variables X, Y consider a random variable $U = X + Y$. The distribution function F_U can be obtain as follows:

$$F_U(s) = \mathbb{P}(X + Y \leq s) = \mathbb{E}\mathbf{1}_{\{X+Y\leq s\}} = \mathbb{E}g(X, Y),$$

where $g = \mathbf{1}_{\{x+y\leq s\}}$. Hence, we obtain from (5.18)

$$F_U(s) = \int_{\mathbb{R}^2} \mathbf{1}_{\{x+y\leq s\}} dP_{XY} = \int_{\{x+y\leq s\}} dP_{XY}.$$

Assume in addition that X, Y have the joint density $f(x, y)$. Then by (5.19)

$$F_U(s) = \int_{\mathbb{R}^2} \mathbf{1}_{\{x+y\leq s\}} f(x, y)\, d\lambda_2(x, y).$$

Consider the new coordinates

$$u = x + y, \quad v = y \tag{5.26}$$

that is

$$x = u - v,\ y = v.$$

Noticing that the Jacobi matrix of the latter transformation is

$$\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

and its determinant is 1, we obtain by (5.20)

$$\begin{aligned} F_U(s) &= \int_{\mathbb{R}^2} \mathbf{1}_{\{u\leq s\}} f(u - v, v)\, d\lambda_2(u, v) \\ &= \int_{\{u\leq s\}} \left(\int_{\mathbb{R}} f(u - v, v)\, d\lambda(v)\right) d\lambda(u). \end{aligned} \tag{5.27}$$

Since the density function $f_U(u)$ is determined by the identity

$$F_U(s) = \int_{(-\infty,s]} f_U d\lambda$$

we see that the interior integral in (5.27) is $f_U(u)$, that is,

$$f_U(u) = \int_{\mathbb{R}} f(u-v, v)\, d\lambda(v). \tag{5.28}$$

Similarly, $V = X - Y$ has the density

$$f_V(v) = \int_{\mathbb{R}} f(u+v, v)\, d\lambda(v).$$

Alternatively, we can use instead of (5.26) another change

$$u = x + y, \quad v = x - y,$$

that is equivalent to

$$x = \frac{u+v}{2},\ y = \frac{u-v}{2}.$$

Then the Jacobi matrix

$$\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}$$

has determinant $-\frac{1}{2}$, whence it follows that

$$F_U(s) = \frac{1}{2}\int_{\mathbb{R}^2} \mathbf{1}_{\{u\le s\}} f\left(\frac{u+v}{2}, \frac{u-v}{2}\right) d\lambda_2(u,v).$$

Hence, we obtain another formula for the density function of U:

$$f_U(u) = \frac{1}{2}\int_{\mathbb{R}} f\left(\frac{u+v}{2}, \frac{u-v}{2}\right) d\lambda(v). \tag{5.29}$$

Similarly, we have

$$f_V(v) = \frac{1}{2}\int_{\mathbb{R}} f\left(\frac{u+v}{2}, \frac{u-v}{2}\right) d\lambda(u) \tag{5.30}$$

Example. Let X, Y be again random vectors with joint density (5.25). Then by (5.29) we obtain the density of $U = X + Y$:

$$\begin{aligned} f_U(u) &= \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{(u+v)^2 + (u-v)^2}{8}\right) dv \\ &= \frac{1}{4\pi}\int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{4} - \frac{v^2}{4}\right) dv \\ &= \frac{1}{\sqrt{4\pi}} e^{-\frac{1}{4}u^2}. \end{aligned}$$

Hence, $X+Y\sim\mathcal{N}(0,2)$ and in the same way $X-Y\sim\mathcal{N}(0,2)$.

Example. Similarly one can handle the sum of n random variables $X_1,...,X_n$.. Indeed, the distribution function of the random variable $U=X_1+X_2+...+X_n$ can be obtained as follows:

$$F_U(u)=\mathbb{P}(X_1+X_2+...+X_n\leq u)=\mathbb{E}\mathbf{1}_{\{X_1+X_2+...+X_n\leq u\}}=\mathbb{E}g(X_1,...,X_n)$$

where $g=\mathbf{1}_{\{x_1+x_2+...+x_n\leq u\}}$. Hence, we obtain from (5.18)

$$F_U(u)=\int_{\mathbb{R}^n}\mathbf{1}_{\{x_1+x_2+...+x_n\leq u\}}dP_{X_1...X_n}=\int_{\{x_1+x_2+...+x_n\leq u\}}dP_{X_1...X_n}.$$

Assume in addition that $X_1,...,X_n$ have the joint density $f(x_1,...,x_n)$. Then by (5.19)

$$F_U(u)=\int_{\mathbb{R}^n}\mathbf{1}_{\{x_1+x_2+...+x_n\leq u\}}f(x_1,...x_n)\,d\lambda_n(x_1,...,x_n).$$

Passing to the new coordinates

$$u_1=x_1+...+x_n,\ u_2=x_2,...,\ u_n=x_n$$

and noticing that the determinant of this change is 1, we obtain

$$\begin{aligned}F_U(u)&=\int_{\mathbb{R}^n}\mathbf{1}_{\{u_1\leq u\}}f(u_1-u_2-...-u_n,u_2,...,u_n)\,d\lambda_n(u_1,....,u_n)\\&=\int_{\{u_1\leq u\}}\left(\int_{\mathbb{R}^{n-1}}f(u_1-u_2-...-u_n,u_2,...,u_n)\,d\lambda_{n-1}(u_2,....,u_n)\right)d\lambda(u_1),\end{aligned}$$

whence we obtain

$$f_U(u)=\int_{\mathbb{R}^{n-1}}f(u-u_2-...-u_n,u_2,...,u_n)\,d\lambda_{n-1}(u_2,....,u_n).$$

5.5 Independent random variables

Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space as before.

Definition. Two random vectors $X:\Omega\to\mathbb{R}^n$ and $Y:\Omega\to\mathbb{R}^m$ are called *independent* if, for all Borel sets $A\in\mathcal{B}(\mathbb{R}^n)$ and $B\in\mathcal{B}(\mathbb{R}^m)$, the events $\{X\in A\}$ and $\{Y\in B\}$ are independent, that is

$$\mathbb{P}(X\in A\text{ and }Y\in B)=\mathbb{P}(X\in A)\mathbb{P}(Y\in B).$$

Similarly, a sequence $\{X_i\}$ of random vectors $X_i:\Omega\to\mathbb{R}^{n_i}$ is called independent if, for any sequence $\{A_i\}$ of Borel sets $A_i\in\mathcal{B}(\mathbb{R}^n)$, the events $\{X_i\in A_i\}$ are independent. Here the index i runs in any index set (which may be finite, countable, or even uncountable).

If $X_1,...,X_k$ is a finite sequence of random vectors, such that $X_i:\Omega\to\mathbb{R}^{n_i}$ then we can form a vector $X=(X_1,...,X_k)$ whose components are those of all X_i; that is, X is an n-dimensional random vector where $n=n_1+...+n_k$. The distribution measure P_X of X is called the joint distribution of the sequence $X_1,...,X_k$ and is also denoted by $P_{X_1...X_k}$.

A particular case of the notion of a joint distribution for the case when all X_i are random variables, that is, $n_i=1$, was considered above.

Theorem 5.16 *Let X_i be a n_i-dimensional random vector, $i = 1, ..., k$. The sequence $X_1, X_2, ..., X_k$ is independent if and only if their joint distribution $P_{X_1...X_k}$ coincides with the product measure of $P_{X_1}, P_{X_2}, ..., P_{X_k}$, that is*

$$P_{X_1...X_k} = P_{X_1} \times ... \times P_{X_k}. \tag{5.31}$$

If $X_1, X_2, ..., X_k$ are independent and in addition X_i has the density function $f_{X_i}(x)$ then the sequence $X_1, ..., X_k$ has the joint density function

$$f(x) = f_{X_1}(x_1) f_{X_2}(x_2) ... f_{X_k}(x_k),$$

where $x_i \in \mathbb{R}^{n_i}$ and $x = (x_1, ..., x_k) \in \mathbb{R}^n$.

Proof. If (5.31) holds then, for any sequence $\{A_i\}_{i=1}^k$ of Borel sets $A_i \subset \mathbb{R}^{n_i}$, consider their product

$$A = A_1 \times A_2 \times ... \times A_k \subset \mathbb{R}^n \tag{5.32}$$

and observe that

$$\begin{aligned} \mathbb{P}(X_1 \in A_1, ..., X_k \in A_k) &= \mathbb{P}(X \in A) \\ &= P_X(A) \\ &= P_{X_1} \times ... \times P_{X_k}(A_1 \times ... \times A_k) \\ &= P_{X_1}(A_1) ... P_{X_k}(A_k) \\ &= \mathbb{P}(X_1 \in A_1) ... \mathbb{P}(X_k \in A_k). \end{aligned}$$

Hence, $X_1, ..., X_k$ are independent.

Conversely, if $X_1, ..., X_k$ are independent then, for any set A of the product form (5.32), we obtain

$$\begin{aligned} P_X(A) &= \mathbb{P}(X \in A) \\ &= \mathbb{P}(X_1 \in A_1,\ X_2 \in A_2,\ ...,\ X_k \in A_k) \\ &= \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2) ... \mathbb{P}(X_k \in A_k) \\ &= P_{X_1} \times ... \times P_{X_k}(A). \end{aligned}$$

Hence, the measure P_X and the product measure $P_{X_1} \times ... \times P_{X_k}$ coincide on the sets of the form (5.32). Since $\mathcal{B}(\mathbb{R}^n)$ is the minimal σ-algebra containing the sets (5.32), the uniqueness part of the Carathéodory extension theorem implies that these two measures coincide on $\mathcal{B}(\mathbb{R}^n)$, which was to be proved.

To prove the second claim, observe that, for any set A of the form (5.32), we have by Fubini's theorem

$$\begin{aligned} \mathbb{P}_X(A) &= P_{X_1}(A_1) ... P_{X_k}(A_k) \\ &= \int_{A_1} f_{X_1}(x_1)\, d\lambda_{n_1}(x_1) ... \int_{A_k} f_{X_k}(x_k)\, d\lambda_{n_k}(x_k) \\ &= \int_{A_1 \times ... \times A_k} f_1(x_1) ... f_k(x_k)\, d\lambda_n(x) \\ &= \int_A f(x)\, d\lambda_n(x) \\ &= : \mu(A), \end{aligned}$$

where $x = (x_1, ..., x_k) \in \mathbb{R}^n$. By Exercise 44, $\mu(A)$ is a measure on $\mathcal{B}(\mathbb{R}^n)$. Hence, the two measures $P_X(A)$ and $\mu(A)$ coincide on all boxes, which implies by the uniqueness part of the Carathéodory extension theorem that they coincide on all Borel sets $A \subset \mathbb{R}^n$. It follows that the function $f(x)$ is indeed the joint density. ■ Lecture 15 08.11.10

Corollary 5.17 *Let $\{\mu_i\}_{i=1}^k$ be a sequence probability measures on the spaces $\mathcal{B}(\mathbb{R}^{n_i})$. Then there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sequence $\{X_i\}_{i=1}^k$ of random vectors such that $X_i : \Omega \to \mathbb{R}^{n_i}$, $P_{X_i} = \mu_i$, and $\{X_i\}_{i=1}^k$ are independent.*

Proof. Consider product measure $\mu = \mu_1 \times ... \times \mu_k$ that is defined on $\mathcal{B}(\mathbb{R}^n)$ with $n = n_1 + ...+ n_k$. Since μ is a probability measure, by Theorem 5.10 there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random vector $X : \Omega \to \mathbb{R}^n$ such that $P_X = \mu$. We can represent X in the form $X = (X_1, ..., X_k)$ where X_1 is a vector that consists of the first n_1 components of X, X_2 consists of the next n_2 components, etc. Then $X_1, ..., X_k$ are random vectors with the joint distribution μ. Let us show that $P_{X_i} = \mu_i$. For example, for $i = 1$ we have for any $A \in \mathcal{B}(\mathbb{R}^{n_1})$

$$\begin{aligned} \mathbb{P}_{X_1}(A) &= \mathbb{P}(X_1 \in A) = \mathbb{P}(X_1 \in A, X_2 \in \Omega, ..., X_k \in \Omega) \\ &= \mathbb{P}(X \in A \times \Omega \times ... \times \Omega) \\ &= \mu(A \times \Omega \times ... \times \Omega) \\ &= \mu_1(A) \end{aligned}$$

In the same way one treats arbitrary i. Since μ is the product of $\mu_1, ..., \mu_k$, it follows that P_X is the product of $P_{X_1}, ..., P_{X_k}$. By Theorem 5.16 we obtain that $X_1, ..., X_k$ are independent. ■

In fact, the statement of Corollary 5.17 is true also for *infinite* sequences $\{\mu_i\}_{i=1}^\infty$, although the proof is much harder since one has to consider *infinite* products of probability measures and, hence, infinite dimensional probability spaces. Nevertheless, this can be done and results in the existence of an infinite sequence of independent random vectors with prescribed distributions.

In the next statement, we collect some more useful properties of independent random variables.

Theorem 5.18 *(a) If $X_i : \Omega \to \mathbb{R}^{n_i}$ is a sequence of independent random vectors and $f_i : \mathbb{R}^{n_i} \to \mathbb{R}^{m_i}$ is a sequence of Borel functions then the random vectors $\{f_i(X_i)\}$ are independent.*

(b) If $\{X_1, X_2, ..., X_n, Y_1, Y_2, .., Y_m\}$ is a sequence of independent random variables then the random vectors

$$X = (X_1, X_2, ..., X_n) \quad \text{and} \quad Y = (Y_1, Y_2, ..., Y_m)$$

are independent.

(c) Under conditions of (b), for all Borel functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$, the random variables $f(X_1, ..., X_n)$ and $g(Y_1, ..., Y_m)$ are independent.

Proof. (a) Let $\{A_i\}$ be a sequence of Borel sets, $A_i \in \mathcal{B}(\mathbb{R}^{m_i})$. We need to show that the events $\{f_i(X_i) \in A_i\}$ are independent. Since

$$\{f_i(X_i) \in A_i\} = \left\{X_i \in f_i^{-1}(A_i)\right\} = \{X_i \in B_i\},$$

where $B_i = f^{-1}(A_i) \in \mathcal{B}(\mathbb{R}^{n_i})$, these sets are independent by the definition of the independence of $\{X_i\}$.

(b) We have by Theorem 5.16

$$P_{XY} = P_{X_1} \times ... \times P_{X_n} \times P_{Y_1} \times ... \times P_{Y_m} = P_X \times P_Y.$$

Hence, X and Y are independent. Here we have used the fact that the product of measures is associative.

(c) The claim is an obvious combination of (a) and (b). ■

Theorem 5.19 *If X and Y are independent integrable random variables then*

$$\mathbb{E}(XY) = \mathbb{E}X\ \mathbb{E}Y \tag{5.33}$$

and

$$\operatorname{var}(X+Y) = \operatorname{var}X + \operatorname{var}Y. \tag{5.34}$$

Proof. Let us first show that XY is integrable, that is, $\mathbb{E}|XY| < \infty$. Applying Theorem 5.11 with $g(x,y) = |xy|$, Theorem 5.16, and Fubini's theorem, we obtain

$$\begin{aligned}\mathbb{E}|XY| &= \int_{\mathbb{R}^2} |xy|\, dP_{XY} = \int_{\mathbb{R}^2} |x|\,|y|\, d(P_X \times P_Y) = \left(\int_{\mathbb{R}} |y| \left(\int_{\mathbb{R}} |x|\, dP_X\right) dP_Y\right) \\ &= \left(\int_{\mathbb{R}} |x|\, dP_X\right)\left(\int_{\mathbb{R}} |y|\, dP_Y\right) = \mathbb{E}|X|\ \ \mathbb{E}|Y| < \infty.\end{aligned}$$

Now repeating the same computation with $g(x,y) = xy$, we obtain (5.33).

To prove the second claim, first observe that $\operatorname{var}(X+c) = \operatorname{var}X$ for any constant c. Hence, subtracting constants from X and Y, we can assume that $\mathbb{E}X = \mathbb{E}Y = 0$. In this case, we have $\operatorname{var}X = \mathbb{E}X^2$ and $\operatorname{var}Y = \mathbb{E}Y^2$. Using (5.33) we obtain

$$\operatorname{var}(X+Y) = \mathbb{E}(X+Y)^2 = \mathbb{E}X^2 + 2\mathbb{E}(XY) + \mathbb{E}Y^2 = \mathbb{E}X^2 + \mathbb{E}Y^2 = \operatorname{var}X + \operatorname{var}Y.$$

■

Remark. As it follows from the proof, the identity (5.33) holds for all independent non-negative random variables X, Y.

The identities (5.33) and (5.34) extend by induction to arbitrary finite sequence of independent integrable random variables $X_1, ..., X_n$ as follows:

$$\mathbb{E}(X_1...X_n) = \mathbb{E}X_1...\mathbb{E}X_n,$$

$$\operatorname{var}(X_1 + ... + X_n) = \operatorname{var}X_1 + ... + \operatorname{var}X_n\ .$$

Example. Without independence the identities (5.33) and (5.34) do not hold. Indeed, choose X to be a random variable with $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = 1$ and set $Y = X$. Then $\mathbb{E}(XY) = \mathbb{E}X^2 = 1$ whereas $\mathbb{E}X\ \mathbb{E}Y = 0$. Similarly,

$$\operatorname{var}(X+Y) = \mathbb{E}(X+Y)^2 = \mathbb{E}(2X)^2 = 4$$

whereas $\operatorname{var}X + \operatorname{var}Y = \mathbb{E}X^2 + \mathbb{E}Y^2 = 2$.

Example. Recall that if $X \sim \mathcal{N}(a,b)$ then $\mathbb{E}X = a$ and $\operatorname{var}X = b$. Let Y another random variable such that X and Y are independent and $Y \sim \mathcal{N}(a',b')$. Then we have

$$\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y = a + a',$$

and, using the independence of X, Y and (5.34),

$$\operatorname{var}(X+Y) = \operatorname{var}X + \operatorname{var}Y = b + b'.$$

In fact, one can prove that

$$X + Y \sim \mathcal{N}(a+a', b+b')$$

(see Exercise 42). In other words, the sum of two independent normally distributed random variables is again a normal random variable.

Example. A random variable X is called a Bernoulli random variable with parameter $p \in [0,1]$ if

$$\mathbb{P}(X=0) = 1-p \quad \text{and} \quad \mathbb{P}(X=1) = p.$$

This is equivalent to say that P_X is a discrete distribution with atoms 0 and 1 and with stochastic sequence $\{1-p, p\}$, that is, $P_X \sim B(p,1)$. We have then

$$\begin{aligned} \mathbb{E}X &= 1\cdot p + 0\cdot(1-p) = p, \\ \operatorname{var}X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = p - p^2. \end{aligned}$$

Consider a sequence $\{X_i\}_{i=1}^n$ of independent Bernoulli variables with the same parameter p (recall that such a sequence exists by Corollary 5.17). Set $S = X_1 + ... + X_n$ and prove that S has the binomial distribution $B(n,p)$, that is, for any $k = 0, 1, ..., n$

$$\mathbb{P}(S=k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{5.35}$$

Indeed, the sum S is equal to k if and only if exactly k of the values $X_1, ..., X_n$ are equal to 1 and $n-k$ are equal to 0. The probability, that the given k variables from $\{X_i\}$, say $X_{i_1}, ..., X_{i_k}$ are equal to 1 and the rest are equal to 0, is equal to $p^k(1-p)^{n-k}$. Since the sequence $(i_1, ..., i_k)$ can be chosen in $\binom{n}{k}$ ways, we obtain (5.35). Hence, $S \sim B(n,p)$.

On the other hand, we have

$$\mathbb{E}S = \sum_{i=1}^n \mathbb{E}X_i = pn$$

and by Theorem 5.19

$$\operatorname{var} S = \sum_{i=1}^{n} \operatorname{var} X_i = np\,(1-p)\,.$$

Hence, the above argument allows to obtain easily the expectation and variance of the binomial distribution $B\,(n,p)$.

5.6 Sequences of random variables

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables on a space $(\Omega, \mathcal{F}, \mathbb{P})$. By Theorem 4.4, if the sequence $\{X_n\,(\omega)\}$ converges for any $\omega \in \Omega$ then the limit function

$$X\,(\omega) = \lim_{n\to\infty} X_n\,(\omega)$$

is $\mathcal{F}$-measurable and, hence, is a random variable. In this case we say that X_n converges pointwise to X and write $X_n \to X$.

There are other *modes of convergence* of a sequence $\{X_n\}$ to a random variable X, which will be investigated here.

Definition. We say that X_n converges to X *in probability* and write

$$X_n \xrightarrow{\mathrm{P}} X,$$

if, for any $\varepsilon > 0$,

$$\mathbb{P}\,(|X_n - X| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

If some event A has probability 1 then one says that A occurs *almost surely* (write a.s.) or A occurs for *almost all* $\omega \in \Omega$ (write a.a.).

Definition. We say that X_n converges to X *almost surely* and write

$$X_n \xrightarrow{\text{a.s.}} X\,,$$

if

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

In other words, $X_n \xrightarrow{\text{a.s.}} X$ if

$$X_n(\omega) \to X(\omega) \text{ as } n \to \infty \text{ for almost all } \omega.$$

Clearly, the pointwise convergence $X_n \to X$ implies the convergence a.s..

Definition. We say that X_n converges to X in *the sense of Borel-Cantelli* and write $X_n \xrightarrow{\text{BC}} X$ if, for any $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}\,(|X_n - X| > \varepsilon) < \infty. \tag{5.36}$$

The condition (5.36) is called *the Borel-Cantelli condition.*

It is clear from comparison of the definitions that

$$X_n \xrightarrow{\text{BC}} X \;\Rightarrow\; X_n \xrightarrow{\text{P}} X.$$

Lecture 16
09.11.10

The following theorem states more detailed relations between the above modes of convergence.

Theorem 5.20 *The following is true:*

(a) *if* $X_n \xrightarrow{\text{BC}} X$ *then* $X_n \xrightarrow{\text{a.s.}} X$

(b) *if* $X_n \xrightarrow{\text{a.s.}} X$ *then* $X_n \xrightarrow{\text{P}} X$.

Before we prove the theorem, consider the following example.

Example. Let us show that in general

$$X_n \xrightarrow{\text{P}} X \not\Rightarrow X_n \xrightarrow{\text{a.s.}} X \not\Rightarrow X_n \xrightarrow{\text{BC}} X.$$

Let $\Omega = [0,1]$, $\mathcal{F}$ be σ-algebra of Borel sets and $\mathbb{P}$ be the Lebesgue measure on $[0,1]$. Let I_n be an interval in $[0,1]$ and define $X_n = \mathbf{1}_{I_n}$. For any $\varepsilon \in (0,1)$ we have

$$\mathbb{P}\left(|X_n| > \varepsilon\right) = \ell\left(I_n\right).$$

It follows that

$$X_n \xrightarrow{\text{P}} 0 \Leftrightarrow \lim_{n\to\infty} \ell\left(I_n\right) = 0,$$

$$X_n \xrightarrow{\text{BC}} 0 \Leftrightarrow \sum_{n=1}^{\infty} \ell\left(I_n\right) < \infty,$$

and $X_n \xrightarrow{\text{a.s.}} 0$ if almost all points $\omega \in I$ are covered by finitely many intervals I_n.

Choose the sequence of intervals I_n as follows:

$$\begin{gathered}
[0,1], \\
[0,\tfrac{1}{2}],[\tfrac{1}{2},1], \\
[0,\tfrac{1}{4}],[\tfrac{1}{4},\tfrac{2}{4}],[\tfrac{2}{4},\tfrac{3}{4}],[\tfrac{3}{4},1], \\
\ldots \\
[0,\tfrac{1}{2^m}],\ldots,[\tfrac{k}{2^m},\tfrac{k+1}{2^{m+1}}],\ldots,[\tfrac{2^m-1}{2^m},1]. \\
\ldots
\end{gathered}$$

Clearly, $\ell\left(I_n\right) \to 0$ as $n \to \infty$ so that $X_n \xrightarrow{\text{P}} 0$. On the other hand, each point $\omega \in [0,1]$ is covered by infinitely many of the intervals I_n so that $X_n(\omega)$ does not converge to 0, that is, the convergence $X_n \xrightarrow{\text{a.s.}} 0$ fails.

Consider another sequence of intervals $I_n = [0, 1/n]$. Since $\ell\left(I_n\right) \to 0$, we have $X_n \xrightarrow{\text{P}} 0$, but since $\sum_n \ell\left(I_n\right) = \infty$, the convergence $X_n \xrightarrow{\text{BC}} 0$ fails.

Before the proof of Theorem 5.20, we prove two lemmas.

Lemma 5.21 (1st Lemma of Borel-Cantelli) *Let $\{A_n\}$ be a sequence of events and consider the event*

$$\begin{aligned} I &= \{A_n \text{ occurs infinitely often}\} \\ &= \{\omega \in \Omega : \omega \in A_n \text{ for infinitely many values of } n\}. \end{aligned}$$

If

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \tag{5.37}$$

then $\mathbb{P}(I) = 0$.

Remark. This lemma is equivalent to Exercise 12(b).

Remark. Exercise 38 contains the 2nd lemma of Borel-Cantelli: if events A_n are independent and

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$$

then $\mathbb{P}(I) = 1$. Hence, for independent events $\{A_n\}$, the event "A_n occurs infinitely often" has the probability either 0 or 1.

For applications in this section we need only 1st lemma of Borel-Cantelli.

Proof of Lemma 5.21 It follows from the definition of I that

$$I = \{\omega \in \Omega : \forall N \in \mathbb{N} \ \ \exists n \geq N \ \ \omega \in A_n\} = \bigcap_{N \in \mathbb{N}} \bigcup_{n \geq N} A_n. \tag{5.38}$$

Since the sequence of events $\left\{ \bigcup_{n \geq N} A_n \right\}_{N=1}^{\infty}$ is monotone decreasing, we obtain by the continuity of $\mathbb{P}$

$$\mathbb{P}\left(\bigcap_{N} \bigcup_{n \geq N} A_n \right) = \lim_{N \to \infty} \mathbb{P}\left(\bigcup_{n \geq N} A_n \right).$$

By the condition (5.37), we have

$$\mathbb{P}\left(\bigcup_{n \geq N} A_n \right) \leq \sum_{n=N}^{\infty} \mathbb{P}(A_n) \to 0 \quad \text{as } N \to \infty,$$

whence it follows that $\mathbb{P}(I) = 0$, which was to be proved.

Lemma 5.22 *Given $\varepsilon > 0$, define for any positive integer n the event*

$$A_n(\varepsilon) = \{|X_n - X| > \varepsilon\} = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\}, \tag{5.39}$$

as well as the event

$$I(\varepsilon) = \{A_n(\varepsilon) \text{ occurs infinitely often}\}. \tag{5.40}$$

Then convergence $X_n \xrightarrow{\text{a.s.}} X$ is equivalent to the fact that $\mathbb{P}(I(\varepsilon)) = 0$ for any $\varepsilon > 0$.

Proof. We have by the definition of limit

$$\{X_n \to X\} = \{\forall \varepsilon > 0 \; \exists N \; \forall n \geq N \quad |X_n(\omega) - X(\omega)| \leq \varepsilon\}$$

whence by (5.39), (5.40) and (5.38)

$$\begin{aligned} \{X_n \not\to X\} &= \{\exists \varepsilon > 0 \; \forall N \; \exists n \geq N \quad |X_n(\omega) - X(\omega)| > \varepsilon\} \\ &= \{\exists \varepsilon > 0 \; \forall N \; \exists n \geq N \quad A_n(\varepsilon) \text{ occurs}\} \\ &= \{\exists \varepsilon > 0 \; I(\varepsilon) \text{ occurs}\} \\ &= \bigcup_{\varepsilon > 0} I(\varepsilon). \end{aligned}$$

Since $I(\varepsilon)$ is monotone decreasing in ε, the union in all $\varepsilon > 0$ can be replaced by the union in all rational ε so that it is a countable union. Hence,

$$X_n \xrightarrow{\text{a.s.}} X \Leftrightarrow \mathbb{P}(X_n \not\to X) = 0 \Leftrightarrow \mathbb{P}\left(\bigcup_{\varepsilon > 0} I(\varepsilon)\right) = 0 \Leftrightarrow \mathbb{P}(I(\varepsilon)) = 0 \text{ for all } \varepsilon > 0,$$

which was to be proved. ■

Proof of Theorem 5.20. We use in the proof the notation of Lemma 5.22. In these terms, we have the following equivalences:

$$X_n \xrightarrow{\text{BC}} X \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} \mathbb{P}(A_n(\varepsilon)) < \infty \tag{5.41}$$

$$X_n \xrightarrow{\text{a.s.}} X \quad \Leftrightarrow \quad \mathbb{P}(I(\varepsilon)) = 0 \text{ for all } \varepsilon > 0 \tag{5.42}$$

$$X_n \xrightarrow{\text{P}} X \quad \Leftrightarrow \quad \lim_{n \to \infty} \mathbb{P}(A_n(\varepsilon)) = 0 \text{ for all } \varepsilon > 0. \tag{5.43}$$

(a) By Lemma 5.21, (5.41) implies (5.42).
(b) By (5.42) and (5.38), we have

$$\mathbb{P}\left(\bigcap_N \bigcup_{n \geq N} A_n(\varepsilon)\right) = 0 \tag{5.44}$$

that is,

$$\lim_{N \to \infty} \mathbb{P}\left(\bigcup_{n \geq N} A_n(\varepsilon)\right) = 0, \tag{5.45}$$

which clearly implies (5.43). ■

A partial converse to Theorem 5.20 is given in the following statement.

Theorem 5.23 *If* $X_n \xrightarrow{\text{P}} X$ *then there exists a subsequence* $\{X_{n_k}\}$ *such that* $X_{n_k} \xrightarrow{\text{BC}} X$*; in particular,* $X_{n_k} \xrightarrow{\text{a.s.}} X$.

Proof. For any positive integer k, we have by (5.43)

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| > \frac{1}{k}\right) = 0.$$

It follows that, for any $k \in \mathbb{N}$ there exists N_k such that

$$\mathbb{P}\left(|X_n - X| > \frac{1}{k}\right) < 2^{-k} \quad \text{for all } n \geq N_k.$$

Therefore, there exists a sequence of positive integers $n_1 < n_2 < ... < n_k < ...$ such that

$$\mathbb{P}\left(|X_{n_k} - X| > \frac{1}{k}\right) < 2^{-k} \quad \text{for all } k = 1, 2, ...$$

(indeed, just set $n_k = N_1 + N_2 + ... + N_k$). Let us show that

$$X_{n_k} \xrightarrow{\text{BC}} X \text{ as } k \to \infty.$$

For any $\varepsilon > 0$, we have

$$\sum_{k=1}^{\infty} \mathbb{P}\left(|X_{n_k} - X| > \varepsilon\right) = \left(\sum_{k\leq 1/\varepsilon} + \sum_{k>1/\varepsilon}\right) \mathbb{P}\left(|X_{n_k} - X| > \varepsilon\right). \tag{5.46}$$

The first sum is finite, and the second sum converges because

$$\sum_{k>1/\varepsilon} \mathbb{P}\left(|X_{n_k} - X| > \varepsilon\right) \leq \sum_{k>1/\varepsilon} \mathbb{P}\left(|X_{n_k} - X| > \frac{1}{k}\right) \leq \sum_{k=1}^{\infty} 2^{-k} < \infty.$$

Hence, the series (5.46) converges, which was to be proved. ■

As an example of application of Theorem 5.20, let us prove the following useful theorem.

Theorem 5.24 (The bounded convergence theorem) *If* $X_n \xrightarrow{\text{a.s.}} X$ *and the sequence* $\{X_n\}$ *is bounded* a.s. *then* $\mathbb{E}\,|X_n - X| \to 0$ *and, hence,* $\mathbb{E}X_n \to \mathbb{E}X$.

Proof. The boundedness of the sequence $\{X_n\}$ a.s. means that, for some constant C, the inequality $|X_n| \leq C$ holds a.s. for every n. Note that also $|X| \leq C$ a.s., which follows from the observation that, for any $\varepsilon > 0$,

$$\mathbb{P}\left(|X| > C + \varepsilon\right) \leq \mathbb{P}\left(|X_n| > C\right) + \mathbb{P}\left(|X_n - X| > \varepsilon\right) \to 0 \text{ as } n \to \infty,$$

where we have used that, by Theorem 5.20, $X_n \xrightarrow{\text{P}} X$. Denoting $Y_n = |X_n - X|$ and noticing that $0 \leq Y_n \leq 2C$ a.s., we obtain using the monotonicity of expectation

$$\mathbb{E}Y_n = \mathbb{E}\left(Y_n \mathbf{1}_{\{Y_n > \varepsilon\}}\right) + \mathbb{E}\left(Y_n \mathbf{1}_{\{Y_n \leq \varepsilon\}}\right) \leq 2C\mathbb{P}\left(Y_n > \varepsilon\right) + \varepsilon, \tag{5.47}$$

where $\varepsilon > 0$ is arbitrary. Letting $n \to \infty$ and using that $\mathbb{P}\left(Y_n > \varepsilon\right) \to 0$, we obtain

$$\limsup_{n\to\infty} \mathbb{E}Y_n \leq \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, it follows that $\mathbb{E}Y_n \to 0$, that is, $\mathbb{E}\,|X_n - X| \to 0$. Since

$$|\mathbb{E}X_n - \mathbb{E}X| = |\mathbb{E}\,(X_n - X)| \leq \mathbb{E}\,|X_n - X|\,,$$

we obtain that $\mathbb{E}X_n \to \mathbb{E}X$. ■

Remark. In general (without boundedness) the condition $X_n \xrightarrow{\text{a.s.}} X$ or even $X_n \xrightarrow{\text{BC}} X$ does not imply that $\mathbb{E}X_n \to \mathbb{E}X$. Indeed, consider the following random variables on the interval $[0, 1]$:

$$X_n = 2^n \mathbf{1}_{(0,2^{-n})}.$$

Then $X_n \xrightarrow{\text{BC}} 0$ because, for any $\varepsilon > 0$, $\mathbb{P}\,(|X_n| > \varepsilon) \leq 2^{-n}$ so that (5.36) is satisfied. However, $\mathbb{E}X_n = 1$ while $\mathbb{E}X = 0$.

Chapter 6

Laws of large numbers

Lecture 17
15.11.10

In this chapter we will be concerned with sums

$$S_n := X_1 + X_2 + ... + X_n$$

where $\{X_n\}$ is a sequence of independent random variables. The sums of independent random variables arise in many application, and their investigation constitutes a large portion of Probability Theory.

Consider some examples.

1. Suppose $X_n = 1$ if at n-th flipping the coin shows heads, and $X_n = 0$ otherwise. Assuming that heads show with probability p and tails with $1 - p$, we see that $\{X_n\}$ is independent sequence of Bernoulli variables, having the same distribution: $X_n = 1$ with probability p and $X_n = 0$ with probability $1-p$. Then S_n is just a number of heads in a series of n trials. One may conjecture that $S_n \approx pn$, which will be justified below.

2. Assume that one gambles on coin flipping and wins a euro each time the coin shows heads and looses b euro in the case of tails. Then define the random variable X_n as follows: $X_n = a$ in if the n-th flipping shows heads and $X_n = -b$ in the case of tails. Then the sum S_n is the amount won after n flips (or lost if $S_n < 0$). Obviously, the behavior of S_n for large n determines the outcome of gambling.

3. Consider a *random walk* on the set of integers $\mathbb{Z}$, which is a simple model of Brownian motion. A particle moves on the nodes of $\mathbb{Z}$ as follows. At time 0 it is at 0. At each integer time $n \geq 1$, it jumps from the current position s to $s + 1$ with probability p, and to $s - 1$ with probability $1 - p$. Let us say that $X_n = 1$ for the first possibility, and $X_n = -1$ for the second. Then S_n is a current position of a particle at time n. Investigating the sums S_n, we may be able to predict the behavior of the random walk as $n \to \infty$.

4. Suppose that a computer performs a task of adding up a long sequence of n real numbers. At each operation there is a rounding error; denote by X_k the error after k-th operation. Then the error after n operations is S_n. Assuming that $|X_k| \leq \varepsilon$ for all k, normally one estimates S_n very roughly by $|S_n| \leq n\varepsilon$. However, if the errors at different operations are independent (which seems a reasonable assumption) then S_n can be significantly smaller than $n\varepsilon$ due to cancellations. As we will see later, it is reasonable to expect $|S_n| \approx \sqrt{n}$.

In this Chapter we investigate the behavior of S_n for large n for general random variables X_k and prove the results that are referred to as *laws of large numbers.*

6.1 The weak law of large numbers

The first result about the sums of independent random variables is the following.

Theorem 6.1 *Let $\{X_n\}$ be independent sequence of random variables having a common finite expectation $\mathbb{E}X_n = a$ and a common finite variance $\operatorname{var} X_n = b^2$. Then*

$$\frac{S_n}{n} \xrightarrow{\text{P}} a \quad \text{as } n \to \infty. \tag{6.1}$$

Recall that, by the definition of convergence in probability, (6.1) is equivalent to the following: for any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - a\right| > \varepsilon\right) = 0. \tag{6.2}$$

The latter can be equivalently rewritten in one of the following forms:

$$\lim_{n\to\infty} \mathbb{P}\left(|S_n - an| > \varepsilon n\right) = 0$$

or

$$\lim_{n\to\infty} \mathbb{P}\left((a-\varepsilon)\, n \le S_n \le (a+\varepsilon)\, n\right) = 1. \tag{6.3}$$

On can interpret (6.3) as follows: for large n one has $S_n \approx an$ with the error $\le \varepsilon n$. For example, in the case of coin flipping, we have $a = p$ so that $S_n \approx pn$. In the case of random walk (or a long computation) we have $a = 0$, and (6.3) says that it is likely that $|S_n| \le \varepsilon n$. In other words, even if a particle can move away, the rate of that is sublinear.

Example. Assume that all X_k have the same normal distribution $\mathcal{N}(0,1)$, that is, $\mathbb{E}X_k = 0$ and $\operatorname{var} X_k = \mathbb{E}X_k^2 = 1$. Then $S_n \sim \mathcal{N}(0,n)$ so that

$$f_{S_n}(x) = \frac{1}{\sqrt{2\pi n}} \exp\left(-\frac{x^2}{2n}\right).$$

It follows that

$$\begin{aligned}
\mathbb{P}\left(|S_n| > \varepsilon n\right) &= 2\int_{\varepsilon n}^{\infty} \frac{1}{(2\pi n)^{1/2}} \exp\left(-\frac{x^2}{2n}\right) dx \quad \left(\text{change } y = x/\sqrt{n}\right) \\
&= 2\int_{\varepsilon\sqrt{n}}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{y^2}{2}\right) dy \\
&\to 0 \text{ as } n \to \infty,
\end{aligned}$$

whence (6.1) follows. Moreover, for any $\delta > 0$, we have in the same way

$$\mathbb{P}\left(|S_n| > \varepsilon n^{1/2+\delta}\right) = 2\int_{\varepsilon n^{\delta}}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{y^2}{2}\right) dy \to 0 \quad \text{as } n \to \infty$$

so that also

$$\frac{S_n}{n^{1/2+\delta}} \xrightarrow{\mathrm{P}} 0 \text{ as } n \to \infty.$$

For example, if X_k is the computational error at step k then, assuming that all errors are independent and identically normally distributed, we obtain that the error of the sum is $o\left(n^{1/2+\delta}\right)$ for any $\delta > 0$. However, $\frac{S_n}{\sqrt{n}}$ does not go to 0 as $n \to \infty$ because

$$\mathbb{P}\left(|S_n| > \varepsilon n^{1/2}\right) = 2\int_{\varepsilon}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{y^2}{2}\right) dy = \text{const} > 0.$$

Example. Recall that a Cauchy distribution $Cauchy\,(a)$ with parameter a has the density

$$f\,(x) = \frac{1}{\pi}\frac{a}{a^2+x^2}.$$

Note that the Cauchy distribution has no expectation but nevertheless in some sense its mean is 0 because it is even. It is possible to prove that if $X \sim Cauchy\,(a)$ and $Y \sim Cauchy\,(b)$ and X, Y are independent then $X + Y \sim Cauchy\,(a+b)$. Assume now that all $X_k \sim Cauchy\,(1)$. Then $S_n \sim Cauchy\,(n)$. Then

$$\begin{aligned}\mathbb{P}\left(|S_n| > \varepsilon n\right) &= \frac{2}{\pi}\int_{\varepsilon n}^{\infty} \frac{n}{n^2+x^2} dx \quad (\text{change } y = x/n) \\ &= \frac{2}{\pi}\int_{\varepsilon}^{\infty} \frac{1}{1+y^2} dy = \text{const} > 0.\end{aligned}$$

Hence, $\frac{S_n}{n}$ does not converge to 0 as $n \to \infty$. It is easy to see that in fact

$$\frac{S_n}{n^{1+\delta}} \xrightarrow{\mathrm{P}} 0 \text{ as } n \to \infty$$

for any $\delta > 0$.

Before the proof of Theorem 6.1, let us prove the following lemma.

Lemma 6.2 (Chebyshev inequality) *Let Y be a random variable. Then, for all positive k and t,*

$$\mathbb{P}(|Y| \geq t) = \frac{1}{t^k}\mathbb{E}\,|Y|^k\,. \tag{6.4}$$

Recall that the quantity $\mathbb{E}\,|Y|^k$ is called the k-th moment of Y.

Proof. Renaming $|Y|$ to Y, we can assume that Y is non-negative. Let us prove (6.4) first in the case $k = 1$, that is,

$$\mathbb{P}\,(Y \geq t) \leq \frac{1}{t}\mathbb{E}Y. \tag{6.5}$$

Indeed, we have

$$Y \geq \mathbf{1}_{\{Y \geq t\}} Y \geq \mathbf{1}_{\{Y \geq t\}} t$$

whence it follows that

$$\mathbb{E}Y \geq \mathbb{E}\left(\mathbf{1}_{\{Y\geq t\}}t\right) = t\mathbb{P}\left(Y \geq t\right),$$

which proves (6.5). Applying (6.5) to Y^k instead of Y and to t^k instead of t, we obtain

$$\mathbb{P}\left(Y \geq t\right) = \mathbb{P}\left(Y^k \geq t^k\right) \leq \frac{1}{t^k}\mathbb{E}Y^k,$$

which was to be proved. ■

Proof of Theorem 6.1. We have

$$\mathbb{E}S_n = \sum_{k=1}^{n} \mathbb{E}X_k = an$$

and, by Theorem 5.19,

$$\operatorname{var} S_n = \sum_{k=1}^{n} \operatorname{var} X_k = b^2 n.$$

Let us apply (6.4) with $Y = |S_n - an|$ and the second moment. Observing that

$$\mathbb{E}Y^2 = \mathbb{E}\left(S_n - an\right)^2 = \operatorname{var} S_n$$

we obtain

$$\mathbb{P}\left(\left|\frac{S_n}{n} - a\right| \geq \varepsilon\right) = \mathbb{P}\left(Y \geq \varepsilon n\right) \leq \frac{1}{\left(\varepsilon n\right)^2}\mathbb{E}Y^2 = \frac{\operatorname{var} S_n}{\left(\varepsilon n\right)^2} = \frac{b^2 n}{\left(\varepsilon n\right)^2} = \frac{b^2}{\varepsilon^2 n}.$$

Hence,

$$\boxed{\mathbb{P}\left(\left|\frac{S_n}{n} - a\right| \geq \varepsilon\right) \leq \frac{b^2}{\varepsilon^2 n}}, \tag{6.6}$$

whence (6.2) follows. ■

The weak law of large numbers has the following more sophisticated version which will be proved in Section 8.8.

THEOREM . *If $\{X_n\}$ are independent identically distributed random variables with a common finite expectation a then*

$$\frac{S_n}{n} \xrightarrow{\text{P}} a \quad \textit{as} \quad n \to \infty.$$

The difference with Theorem 6.1 is that one does not need here the finiteness of the variance, at the expense of having the same distribution function for all random variables X_n.

6.2 The Weierstrass approximation theorem

Here we show how the weak law of large numbers allows to prove the following purely analytic theorem.

Theorem 6.3 (The Weierstrass approximation theorem) *Let f be a continuous function on a bounded closed interval $[a,b]$. Then, for any $\varepsilon > 0$, there exists a polynomial $P(x)$ such that*

$$\sup_{x\in[a,b]} |f(x) - P(x)| < \varepsilon.$$

Proof. Suffices to consider the case of the interval $[0,1]$. Consider the sequence $\{X_n\}$ of independent Bernoulli variables taking 1 with probability p, and 0 with probability $1-p$. As we know, the sum $S_n = X_1 + ... + X_n$ has the binomial distribution $B(n,p)$, that is,

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1..., n.$$

It follows that

$$\mathbb{E} f(\frac{S_n}{n}) = \sum_{k=0}^{n} f(\frac{k}{n}) \mathbb{P}(S_n = k) = \sum_{k=0}^{n} f(\frac{k}{n}) \binom{n}{k} p^k (1-p)^{n-k}.$$

The right hand side here can be considered as a polynomial in p. Denote it by

$$B_n(p) = \sum_{k=0}^{n} f(\frac{k}{n}) \binom{n}{k} p^k (1-p)^{n-k}.$$

The polynomial $B_n(p)$ is called the *Bernstein polynomial* of f. It turns out to be a good approximation for $f(p)$. The idea is that S_n/n converges in some sense to p as $n \to \infty$. Therefore, we may expect that $\mathbb{E} f(\frac{S_n}{n})$ converges to $f(p)$. To be precise, we will prove that

$$\lim_{n\to\infty} \sup_{p\in[0,1]} |f(p) - B_n(p)| = 0, \tag{6.7}$$

which will settle the claim.

To prove (6.7), first observe that any continuous function f on $[0,1]$ is uniformly continuous, that is, for any $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$|x - y| \le \delta \Rightarrow |f(x) - f(y)| \le \varepsilon.$$

Using the binomial theorem, we obtain

$$f(p) = f(p)\,(p + (1-p))^n = \sum_{k=0}^{n} f(p) \binom{n}{k} p^k (1-p)^{n-k}.$$

Therefore,

$$\begin{aligned} |f(p) - B_n(p)| &\leq \sum_{k=0}^{n} \left| f(p) - f(\frac{k}{n}) \right| \binom{n}{k} p^k (1-p)^{n-k} \\ &= \left(\sum_{\left|\frac{k}{n}-p\right| \leq \delta} + \sum_{\left|\frac{k}{n}-p\right| > \delta} \right) \left| f(p) - f(\frac{k}{n}) \right| \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned}$$

In the first sum, we have by the choice of δ that

$$\left| f(p) - f(\frac{k}{n}) \right| \leq \varepsilon,$$

so that the sum is bounded by ε.

In the second sum, we use the fact that f is bounded, that is, $C := \sup |f| < \infty$, whence

$$\left| f(p) - f(\frac{k}{n}) \right| \leq 2C.$$

Therefore, the second sum is bounded by

$$2C \sum_{\left|\frac{k}{n}-p\right| > \delta} \binom{n}{k} p^k (1-p)^{n-k} = 2C \sum_{\left|\frac{k}{n}-p\right| > \delta} \mathbb{P}(S_n = k) = 2C\mathbb{P}\left(\left| \frac{S_n}{n} - p \right| > \delta \right).$$

Using the estimate (6.6) and $\mathbb{E}X_i = p$, we obtain

$$\mathbb{P}\left(\left| \frac{S_n}{n} - p \right| > \delta \right) \leq \frac{b^2}{\delta^2 n},$$

where $b^2 = \operatorname{var} X_i = p(1-p) < 1$.

Hence, for all $p \in [0,1]$,

$$|f(p) - B_n(p)| \leq \varepsilon + \frac{2C}{\delta^2 n}.$$

Letting $n \to \infty$, we obtain

$$\limsup_{n\to\infty} |f(p) - B_n(p)| \leq \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, it follows that

$$\lim_{n\to\infty} |f(p) - B_n(p)| = 0,$$

which was to be proved. ■

Remark. An upper bound for the sum

$$\sum_{\left|\frac{k}{n}-p\right| > \delta} \binom{n}{k} p^k (1-p)^{n-k}$$

can be proved also analytically, which gives also another proof of the weak law of large numbers in the case when all X_n are Bernoulli variables. Such a proof was found by Jacob Bernoulli and historically was the first proof of the weak law of large numbers (in this particular case).

Lecture 18
16.11.10

6.3 The strong law of large numbers

As before, let $\{X_n\}$ be a sequence of random variables and

$$S_n = X_1 + X_2 + ... + X_n.$$

Theorem 6.4 (The strong law of large numbers) *Let $\{X_n\}$ be independent identically distributed random variables with a finite expectation $\mathbb{E}X_n = a$ and a finite variance* $\operatorname{var} X_n = b^2$. *Then*

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} a.$$

The term "identically distributed" means that all random variables X_n have the same distributions. This of course implies that they have the same expectation and the same variance.

The statement of Theorem 6.4 remains true if one drops the assumption of the finiteness of $\operatorname{var} X_n$. Moreover, the finiteness of the mean $\mathbb{E}X_n$ is not only sufficient but also necessary condition for the existence of the limit $\lim \frac{s_n}{n}$ a.s. Another possibility to relax the hypotheses is to drop the assumption that X_n are identically distributed but still require that X_n have a common finite mean and a common finite variance.

The proofs of these stronger results are much longer and will be omitted.

Proof of Theorem 6.4. By Theorem 5.20, it would be sufficient to know that

$$\frac{S_n}{n} \xrightarrow{\text{BC}} a,$$

that is, for any $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - a\right| > \varepsilon\right) < \infty. \tag{6.8}$$

In the proof of Theorem 6.1, we have obtained the estimate (6.6)

$$\mathbb{P}\left(\left|\frac{S_n}{n} - a\right| \geq \varepsilon\right) \leq \frac{b^2}{\varepsilon^2 n}, \tag{6.9}$$

that however is not enough to prove (6.8), because

$$\sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

However, taking in (6.9) n to be a perfect square k^2, we obtain

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{k^2}}{k^2} - a\right| > \varepsilon\right) \leq \sum_{k=1}^{\infty} \frac{\text{const}}{k^2} < \infty,$$

which implies by Theorem 5.20 that

$$\frac{S_{k^2}}{k^2} \xrightarrow{\text{a.s.}} a. \tag{6.10}$$

Now we need to extend this convergence to the whole sequence S_n, that is to "fill gaps" between perfect squares. Assume first that all $X_n \geq 0$. Then the sequence S_n is increasing. For any positive integer n, find k so that

$$k^2 \leq n < (k+1)^2.$$

Then

$$\frac{S_{k^2}}{(k+1)^2} \leq \frac{S_{k^2}}{n} \leq \frac{S_n}{n} \leq \frac{S_{(k+1)^2}}{n} \leq \frac{S_{(k+1)^2}}{k^2},$$

and since $k^2 \sim (k+1)^2$ as $k \to \infty$, we see that by (6.10)

$$\frac{S_{k^2}}{(k+1)^2} \xrightarrow{\text{a.s.}} a \quad \text{and} \quad \frac{S_{(k+1)^2}}{k^2} \xrightarrow{\text{a.s.}} a\,,$$

whence

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} a.$$

Finally, we get rid of the restriction $X_n \geq 0$. For a general X_n, consider its positive and negative parts X_n^+ and X_n^-. More precisely, set

$$X_n^+ = f(X_n) \quad \text{and} \quad X_n^- = g(X_n)$$

where

$$f(x) = \max(x, 0) \quad \text{and} \quad g(x) = \max(-x, 0).$$

Note that $X_n = X_n^+ - X_n^-$ and $|X_n| = X_n^+ + X_n^-$.

We claim that the sequence $\{X_n^+\}$ (and similarly $\{X_n^-\}$) satisfies the hypothesis of the present theorem. Firstly, the sequence $\{X_n^+\}$ is independent by Theorem 5.18 just because $X_n^+ = f(X_n)$. For the same reason, all X_n^+ are identically distributed. Also, X_n^+ is square integrable by $X_n^+ \leq |X_n|$ so that X_n^+ has a finite variance.

By the first part of the proof, we have

$$\frac{X_1^+ + ... + X_n^+}{n} \xrightarrow{\text{a.s.}} \mathbb{E}X_1^+$$
$$\frac{X_1^- + ... + X_n^-}{n} \xrightarrow{\text{a.s.}} \mathbb{E}X_1^-.$$

Subtracting these two identities, we obtain

$$\frac{X_1 + ... + X_n}{n} \xrightarrow{\text{a.s.}} \mathbb{E}X_1 = a,$$

which was to be proved. ■

6.4 Random walks

Let us apply the strong law to a random walk. Let $\{X_n\}$ be independent sequence random variables taking values 1 and -1 with probabilities p and $1-p$, respectively. Define $S_0 = 0$ and

$$S_n = X_1 + X_2 + ... + X_n.$$

Then S_n can be interpreted as a position at time n of a particle performing a random walk on integers. Since $\mathbb{E}X_n = 2p - 1$ and $\operatorname{var} X_n < \infty$, Theorem 6.4 says that

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} 2p - 1. \tag{6.11}$$

If $p \neq \frac{1}{2}$ then this means that S_n behaves approximately as $(2p-1)n$ for large n, which can be interpreted as the particle moving towards the infinity with a constant speed $2p - 1$.

If $p = \frac{1}{2}$ then $\mathbb{E}X_n = 0$ and $\mathbb{E}S_n = 0$. In this case S_n is called *the simple random walk.* For the simple random walk, (6.11) implies

$$S_n = o(n) \quad \text{a.s.}$$

By a more careful analysis, one can say much more about the asymptotic behavior of S_n.

THEOREM. (Khinchin's law of the iterated logarithm) *Let S_n be the simple random walk. Then*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log n}} = 1 \quad \textit{and} \quad \liminf_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log n}} = -1.$$

The proof of this theorem is rather long and is outside the scope of this course. Instead, we will prove the following result, which although weaker than Khinchin's theorem, still introduces the $\sqrt{n}$.

Theorem 6.5 (Hausdorff's theorem) *Let $\{X_n\}$ be independent random variables with a common expectation $\mathbb{E}X_n = 0$ and a common finite k-*th *moment $M_k = \mathbb{E}\,|X_n|^k$ for all $k = 1, 2, 3, \ldots$. Then, for all $\varepsilon > 0$,*

$$S_n = o(n^{\frac{1}{2}+\varepsilon}) \quad \text{a.s.}$$

The hypothesis of finiteness of all moments is trivially satisfied for all random variables taking finitely many values as well as for normal distribution. Hence, Theorem 6.5 applies to such random variables.

We first prove the following lemma.

Lemma 6.6 (The Hölder inequality) *Let $p, q > 1$ be such that*

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{6.12}$$

Then, for all random variables X, Y,

$$\mathbb{E}\,|XY| \leq (\mathbb{E}\,|X|^p)^{1/p} (\mathbb{E}\,|Y|^q)^{1/q} \tag{6.13}$$

(the undefined product $0 \cdot \infty$ is understood as 0).

Remark. Numbers p, q satisfying (6.12) are called the *Hölder conjugate*, and the couple (p, q) is called a Hölder couple. In particular, for $p = q = 2$ we obtain the Cauchy-Schwarz inequality.

Proof. If $\mathbb{E}|X|^p = 0$ then by Theorem 4.12 we have $X = 0$ a.e.so that the left hand side of (6.13) vanishes and, hence, (6.13) is trivially satisfied. In the same way (6.13) is satisfied if $\mathbb{E}|Y|^q = 0$. Let us assume in the sequel that $\mathbb{E}|X|^p$ and $\mathbb{E}|Y|^q$ are positive. If one of these values is equal to ∞ then again (6.13) is trivially satisfied. Hence, we can assume that both $\mathbb{E}|X|^p$ and $\mathbb{E}|Y|^q$ are positive and finite.

Next, observe that inequality (6.13) is scaling invariant: if X is replaced by αX when $\alpha \in \mathbb{R}$, then the validity of (6.13) does not change (indeed, when multiplying X by α, the both sides of (6.13) are multiplied by $|\alpha|$). Hence, by normalizing X and Y we can assume that $\mathbb{E}|X|^p = 1 = \mathbb{E}|Y|^q$. Then it remains to prove that

$$\mathbb{E}|XY| \leq 1.$$

For that we use the Young inequality:

$$ab \leq \frac{a^p}{p} + \frac{b^p}{q}, \tag{6.14}$$

which is true for all non-negative reals a, b and all Hölder couples p, q. Indeed, consider the graph of function $y = x^{p-1}$ over the interval $x \in [0, a]$. The area of the subgraph is

$$A = \int_0^a x^{p-1} dx = \frac{a^p}{p}.$$

Consider the inverse function $x = y^{\frac{1}{p-1}} = y^{q-1}$ over the interval $y \in [0, b]$. The area of its subgraph is

$$B = \int_0^b y^{q-1} dy = \frac{b^q}{q}.$$

It is easy to see that the two subgraphs cover the rectangle $[0, a] \times [0, b]$: if (x, y) is not in the first subgraph, that is, $y > x^{p-1}$ then $x < y^{\frac{1}{p-1}} = y^{q-1}$ so that (x, y) is in the second subgraph. Since the area of the rectangle is ab, it follows that $ab \leq A+B$, whence (6.14) follows.

Applying (6.14) to $|X|, |Y|$ and integrating against $d\mathbb{P}$, we obtain

$$\mathbb{E}|XY| \leq \frac{1}{p}\mathbb{E}|X|^p + \frac{1}{q}\mathbb{E}|Y|^q = \frac{1}{p} + \frac{1}{q} = 1,$$

which finishes the proof. ■

Taking $Y = 1$ in (6.13) we obtain

$$\mathbb{E}|X| \leq (\mathbb{E}|X|^p)^{1/p}$$

for any $p \geq 1$. Replacing X by $|X|^\alpha$ and setting $\beta = \alpha p$, we obtain

$$\mathbb{E}|X|^\alpha \leq \left(\mathbb{E}|X|^\beta\right)^{\alpha/\beta} \tag{6.15}$$

for all $\beta \geq \alpha > 0$.

Proof of Theorem 6.5. We will use the Chebyshev's inequality (6.4) in the form

$$\mathbb{P}(|S_n| > t) \leq \frac{1}{t^{2k}} \mathbb{E} S_n^{2k}, \tag{6.16}$$

where $t > 0$ and an integer k will be chosen later. Let us first estimate $\mathbb{E}\left(S_n^{2k}\right)$. Denote for simplicity $m = 2k$ and observe that

$$S_n^m = \left(\sum_{i_1=1}^n X_{i_1}\right)\left(\sum_{i_2=1}^n X_{i_2}\right)...\left(\sum_{i_m=1}^n X_{i_m}\right) = \sum_{i_1,i_2,...,i_m} X_{i_1}X_{i_2}...X_{i_m}. \tag{6.17}$$

Here $i_1, i_2, ..., i_m$ are indices varying between 1 and n. In particular, the total number of terms in the sum (6.17) is equal to n^m.

Let us call a term $X_{i_1}X_{i_2}...X_{i_m}$ in the sum (6.17) *trivial* if the value of one of the indices $i_1, ..., i_m$ is different from all the others. For example, $X_2X_1X_3X_3X_2X_2$ is trivial whereas $X_1X_2X_1X_3X_2X_3$ is non-trivial. Observe that the expectation of any trivial term is 0. For example, if the value of i_1 is different from $i_2, ..., i_n$ then, using the independence of $\{X_i\}$ we obtain

$$\mathbb{E}(X_{i_1}X_{i_2}...X_{i_m}) = \mathbb{E}(X_{i_1})\,\mathbb{E}(X_{i_2}...X_{i_m}) = 0,$$

because $\mathbb{E}X_{i_1} = 0$.

Let us estimate the expectation of any non-trivial term. Denote by l_j the number of occurrences of the value $j = 1, ..., n$ in the sequence $i_1, i_2, ..., i_m$. Then

$$l_1 + l_2 + ... + l_n = m,$$

and

$$\begin{aligned}
|\mathbb{E}(X_{i_1}X_{i_2}...X_{i_m})| &= \left|\mathbb{E}\left(X_1^{l_1}X_2^{l_2}...X_n^{l_n}\right)\right| \\
&= \left|\mathbb{E}X_1^{l_1}\mathbb{E}X_2^{l_2}...\mathbb{E}X_n^{l_n}\right| \\
&\leq (\mathbb{E}X_1^m)^{l_1/m}(\mathbb{E}X_2^m)^{l_2/m}...(\mathbb{E}X_n^m)^{l_n/m} \\
&= M_m^{\frac{l_1+...+l_m}{m}} \\
&= M_m,
\end{aligned}$$

where $M_m = \mathbb{E}X_i^m$. Here we have used the Hölder inequality (6.15)

$$\left|\mathbb{E}X^l\right| \leq (\mathbb{E}X^m)^{l/m}$$

and the fact that m is even, which simplifies notation.

Now let us estimate the number of non-trivial sequences $i_1, ..., i_m$, that is, the sequences where each value of the index is repeated at least twice. The number of different values of indices $i_1, i_2, ..., i_m$ in a non-trivial sequence is at most $m/2 = k$. Let us mark all positions in the sequence $i_1, ..., i_m$ with the distinct values and then

mark also some other positions so that the total number of the marked positions is exactly k. Here is an example of marking with $k = 4, m = 8$:

$$\underset{}{i_1},\ \underset{}{i_2},\ \underset{\times}{i_3},\ \underset{\times}{i_4},\ \underset{}{i_5},\ \underset{\times}{i_6},\ \underset{\times}{i_7},\ \underset{}{i_8}.$$

In each of k marked position the index can take any value from 1 to n. In each of k unmarked position the value of the index must repeat one of the marked values, hence, giving at most k values. Hence, the number of non-trivial sequences with fixed marking is at most $n^k k^k$. Since the marked k position can chosen by $\binom{m}{k} = \binom{2k}{k}$ ways, the total number of non-trivial sequences $i_1, ..., i_m$ is bounded from above by $\binom{2k}{k} k^k n^k = C_k n^k$ where $C_k = \binom{2k}{k} k^k$. For comparison, let us recall that the total number of all sequences $i_1, ..., i_m$ is n^{2k}, which means that the number of non-trivial terms is a very small fraction of all terms if n is large enough.

Given that the number of non-trivial terms is at most $C_k n^k$ and the expectation of each of then is bounded by M_{2k}, we can estimate $\mathbb{E}\left(S_n^{2k}\right)$ as follows:

$$\mathbb{E}\left(S_n^{2k}\right) \le M_{2k} C_k n^k = C'_k n^k,$$

where $C'_k = M_{2k} C_k$. By Chebyshev's inequality (6.16), we have for any $t > 0$,

$$\mathbb{P}\left(|S_n| > t\right) \le \frac{C'_k n^k}{t^{2k}}.$$

Fix some $\varepsilon, \delta > 0$ and choose $t = \delta n^{\frac{1}{2}+\varepsilon}$. Then we have

$$\mathbb{P}\left(\left|\frac{S_n}{n^{\frac{1}{2}+\varepsilon}}\right| > \delta\right) = \mathbb{P}\left(|S_n| > \delta n^{\frac{1}{2}+\varepsilon}\right) \le \frac{C'_k n^k}{\delta^{2k} n^{k+2\varepsilon k}} = \frac{\text{const}}{n^{2\varepsilon k}}.$$

Choosing $k > \frac{1}{2\varepsilon}$, we see that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{S_n}{n^{1/2+\varepsilon}}\right| > \delta\right) < \infty,$$

which means that

$$\frac{S_n}{n^{1/2+\varepsilon}} \xrightarrow{\text{BC}} 0.$$

By Theorem 5.20, we conclude that

$$\frac{S_n}{n^{1/2+\varepsilon}} \xrightarrow{\text{a.s.}} 0,$$

Lecture 19
22.11.10

which was to be proved. ■

6.5 The tail events and Kolmogorov's $0-1$ law

Definition. For any family $\{X_i\}_{i\in I}$ of random variables define $\sigma\left(\{X_i\}\right)$ as the minimal σ-algebra containing all the events of the form $\{X_i \le c\}$ where $i \in I$ and

$c \in \mathbb{R}$. In other words, $\sigma(\{X_i\})$ is the minimal σ-algebra such that all X_i are measurable with respect to it. One says that $\sigma(\{X_i\})$ is the σ-algebra generated by the family $\{X_i\}$.

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. For any $m \in \mathbb{N}$, set

$$\sigma_m = \sigma(X_m, X_{m+1}, ...).$$

Clearly, σ_1 is σ-algebra generated by the whole sequence $\{X_n\}$, and the sequence $\{\sigma_m\}$ is decreasing in m.

Definition. The *tail* σ-algebra of $\{X_n\}$ is defined by

$$\sigma_\infty = \bigcap_{m=1}^{\infty} \sigma_m = \lim_{m\to\infty} \sigma_m.$$

The elements of σ_∞ are called *tail events*. A function on Ω that is σ_∞-measurable, is called a *tail function*.

Since $\sigma_\infty \subset \mathcal{F}$, all tail events are events, and all tail functions are random variables. Here are some examples of tail events and functions.

Example. We claim that the set

$$\left\{\lim_{n\to\infty} X_n \quad \text{exists}\right\} \tag{6.18}$$

is a tail event. Let us first understand why it is in σ_1. Indeed, the existence of the limit is equivalent to the fact that the sequence $\{X_n\}$ is Cauchy, that is

$$\begin{aligned}\left\{\lim_{n\to\infty} X_n \quad \text{exists}\right\} &= \{\forall \varepsilon > 0 \quad \exists N \quad \forall n, k \geq N \quad |X_n - X_k| < \varepsilon\} \\ &= \bigcap_{\varepsilon>0} \bigcup_{N} \bigcap_{n,k\geq N} \{|X_n - X_k| < \varepsilon\}.\end{aligned}$$

Since X_n and X_k are σ_1-measurable, $|X_n - X_k|$ is also σ_1-measurable, whence it follows that

$$\left\{\lim_{n\to\infty} X_n \quad \text{exists}\right\} \in \sigma_1.$$

The event (6.18) is defined by the tail of the sequence, that is, the existence of the limit of the sequence $X_1, X_2, ...$ and of the *shifted* sequence $X_m, X_{m+1}, ...$ is the same event. Hence, applying the above argument to the shifter sequence, we obtain

$$\left\{\lim_{n\to\infty} X_n \quad \text{exists}\right\} \in \sigma_m.$$

Intersecting over all m, we obtain the claim.

Example. We claim that the function $\limsup_{n\to\infty} X_n$ is a tail function. Indeed, for that it suffices to verify that the set

$$\left\{\limsup_{n\to\infty} X_n \leq x\right\} \tag{6.19}$$

is a tail event for any real x. Using the definition of lim sup, we have

$$\begin{aligned}\left\{\limsup_{n\to\infty} X_n \le x\right\} &= \{\forall \varepsilon > 0 \quad \exists N \quad \forall n \ge N \quad X_n < x + \varepsilon\} \\ &= \bigcap_{\varepsilon>0} \bigcup_{N} \bigcap_{n\ge N} \{X_n < x + \varepsilon\},\end{aligned}$$

which is in σ_1. Since the event (6.19) is defined by the tail of the sequence, we obtain as in the previous example, that (6.19) is a tail event.

Example. Denote $S_n = X_1 + X_2 + ...X_n$, and let $a \in [-\infty, +\infty]$. In general, the event

$$\left\{\lim_{n\to\infty} S_n = a\right\}$$

is not a tail event[1] because it cannot be expressed without using, say, X_1. However, we claim that the event

$$\left\{\lim_{n\to\infty} \frac{S_n}{n} = a\right\}$$

is a tail event. Indeed, one proves as above that $\lim S_n/n$ is σ_1-measurable. To prove that this event is defined by the tail, observe that, for any index m,

$$\lim_{n\to\infty} \frac{S_n}{n} = \lim_{n\to\infty} \left(\frac{S_{m-1}}{n} + \frac{X_m + X_{m+1} + ... + X_n}{n}\right) = \lim_{n\to\infty} \frac{X_m + X_{m+1} + ... + X_n}{n-m+1},$$

because $S_{m-1}/n \to 0$ and $n \sim n - m + 1$ as $n \to \infty$.

In the same way, one can prove that the following are tail events (for $\alpha > 0$):

$$\left\{\lim_{n\to\infty} \frac{S_n}{n^\alpha} \quad \text{exists}\right\} \quad \text{and} \quad \left\{\lim_{n\to\infty} \frac{S_n}{n^\alpha} = a\right\}.$$

In particular, in the strong law of large numbers, we considered the tail event $\{S_n/n \to a\}$, and in the Hausdorff theorem – the tail event $S_n = o(n^{\frac{1}{2}+\varepsilon})$. Basically, most limit theorems have to do with some tail events.

Here are some example of tail functions that arise in this way:

$$\limsup_{n\to\infty} \frac{S_n}{n^\alpha} \quad \text{and} \quad \liminf_{n\to\infty} \frac{S_n}{n^\alpha}.$$

Theorem 6.7 (Kolmogorov's $0-1$ law) *If the sequence $\{X_n\}$ is independent then the tail σ-algebra σ_∞ of $\{X_n\}$ is trivial, that is, each event in σ_∞ has probability* 0 *or* 1.

[1] More precisely, $\{\lim S_n = a\}$ is *not* a tail event for the sequence $\{X_n\}$ but *is* a tail event for the sequence $\{S_n\}$.

Proof. Let us first prove the following claim.

CLAIM. *If $\{X_1, ..., X_m, Y_1, .., Y_n\}$ is a sequence of independent random variables then the following two σ-algebras*

$$\sigma(X_1, ..., X_m) \quad \text{and} \quad \sigma(Y_1, ..., Y_n)$$

are independent.

Denote by $\mathcal{A}$ the family of events of the form

$$A = \{X_1 \leq a_1, X_2 \leq a_2, ..., X_m \leq a_m\}$$

where $a_i \in \mathbb{R}$ and by $\mathcal{B}$ the family of events of the form

$$B = \{Y_1 \leq b_1, Y_2 \leq b_2, ..., Y_n \leq b_n\}$$

where $b_j \in \mathbb{R}$. By hypotheses, every two events $A \in \mathcal{A}$ and $B \in \mathcal{B}$ are independent. Indeed, since the events

$$\{X_1 \leq a_1\}, ..., \{X_m \leq a_m\}, \{Y_1 \leq b_1\}, ..., \{Y_n \leq b_n\}$$

are independent, the events A and B are independent by Lemma 3.5.

Since both families $\mathcal{A}$, $\mathcal{B}$ are closed under finite intersection, by obtain by Theorem 3.6 that $\sigma(\mathcal{A})$ and $\sigma(\mathcal{B})$ are also independent. By definition we have $\sigma(X_1, ..., X_m) = \sigma(\mathcal{A})$ and $\sigma(Y_1, ..., Y_n) = \sigma(\mathcal{B})$ whence the claim follows.

Fix some index m and observe that the sequence of σ-algebras

$$\sigma(X_{m+1}, ..., X_{m+n}), \quad n = 1, 2, ...$$

increases with n. It follows that their union

$$\mathcal{U} = \bigcup_{n=1}^{\infty} \sigma(X_{m+1}, ..., X_{m+n})$$

is an algebra (but not necessarily a σ-algebra). By the above claim, $\sigma(X_1, ..., X_m)$ and $\sigma(X_{m+1}, ..., X_{m+n})$ are independent, whence it follows that $\sigma(X_1, ..., X_m)$ and $\mathcal{U}$ are independent. By Theorem 3.6, also $\sigma(X_1, ..., X_m)$ and $\sigma(\mathcal{U})$ are independent. Since $\sigma(\mathcal{U})$ is the σ-algebra generated by the infinite sequence $\{X_{m+1}, X_{m+2}, ...\}$, that is,

$$\sigma(\mathcal{U}) = \sigma(X_{m+1}, X_{m+2}, ...) = \sigma_{m+1},$$

we see that $\sigma(X_1, ..., X_m)$ and σ_{m+1} are independent.

Since $\sigma_\infty \subset \sigma_{m+1}$, it follows that $\sigma(X_1, ..., X_m)$ and σ_∞ are independent for all m. Then also the union

$$\mathcal{V} = \bigcup_{m=1}^{\infty} \sigma(X_1, ..., X_m) \tag{6.20}$$

and σ_∞ are independent. Since $\sigma(\mathcal{V}) = \sigma_1$, we obtain that σ_1 and σ_∞ are independent. Since $\sigma_\infty \subset \sigma_1$, it follows that σ_∞ and σ_∞ are independent. Hence, every

event $A \in \sigma_\infty$ is independent of itself, whence $\mathbb{P}(A) = 0$ or 1 (cf. Example in Section 3.4). ■

As a consequence, we see that the events like

$$\left\{\lim_{n\to\infty} X_n \quad \text{exists}\right\}, \quad \left\{\lim_{n\to\infty} \frac{S_n}{n} = a\right\}, \quad \left\{S_n = o(n^{\frac{1}{2}+\varepsilon})\right\}, \quad \text{etc.},$$

occur with probability either 0 or 1, provided $\{X_n\}$ are independent.

Theorem 6.7 implies that if T is a tail function of independent random variables $X_1, X_2, ...$ then $T = \text{const}$ a.s. (see Exercises). For example, we see that such random variables as

$$\limsup X_n, \quad \limsup \frac{S_n}{n^\alpha}, \quad \limsup |X_n|^{1/n},$$

are constants a.s.

Chapter 7

Convergence of sequences of random variables

7.1 Measurability of limits a.s.

Recall that by Theorem 4.4 the pointwise limit of a sequence of measurable functions is measurable. The following lemma extends this property to convergence a.s..

Lemma 7.1 *Let $\{X_n\}$ be a sequence of random variables such that the sequence $\{X_n(\omega)\}$ converges for almost all $\omega \in \Omega$. Define*

$$X(\omega) = \begin{cases} \lim_{n\to\infty} X_n(\omega), & \text{if } \{X_n(\omega)\} \text{ converges} \\ 0, & \text{otherwise.} \end{cases}$$

Then X is a random variable.

Proof. Let us first recall that the set

$$A = \{\omega \in \Omega : \{X_n(\omega)\} \text{ converges}\}$$

is measurable, because it is a tail event. Consider the sequence $Y_n = \mathbf{1}_A X_n$ and observe that, firstly, Y_n is measurable as the product of two measurable functions;, and secondly $Y_n \to X$ a.s.as $n \to \infty$ pointwise. Indeed, if $\omega \in A$ then

$$Y_n(\omega) = X_n(\omega) \to X(\omega)$$

and if $\omega \notin A$ then

$$Y_n(\omega) = 0 \to X(\omega).$$

Hence, by Theorem 4.4, X is measurable as the pointwise limit of a sequence of measurable functions. ■

The point of Lemma 7.1 is that if a sequence $\{X_n\}$ converges a.s. then there is a random variable X such that $X_n \xrightarrow{\text{a.s.}} X$.

7.2 Convergence of the expectations

In the next theorem we collect convenient conditions under which the operations $\mathbb{E}$ and lim can be interchanged.

Theorem 7.2 *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables such that $X_n \xrightarrow{\text{a.s.}} X$.*

(*a*) (The monotone convergence theorem) *If $X_n \geq 0$* a.s. *and the sequence $\{X_n\}$ is monotone increasing* a.s. *then*

$$\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n. \tag{7.1}$$

(*b*) (Fatou's lemma) *If $X_n \geq 0$* a.s. *then*

$$\mathbb{E}X \leq \liminf_{n\to\infty} \mathbb{E}X_n. \tag{7.2}$$

(*c*) (The dominated convergence theorem) *If here exists an integrable random variable Y such that $|X_n| \leq Y$* a.s. *for all n, then* (7.1) *is satisfied.*

Remark. In part (*b*) the identity (7.1) may fail. Indeed, consider on $[0,1]$ the following sequence of functions

$$X_n = n\mathbf{1}_{(0,\frac{1}{n})}.$$

Then $X_n \to 0$ pointwise on $[0,1]$ while $\mathbb{E}X_n = 1 \not\to 0$.

Remark. The condition $|X_n| \leq Y$ in part (*c*) is called the *domination condition*. Theorem 7.2 provides two sufficient conditions for the validity of (7.1): either the sequence $\{X_n\}$ must be non-negative and monotone increasing, or it should satisfy the domination condition.

The bounded convergence theorem of Corollary 5.24 is clearly a consequence of the dominated convergence theorem with $Y = \text{const}$. However, we use the bounded convergence theorem to prove Theorem 7.2.

Proof. (*a*) For any positive integer m, define $X^{(m)} = \min(X, m)$ and $X_n^{(m)} = \min(X_n, m)$. Since

$$0 \leq X^{(m)} - X_n^{(m)} \leq X - X_n,$$

we obtain that $X_n^{(m)} \xrightarrow{\text{a.s.}} X^{(m)}$ as $n \to \infty$ for any m. Since the sequence $\left\{X_n^{(m)}\right\}_{n=1}^{\infty}$ is uniformly bounded by m, we obtain by the bounded convergence theorem (Corollary 5.24) that

$$\mathbb{E}X^{(m)} = \lim_{n\to\infty} \mathbb{E}X_n^{(m)} \leq \lim_{n\to\infty} \mathbb{E}X_n.$$

Letting $m \to \infty$ and using

$$\lim_{m\to\infty} \mathbb{E}X^{(m)} = \mathbb{E}X$$

(see Theorem 4.9), we obtain

$$\mathbb{E}X \leq \lim_{n\to\infty} \mathbb{E}X_n.$$

Since the opposite inequality is trivially satisfied by $X \geq X_n$, we obtain the equality (7.1).

Lecture 20
23.11.10

(b) Consider the sequence

$$Y_n = \inf_{k\geq n} X_k = \lim_{k\to\infty} \min\{X_n, X_{n+1}, ..., X_k\}.$$

Then $\{Y_n\}$ is a monotone increasing sequence of random variables, $Y_n \geq 0$ a.s. and $Y_n \xrightarrow{\text{a.s.}} X$. It follows from part ($a$) that

$$\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}Y_n.$$

On the other hand, $Y_n \leq X_n$ implies

$$\lim_{n\to\infty} \mathbb{E}Y_n \leq \liminf_{n\to\infty} \mathbb{E}X_n,$$

whence (7.2) follows.

(c) Since $X_n + Y \geq 0$ a.s. and $X_n + Y \xrightarrow{\text{a.s.}} X + Y$, we obtain by Fatou's lemma

$$\mathbb{E}(X+Y) \leq \liminf_{n\to\infty} \mathbb{E}(X_n + Y) = \liminf_{n\to\infty} \mathbb{E}X_n + \mathbb{E}Y.$$

Since Y is integrable, cancelling out $\mathbb{E}Y$ we obtain

$$\mathbb{E}X \leq \liminf_{n\to\infty} \mathbb{E}X_n. \tag{7.3}$$

Applying the same argument to the sequence $\{-X_n\}$, we obtain

$$\mathbb{E}(-X) \leq \liminf_{n\to\infty} \mathbb{E}(-X_n) = -\limsup_{n\to\infty} \mathbb{E}X_n$$

whence

$$\mathbb{E}X \geq \limsup_{n\to\infty} \mathbb{E}X_n.$$

Combining with (7.3) we obtain (7.1).

Alternative proof of (c)**.** Replacing Y by $Y_+ + c$ where $c > 0$, we can assume that Y is strictly positive. By dividing all random variables X_n, X, Y by the constant $\mathbb{E}Y$, we can assume that $\mathbb{E}Y = 1$. Then Y determines a new probability measure $\mathbb{P}'$ on $\mathcal{F}$ given by $d\mathbb{P}' = Y d\mathbb{P}$, that is, for any event A,

$$\mathbb{P}'(A) = \int_A Y d\mathbb{P} = \mathbb{E}(\mathbf{1}_A Y).$$

In particular, if $\mathbb{P}(A) = 0$ then also $\mathbb{P}'(A) = 0$. For the associated expectation $\mathbb{E}'$ we have

$$\mathbb{E}'Z = \int_\Omega Z d\mathbb{P}' = \int_\Omega ZY d\mathbb{P} = \mathbb{E}(ZY).$$

Consider random variables $X' = X/Y$ and $X'_n = X_n/Y$. By hypothesis, the sequence $\{X'_n\}$ is uniformly bounded a.s.. Since $X'_n \xrightarrow{\text{a.s.}} X'$, applying the bounded convergence theorem in the space $(\Omega, \mathcal{F}, \mathbb{P}')$, we obtain

$$\mathbb{E}'X' = \lim_{n\to\infty} \mathbb{E}'X'_n,$$

that is

$$\mathbb{E}\,(X'Y) = \lim_{n\to\infty} \mathbb{E}\,(X'_nY)\,,$$

which is equivalent to (7.1). ■

Although we have stated and proved Theorem 7.2 for probability measures, similar statement (with the same proof) is true for an arbitrary measure μ defined on a σ-algebra $\mathcal{F}$ of subsets of a set Ω. Let us restate Theorem 7.2 in this case using different notation.

THEOREM. *Let $\{f_n\}$ be a sequence of measurable functions on Ω and f be a measurable function such that $f_n \xrightarrow{\text{a.s.}} f$.*

(*a*) (The monotone convergence theorem) *If $f_n \geq 0$* a.s. *and the sequence $\{f_n\}$ is monotone increasing* a.s. *then*

$$\int_\Omega f d\mu = \lim_{n\to\infty} \int_\Omega f_n d\mu. \tag{7.4}$$

(*b*) (Fatou's lemma) *If $f_n \geq 0$* a.s. *then*

$$\int_\Omega f d\mu \leq \liminf_{n\to\infty} \int_\Omega f_n d\mu.$$

(*c*) (The dominated convergence theorem) *If there exists an integrable function g such that $|f_n| \leq g$* a.s. *for all n, then* (7.4) *is satisfied.*

For example, the convergence theorems hold for the Lebesgue measure λ_n on bounded boxes in $\mathbb{R}^n$. By taking exhaustion of $\mathbb{R}^n$ by bounded boxes, one can show that the convergence theorems remain true for the Lebesgue measure λ_n on entire $\mathbb{R}^n$.

7.3 Weak convergence of measures

Definition. Given a sequence $\{\mu_n\}$ of probability measures on $\mathcal{B}\,(\mathbb{R}^m)$ and a probability measure μ on $\mathcal{B}(\mathbb{R}^m)$, we say that μ_n converges *weakly* to μ and write

$$\mu_n \Rightarrow \mu$$

if, for any bounded continuous function f on $\mathbb{R}^m$,

$$\int_{\mathbb{R}^m} f d\mu_n \to \int_{\mathbb{R}^m} f d\mu.$$

It would be natural to say that μ_n converges to μ and write $\mu_n \to \mu$ if

$$\mu_n(A) \to \mu(A), \tag{7.5}$$

for all Borel sets $A \in \mathcal{B}(\mathbb{R}^m)$. However, this would be a stronger condition than we need. Consider the following example.

Example. For any point $x \in \mathbb{R}^m$ define the *Dirac measure* δ_x on $\mathcal{B}(\mathbb{R}^m)$ by

$$\delta_x(A) = \begin{cases} 1, x \in A, \\ 0, x \notin A. \end{cases}.$$

Let $\{x_n\}$ be a sequence of points in $\mathbb{R}^m$ that converges to a point x and such that $x_n \neq x$ for all n. We claim that $\delta_{x_n} \Rightarrow \delta_x$ but in general $\delta_{x_n} \not\to \delta_x$. Indeed, taking $A = \{x\}$ we see that $\delta_{x_n}(A) = 0$ while $\delta_x(A) = 1$ so that $\delta_{x_n}(A) \not\to \delta_x(A)$, which proves the second claim. On the other hand, for any bounded continuous function f on $\mathbb{R}^m$,

$$\int_{\mathbb{R}^m} f d\delta_{x_n} = f(x_n) \to f(x) = \int_{\mathbb{R}^m} f d\delta_x,$$

which proves the first claim.

Next we give another characterization of the weak convergence of measures in one-dimensional case. Let μ be a probability measure on $\mathcal{B}(\mathbb{R})$. Recall that the distribution function F_μ of measure μ is defined by

$$F_\mu(x) = \mu(-\infty, x].$$

Recall also that, by Theorem 2.10, function F_μ is monotone increasing and right continuous.

Theorem 7.3 *Let $\{\mu_n\}$ and μ be probability measures on $\mathcal{B}(\mathbb{R})$. Then the following conditions are equivalent:*

(*i*) $\mu_n \Rightarrow \mu$;

(*ii*) $F_{\mu_n}(x) \to F_\mu(x)$ *for any* $x \in \mathbb{R}$ *where* F_μ *is continuous.*

In other words, $\mu_n \Rightarrow \mu$ is equivalent to the pointwise convergence of the distribution functions at all the points of continuity of F_μ.

Remark. Observe that the set of points of discontinuity of F_μ is at most countable, because F_μ is monotone. Indeed, for any point x at which F_μ is discontinuous, there corresponds a non-empty open interval $(F_\mu(x-), F_\mu(x+))$. All such intervals are disjoint, which implies that the set of them is at most countable.

Note also that a monotone right continuous function F is uniquely determined by its values outside a countable set S (in particular, F is uniquely determined by its values at all points of continuity). Indeed, for any real x, there is a sequence $\{x_n\}$ which converges to x from above and such that all x_n are outside S. Therefore, the values $F(x_n)$ are given, whence $F(x)$ can be determined by $F(x) = \lim_{n\to\infty} F(x_n)$.

Example. For a Dirac measure δ_x we have $F_{\delta_x} = \mathbf{1}_{[x,+\infty)}$ so that the only point of discontinuity of F_{δ_x} is x. Let $\{x_n\}$ be a strictly monotone decreasing sequence of reals that converges to x. As we have seen above, $\delta_{x_n} \Rightarrow \delta_x$. On the other hand, we have

$$\lim_{n\to\infty} F_{\delta_{x_n}} = \lim_{n\to\infty} \mathbf{1}_{[x_n,+\infty)} = \mathbf{1}_{(x,+\infty)},$$

so that $F_{\delta_{x_n}} \to F_{\delta_x}$ in $\mathbb{R} \setminus \{x\}$, while at the point x

$$\lim_{n\to\infty} F_{\delta_{x_n}}(x) \neq F_{\delta_x}(x).$$

Proof of Theorem 7.3. Denote for simplicity $F_n = F_{\mu_n}$ and $F = F_\mu$. Denote by C the set of points of continuity of F.

$(i) \Longrightarrow (ii)$. Fix a point $x \in C$ and prove that $F_n(x) \to F(x)$. Given any $\varepsilon > 0$, find a continuous function f_ε such that

$$\mathbf{1}_{(-\infty,x]} \leq f_\varepsilon \leq \mathbf{1}_{(-\infty,x+\varepsilon]}. \tag{7.6}$$

For example, we can define f_ε by

$$f_\varepsilon(t) = \begin{cases} 1, & t \leq x, \\ 0, & t > x + \varepsilon, \end{cases}$$

and $f_\varepsilon(t)$ is linear in $[x, x+\varepsilon]$. It follows from (7.6) that

$$F_n(x) = \mu_n((-\infty, x]) = \int_{\mathbb{R}} \mathbf{1}_{(-\infty,x]} d\mu_n \leq \int_{\mathbb{R}} f_\varepsilon d\mu_n.$$

and

$$F(x+\varepsilon) = \mu((-\infty, x+\varepsilon]) = \int_{\mathbb{R}} \mathbf{1}_{(-\infty,x+\varepsilon]} d\mu \geq \int_{\mathbb{R}} f_\varepsilon d\mu.$$

By hypothesis, we have

$$\int_{\mathbb{R}} f_\varepsilon d\mu_n \to \int_{\mathbb{R}} f_\varepsilon d\mu \quad \text{as } n \to \infty,$$

whence it follows that

$$\limsup_{n\to\infty} F_n(x) \leq \int_{\mathbb{R}} f_\varepsilon d\mu \leq \int_{\mathbb{R}} \mathbf{1}_{(-\infty,x+\varepsilon]} d\mu = F(x+\varepsilon). \tag{7.7}$$

Similarly, considering a continuous function g_ε such that

$$\mathbf{1}_{(-\infty,x-\varepsilon]} \leq g_\varepsilon \leq \mathbf{1}_{(-\infty,x]},$$

we obtain

$$\liminf_{n\to\infty} F_n(x) \geq \lim_{n\to\infty} \int_{\mathbb{R}} g_\varepsilon d\mu_n = \int_{\mathbb{R}} g_\varepsilon d\mu \geq \int_{\mathbb{R}} \mathbf{1}_{(-\infty,x-\varepsilon]} d\mu = F(x-\varepsilon). \tag{7.8}$$

Since F is continuous at x and $\varepsilon > 0$ is arbitrary, we obtain from (7.7) and (7.8) that

$$\lim_{n\to\infty} F_n(x) = F(x).$$

$(ii) \Longrightarrow (i)$ Let $F_n(x) \to F(x)$ for all $x \in C$. We need to prove that, for any bounded continuous function f,

$$\int_{\mathbb{R}} f d\mu_n \to \int_{\mathbb{R}} f d\mu. \tag{7.9}$$

We first prove (7.9) for some simpler functions (not necessarily continuous).

Let $f = \mathbf{1}_{(-\infty,x]}$ where $x \in C$. Then

$$\int_{\mathbb{R}} f d\mu_n = F_n(x) \to F(x) = \int_{\mathbb{R}} f d\mu,$$

so that (7.9) holds for such functions. By additivity, (7.9) extends to any function of the form

$$f = \mathbf{1}_{(y,x]} = \mathbf{1}_{(-\infty,x]} - \mathbf{1}_{(-\infty,y]}$$

where $x, y \in C$ and $y < x$. Moreover, if $(y_i, x_i]$ is a finite sequence of intervals where all x_i, $y_i \in C$ then (7.9) holds for any function of the form

$$f = \sum_i c_i \mathbf{1}_{(y_i x_i]}. \tag{7.10}$$

Let now f be a bounded continuous function on $\mathbb{R}$. Choose a bounded closed interval $[a, b]$ and consider a partition $p = \{x_i\}_{i=0}^N$ of $[a, b]$, that is, a sequence such that

$$x_0 = a < x_1 < x_2 < ... < x_N = b.$$

Associated with this partition is a simple function

$$f_p = \sum_{i=1}^{N} f(x_i) \mathbf{1}_{(x_{i-1},x_i]}.$$

Consider now a sequence of partitions p as $m(p) \to 0$ where $m(p) := \max_i |x_i - x_{i-1}|$ is the mesh of p. Since f is uniformly continuous on $[a, b]$, we have $f_p \rightrightarrows f$ on $[a, b]$ that is,

$$\sup_{[a,b]} |f - f_p| \to 0 \text{ as } m(p) \to 0.$$

Now assume in addition that the points $a, b \in C$ and that all the points x_i in all partitions are also chosen from C (which is possible because C is dense in $\mathbb{R}$). Then any function f_p has the form (7.10) whence we conclude that

$$\int_{\mathbb{R}} f_p d\mu_n \to \int_{\mathbb{R}} f_p d\mu \text{ as } n \to \infty. \tag{7.11}$$

Splitting the domain $\mathbb{R}$ of integration in (7.9) into the intervals $(-\infty, a]$, $(a, b]$, $(b, +\infty]$, we obtain

$$\begin{aligned}\left|\int_{\mathbb{R}} f d\mu_n - \int_{\mathbb{R}} f d\mu\right| &\leq \left|\int_{(a,b]} f d\mu_n - \int_{(a,b]} f d\mu\right| && (7.12)\\ &+ \left|\int_{(-\infty,a]} f d\mu_n - \int_{(-\infty,a]} f d\mu\right| && (7.13)\\ &+ \left|\int_{(b,+\infty)} f d\mu_n - \int_{(b,+\infty)} f d\mu\right| && (7.14)\end{aligned}$$

Let us estimate separately all the terms on the right hand side of (7.12)-(7.14). For the term (7.12) we have

$$\begin{aligned}\left|\int_{(a,b]} f d\mu_n - \int_{(a,b]} f d\mu\right| &\leq \left|\int_{(a,b]} (f - f_p)\, d\mu_n\right| \\ &+ \left|\int_{(a,b]} f_p d\mu_n - \int_{(a,b]} f_p d\mu\right| \\ &+ \left|\int_{(a,b]} (f_p - f)\, d\mu\right| \\ &\leq 2 \sup_{[a,b]} |f - f_p| + \left|\int_{\mathbb{R}} f_p d\mu_n - \int_{\mathbb{R}} f_p d\mu\right|,\end{aligned}$$

where we have used that the total mass of μ_n and μ is 1. It follows from (7.11) that

$$\limsup_{n\to\infty} \left|\int_{(a,b]} f d\mu_n - \int_{(a,b]} f d\mu\right| \leq 2 \sup_{[a,b]} |f - f_p|.$$

Since the right hand side here $\to 0$ as $m(p) \to 0$, we obtain that

$$\lim_{n\to\infty} \left|\int_{(a,b]} f d\mu_n - \int_{(a,b]} f d\mu\right| = 0. \tag{7.15}$$

Lecture 21
29.11.10

For the term (7.13) we have

$$\begin{aligned}\left|\int_{(-\infty,a]} f d\mu_n - \int_{(-\infty,a]} f d\mu\right| &\leq \int_{(-\infty,a]} |f|\, d\mu_n + \int_{(-\infty,a]} |f|\, d\mu \\ &\leq \sup |f| \left(F_n(a) + F(a)\right)\end{aligned}$$

and for the term (7.14)

$$\begin{aligned}\left|\int_{(b,+\infty]} f d\mu_n - \int_{(b,+\infty]} f d\mu\right| &\leq \int_{(b,+\infty]} |f|\, d\mu_n + \int_{(b,+\infty]} |f|\, d\mu \\ &\leq \sup |f| \left(1 - F_n(b) + 1 - F(b)\right).\end{aligned}$$

Fix some $\varepsilon > 0$. Since $F(a) \to 0$ as $a \to -\infty$, we can choose $a \in C$ so that

$$F(a) < \varepsilon.$$

Since $F(b) \to 1$ as $b \to +\infty$, we can choose $b \in C$ so that

$$1 - F(b) < \varepsilon.$$

Since $F_n(a) \to F(a)$ and $F_n(b) \to F(b)$ as $n \to \infty$, we have for large enough n

$$F_n(a) < \varepsilon \text{ and } 1 - F_n(b) < \varepsilon.$$

Hence, for large enough n, the terms (7.13) and (7.14) are bounded by 2ε each. Combining with (7.15) we obtain from (7.12)-(7.14) that

$$\limsup_{n\to\infty} \left| \int_{\mathbb{R}} f d\mu_n - \int_{\mathbb{R}} f d\mu \right| \leq 4\varepsilon \sup |f| .$$

Since $\varepsilon > 0$ is arbitrary, we obtain (7.9). ■

7.4 Convergence in distribution

Definition. We say that a sequence of random variables X_n *converges in distribution* (or in law) to a random variable X and write

$$X_n \xrightarrow{\mathrm{D}} X$$

if

$$P_{X_n} \Rightarrow P_X.$$

Recall that P_X is a probability measure on $\mathcal{B}(\mathbb{R})$ defined by $P_X(A) = \mathbb{P}(X \in A)$. So, the convergence in distribution can be applied even if the random variables $\{X_n\}$ and X are all defined on different probability spaces. Sometimes we will use also notation

$$X_n \xrightarrow{\mathrm{D}} \mu ,$$

which is short for

$$P_{X_n} \Rightarrow \mu,$$

where μ is a probability measure on $\mathcal{B}(\mathbb{R})$.

Theorem 7.4 *The following are equivalent:*

(*i*) $X_n \xrightarrow{\mathrm{D}} X$;

(*ii*) $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$, *for any bounded continuous function* f *on* $\mathbb{R}$.

(*iii*) $F_{X_n}(x) \to F_X(x)$, *for any point* x *where* F_X *is continuous.*

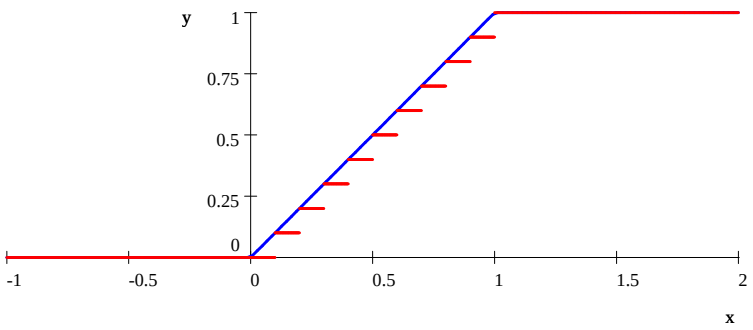


Figure 7.1: Functions $F_n(x)$ and $F(x)$

Proof. Assuming that f is any bounded continuous function on $\mathbb{R}$, we have, by the definition of weak convergence and Theorem 5.7

$$\begin{aligned} X_n \xrightarrow{\mathrm{D}} X \quad &\Longleftrightarrow \quad P_{X_n} \Rightarrow P_X \\ &\Longleftrightarrow \quad \int_{\mathbb{R}} f dP_{X_n} \to \int_{\mathbb{R}} f dP_X \\ &\Longleftrightarrow \quad \mathbb{E}f(X_n) \to \mathbb{E}f(X), \end{aligned}$$

which means that $(i) \Leftrightarrow (ii)$. On the other hand by Theorem 7.3 $P_{X_n} \Rightarrow P_X$ is equivalent to (iii). ■

Example. Let X_n be a random variable taking the values $\left\{\frac{1}{n}, \frac{2}{n}, ..., \frac{n}{n}\right\}$ each with probability $\frac{1}{n}$. Let us show that $X_n \xrightarrow{\mathrm{D}} X$ where X is uniformly distributed on $[0,1]$. The distribution functions F_{X_n} and F_X are equal to 0 for $x \leq 0$ and to 1 for $x \geq 1$. For $x \in (0,1)$, we have

$$F_{X_n}(x) = \mathbb{P}(X_n \leq x) = \frac{1}{n} \# \left\{k \in \mathbb{N} : \frac{k}{n} \leq x\right\} = \frac{[nx]}{n} \underset{n\to\infty}{\longrightarrow} x = F_X(x)$$

(see Fig. 7.1).

Hence, $F_{X_n}(x) \to F_X(x)$ for all $x \in \mathbb{R}$, whence the claim follows by Theorem 7.4(iii). Alternatively, $X_n \xrightarrow{\mathrm{D}} X$ follows from Theorem 7.4(ii) because for any bounded continuous function f

$$\mathbb{E}f(X_n) = \frac{1}{n} \sum_{k=1}^{n} f\left(\frac{k}{n}\right) \underset{n\to\infty}{\longrightarrow} \int_0^1 f(x)\, dx = \mathbb{E}f(X),$$

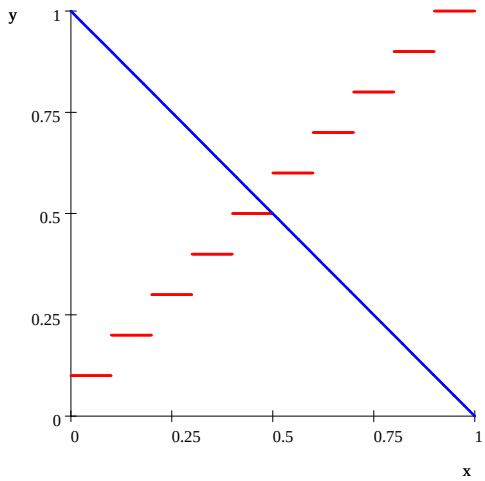
which is true by the Riemann integrability of f.

This example can be used to show that in general $X_n \xrightarrow{\mathrm{D}} X$ does not imply that $X_n \xrightarrow{\mathrm{P}} X$. Indeed, let us specify X_n on $[0,1]$ as follows:

$$X_n(\omega) = \frac{k}{n} \text{ if } \omega \in (\frac{k-1}{n}, \frac{k}{n}]$$

where $k = 1, ..., n$. Clearly, X_n takes each the values $\frac{1}{n}, \frac{2}{n}, ..., \frac{n}{n}$ with probability $\frac{1}{n}$. It is obvious that, for all $\omega \in [0,1]$,

$$\lim_{n\to\infty} X_n(\omega) = \omega.$$



Random variables X_n and X

Consider the random variable $X(\omega) = 1 - \omega$ that is uniformly distributed in $[0,1]$ because for any $c \in [0,1]$

$$\mathbb{P}(X \le c) = \mathbb{P}(1 - \omega \le c) = \mathbb{P}(\omega \ge 1 - c) = 1 - \mathbb{P}(\omega < 1 - c) = 1 - (1 - c) = c.$$

Then X_n cannot converge in probability to X because no subsequence of $\{X_n\}$ converges to X a.s. (cf. Theorem 5.23).

The convergence in distribution is the weakest mode of convergence of random variables, as is stated by the following theorem.

Theorem 7.5 *Let $\{X_n\}$ and X be random variables.*

(a) *If $X_n \xrightarrow{\mathrm{P}} X$ then $X_n \xrightarrow{\mathrm{D}} X$.*

(b) *If c is a constant then $X_n \xrightarrow{\mathrm{D}} c$ implies $X_n \xrightarrow{\mathrm{P}} c$.*

Proof. (a) Denote for simplicity $F_n = F_{X_n}$ and $F = F_X$. By Theorem 7.4, it suffices to prove that

$$F_n(x) \to F(x)$$

for any point x of continuity of F. We use the following general inequality: if A and B are two events then

$$\mathbb{P}(A) - \mathbb{P}(B) \leq \mathbb{P}(A \cap B^c), \tag{7.16}$$

which follows from

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \leq \mathbb{P}(B) + \mathbb{P}(A \cap B^c).$$

Then we have, for any $\varepsilon > 0$,

$$\begin{aligned} F_n(x) - F(x+\varepsilon) &= \mathbb{P}(X_n \leq x) - \mathbb{P}(X \leq x+\varepsilon) \\ &\leq \mathbb{P}(X_n \leq x \text{ and } X > x+\varepsilon) \\ &\leq \mathbb{P}(|X_n - X| > \varepsilon). \end{aligned}$$

As $n \to \infty$, the right hand side $\to 0$ by hypothesis, whence

$$\limsup_{n\to\infty} F_n(x) \leq F(x+\varepsilon). \tag{7.17}$$

Similarly,

$$\begin{aligned} F(x-\varepsilon) - F_n(x) &= \mathbb{P}(X \leq x-\varepsilon) - \mathbb{P}(X_n \leq x) \\ &\leq \mathbb{P}(X \leq x-\varepsilon \text{ and } X_n > x) \\ &\leq \mathbb{P}(|X_n - X| > \varepsilon), \end{aligned}$$

which implies

$$\liminf_{n\to\infty} F_n(x) \geq F(x-\varepsilon). \tag{7.18}$$

Since F is continuous at x, (7.17) and (7.18) yield

$$\lim_{n\to\infty} F_n(x) = F(x).$$

(b) The random variable $X \equiv c$ has the distribution function

$$F = \mathbf{1}_{[c,+\infty)},$$

which is continuous at any $x \neq c$. Hence, by hypothesis, $F_n(x) \to F(x)$ for any $x \neq c$.

For any $\varepsilon > 0$, we have

$$\mathbb{P}(|X_n - c| > \varepsilon) = \mathbb{P}(X_n < c-\varepsilon) + \mathbb{P}(X_n > c+\varepsilon) \leq F_n(c-\varepsilon) + (1 - F_n(c+\varepsilon)).$$

As $n \to \infty$, the right hand side converges to

$$F(c-\varepsilon) + (1 - F(c+\varepsilon)) = 0.$$

We conclude that

$$\mathbb{P}(|X_n - c| > \varepsilon) \underset{n\to\infty}{\longrightarrow} 0,$$

which was to be proved. ■

7.5 A limit distribution of the maximum

We give here a simple example of a statement involving a limit distribution of a functional of random variables.

Theorem 7.6 *Let $\{X_n\}$ be independent identically distributed random variables with a common distribution function F such that*

$$1 - F(x) \sim cx^{-\alpha} \quad \text{as } x \to +\infty, \tag{7.19}$$

where $\alpha > 0$ and $c > 0$. Denote

$$M_n = \max(X_1, X_2, ..., X_n).$$

Then

$$\frac{M_n}{n^{1/\alpha}} \xrightarrow{\mathrm{D}} \mu, \tag{7.20}$$

where μ is the probability measure with the following distribution function

$$F_\mu(x) = \begin{cases} \exp(-cx^{-\alpha}), & x > 0, \\ 0, & x \le 0. \end{cases}$$

Remark. The function F_μ is a distribution function by Theorem 2.10. The density function of μ is

$$f_\mu(x) = F'_\mu(x) = c\alpha x^{-\alpha-1} \exp\left(-cx^{-\alpha}\right), \ x > 0$$

(see Fig. 7.2). Note that if $x \to +\infty$ then

$$F_\mu(x) = 1 - cx^{-\alpha} + o(x^{-\alpha}).$$

Hence, F_μ satisfies (7.19) as well.

Remark. As it has been already explained, (7.20) means that

$$P_{M_n/n^{1/\alpha}} \xrightarrow{\mathrm{D}} \mu.$$

It follows from (7.20) that, for any $\gamma > 1/\alpha$,

$$\frac{M_n}{n^\gamma} \xrightarrow{\mathrm{D}} 0,$$

whence by Theorem 7.5 also

$$\frac{M_n}{n^\gamma} \xrightarrow{\mathrm{P}} 0.$$

Example. Let X_n have the Cauchy distribution, that is

$$F(x) = \frac{1}{\pi} \int_{-\infty}^{x} \frac{dt}{1+t^2}.$$

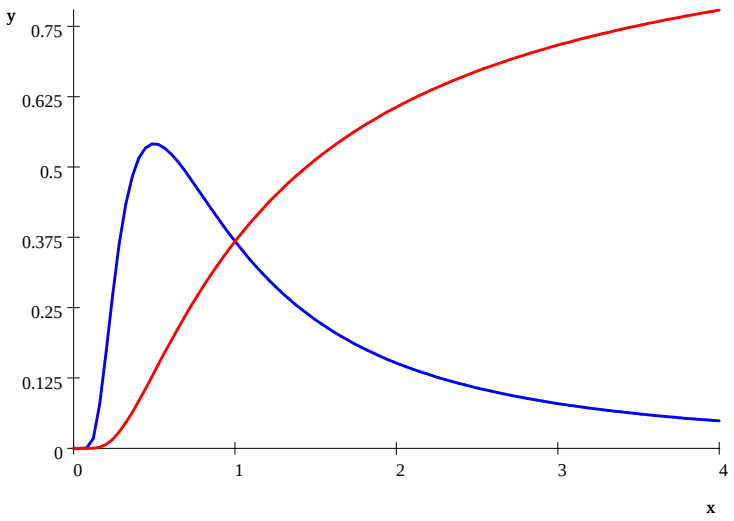


Figure 7.2: Functions F_μ and f_μ for $c = \alpha = 1$.

If $x \to +\infty$ then

$$1 - F(x) = \frac{1}{\pi} \int_x^\infty \frac{dt}{1+t^2} \sim \frac{1}{\pi} \int_x^\infty \frac{dt}{t^2} = \frac{1}{\pi x}.$$

Hence, (7.19) holds with $\alpha = 1$ and $c = \pi^{-1}$. We conclude that $M_n/n \xrightarrow{\mathrm{D}} \mu$ where

$$F_\mu = \begin{cases} \exp\left(-\frac{1}{\pi x}\right), & x > 0, \\ 0, & x \le 0. \end{cases}$$

Example. If $X_n \sim \mathcal{N}(0,1)$ then

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt.$$

If $x \to +\infty$ then

$$1 - F(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \exp\left(-\frac{t^2}{2}\right) dt < \int_x^{+\infty} \exp(-t)\, dt = e^{-x}.$$

Since $e^{-x} = o(x^{-\alpha})$ as $x \to +\infty$ for any $\alpha > 0$, Theorem 7.6 does not apply in this case.

Lecture 22
30.11.10

Proof of Theorem 7.6. Let us first evaluate the distribution function of M_n:

$$F_{M_n}(x) = \mathbb{P}(M_n \le x) = \mathbb{P}(X_1 \le x,\, X_2 \le x,\, ...,\, X_n \le x) = F(x)^n,$$

where we have used the independence of $\{X_k\}$. Therefore, the distribution function of

$$Y_n := \frac{M_n}{n^{1/\alpha}}$$

is

$$F_{Y_n}(x) = \mathbb{P}(Y_n \leq x) = \mathbb{P}\left(M_n \leq n^{1/\alpha}x\right) = F\left(n^{1/\alpha}x\right)^n.$$

If $x \leq 0$ then this is $\leq F(0)^n \to 0$ as $n \to \infty$ (note that $F(0) < 1$, which follows from (7.19)).

To treat the case $x > 0$, first rewriting (7.19) in the form

$$F(y) = 1 - cy^{-\alpha} + o\left(y^{-\alpha}\right)$$

that is

$$\ln F(y) = -cy^{-\alpha} + o\left(y^{-\alpha}\right)$$

as $y \to +\infty$. Substituting $y = n^{1/\alpha}x$, we obtain as $n \to \infty$

$$\begin{aligned} \ln F_{Y_n}(x) &= n \ln F\left(n^{1/\alpha}x\right) \\ &= n\left(-c\frac{x^{-\alpha}}{n} + o\left(\frac{1}{n}\right)\right) \\ &= -cx^{-\alpha} + o(1) \end{aligned}$$

whence

$$F_{Y_n}(x) \underset{n\to\infty}{\longrightarrow} \exp\left(-cx^{-\alpha}\right).$$

Hence, $F_{Y_n}(x) \longrightarrow F_\mu(x)$ for all $x \in \mathbb{R}$. By Theorem 7.4 we conclude that $Y_n \xrightarrow{\mathrm{D}} \mu$. ■

Example. Consider the case when all $X_n \sim \mathcal{N}(0,1)$. Then as $x \to +\infty$

$$1 - F(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} e^{-t^2/2} dt \sim \frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Set $C_n = n\sqrt{\ln n}$ and

$$Y_n = \frac{1}{C_n} \exp\left(M_n^2\right).$$

Then $F_{Y_n}(x) = 0$ for $x \leq 0$ whereas for $x > 0$

$$F_{Y_n}(x) = \mathbb{P}(Y_n \leq x) = \mathbb{P}\left(M_n \leq \sqrt{\ln(C_n x)}\right) = F\left(\sqrt{\ln(C_n x)}\right)^n.$$

As $n \to \infty$, we obtain

$$\begin{aligned} F_{Y_n}(x) &= \left(1 - \frac{\sqrt{\ln(C_n x)}}{\sqrt{2\pi}} e^{-\left(\sqrt{\ln(C_n x)}\right)^2/2} (1 + o(1))\right)^n \\ &= \left(1 - \frac{\sqrt{\ln(C_n x)}}{\sqrt{2\pi} C_n x} (1 + o(1))\right)^n. \end{aligned}$$

Since

$$n\frac{\sqrt{\ln(C_n x)}}{C_n x} \sim \frac{n\sqrt{\ln C_n}}{C_n}\frac{1}{x} \sim \frac{n\sqrt{\ln n}}{n\sqrt{\ln n}}\frac{1}{x} = \frac{1}{x},$$

it follows that

$$F_{Y_n} \to \exp\left(-\frac{1}{\sqrt{2\pi}x}\right).$$

Hence,

$$\frac{1}{n\sqrt{\ln n}}\exp\left(M_n^2\right) \xrightarrow{\mathrm{D}} \mu$$

where

$$F_\mu(x) = \begin{cases} \exp\left(-\frac{1}{\sqrt{2\pi x}}\right), x > 0, \\ 0, \qquad\qquad\qquad x \leq 0. \end{cases}$$

Chapter 8

Characteristic function and central limit theorem

8.1 Complex-valued random variables

In this Chapter we allow complex-valued random variables. Identifying $\mathbb{C}$ with $\mathbb{R}^2$, we can say that a complex-valued random variable is the same as a random vector with values in $\mathbb{R}^2$. Any complex-valued random variable Z can be then represented as a couple (X, Y) where X, Y are real-valued random variables, which is equivalent to $Z = X + iY$. Then Z is called integrable if both X, Y are integrable, and

$$\mathbb{E}Z := \mathbb{E}X + i\mathbb{E}Y.$$

Note that

$$\mathbb{E}(cZ) = c\mathbb{E}Z$$

for any $c \in \mathbb{C}$. Indeed, writing $c = a + ib$ we obtain

$$\begin{aligned} \mathbb{E}(cZ) &= \mathbb{E}((aX - bY) + i(aY + bX)) \\ &= a\mathbb{E}X - b\mathbb{E}Y + ia\mathbb{E}Y + ib\mathbb{E}X \\ &= (a + ib)(\mathbb{E}X + i\mathbb{E}Y) \\ &= c\mathbb{E}Z. \end{aligned}$$

Lemma 8.1 *Z is integrable if and only if $|Z|$ is integrable. If Z is integrable then $|\mathbb{E}Z| \leq \mathbb{E}|Z|$.*

Proof. Indeed, we have $|Z| = \sqrt{X^2 + Y^2}$ whence

$$\max(|X|, |Y|) \leq |Z| \leq |X| + |Y|.$$

It follows that the integrability of both $|X|, |Y|$ is equivalent to that of $|Z|$.

To prove the inequality $|\mathbb{E}Z| \leq \mathbb{E}|Z|$, consider first the case when $\mathbb{E}Z$ is a non-negative real, that is, when $\mathbb{E}X \geq 0$ and $\mathbb{E}Y = 0$. Then we have

$$|\mathbb{E}Z| = \mathbb{E}Z = \mathbb{E}X + i\mathbb{E}Y = \mathbb{E}X \leq \mathbb{E}|X| \leq \mathbb{E}|Z|.$$

In the general case $\mathbb{E}Z$ is complex valued, so let θ be the polar angle of $\mathbb{E}Z$. Then $\mathbb{E}\left(e^{-i\theta}Z\right)$ is non-negative real, and by the above argument we obtain

$$|\mathbb{E}Z| = \left|e^{-i\theta}\mathbb{E}Z\right| = \left|\mathbb{E}\left(e^{-i\theta}Z\right)\right| \leq \mathbb{E}\left|e^{-i\theta}Z\right| = \mathbb{E}\,|Z|\,,$$

which finishes the proof. ■

Remark. Rewriting the inequality $|\mathbb{E}Z| \leq \mathbb{E}\,|Z|$ in terms of X and Y, we obtain the following non-trivial inequality

$$\sqrt{(\mathbb{E}X)^2 + (\mathbb{E}Y)^2} \leq \mathbb{E}\sqrt{X^2 + Y^2}.$$

8.2 Characteristic functions

Definition. Given a (real-valued) random variable X, its characteristic function $\varphi_X(\lambda)$ is a complex-valued function that is defined for all $\lambda \in \mathbb{R}$ by the identity

$$\varphi_X(\lambda) = \mathbb{E}\left(\exp\left(i\lambda X\right)\right). \tag{8.1}$$

Note that the random variable $\exp(i\lambda X)$ is integrable by Lemma 8.1 because $|\exp(i\lambda X)| = 1$. Using the distribution of X, we can rewrite (8.1) as follows

$$\varphi_X(\lambda) = \int_{\mathbb{R}} \exp(i\lambda x)\, dP_X(x). \tag{8.2}$$

Given any measure μ on $\mathcal{B}(\mathbb{R})$ with $\mu(\mathbb{R}) < \infty$, one can similarly define a characteristic function of μ by

$$\varphi_\mu(\lambda) = \int_{\mathbb{R}} \exp(i\lambda x)\, d\mu(x). \tag{8.3}$$

In Analysis, this function is called the *Fourier transform*[1] of μ and is denoted by $\widehat{\mu}$, that is,

$$\widehat{\mu}(\lambda) = \int_{\mathbb{R}} \exp(i\lambda x)\, d\mu(x).$$

For any integrable function f on $\mathbb{R}$, define its Fourier transform by

$$\widehat{f}(\lambda) = \int_{\mathbb{R}} \exp(i\lambda x)\, f(x)dx,$$

where dx stands for the Lebesgue measure. If f is the density function of measure μ then we clearly have $\widehat{\mu} = \widehat{f}$.

[1]Strictly speaking, to obtain the Fourier transform one should replace in (8.3) $\exp(i\lambda x)$ by $\exp(-i\lambda x)$, but we will neglect that.

Example. Let $X = c$ a.s. where c is a constant. Then

$$\varphi_X(\lambda) = e^{ic\lambda}.$$

Alternatively, if μ is a Dirac measure concentrated at c then $\widehat{\mu}(\lambda) = e^{ic\lambda}$.

More generally, let X be a discrete random variable that takes values $\{c_k\}_{k=0}^{\infty}$ with probabilities $\{p_k\}_{k=0}^{\infty}$. Then

$$\varphi_X(\lambda) = \mathbb{E}\exp(i\lambda X) = \sum_k p_k e^{i\lambda c_k}.$$

Clearly, the right hand side here is a Fourier series. Similarly, if μ is a discrete probability measure with atoms c_k and with the stochastic sequence $\{p_k\}$ then

$$\widehat{\mu}(\lambda) = \sum_k p_k e^{i\lambda c_k}.$$

In a particular case, when X takes the values ± 1 with probabilities $\frac{1}{2}$ each, we obtain

$$\varphi_X(\lambda) = \frac{e^{i\lambda} + e^{-i\lambda}}{2} = \cos\lambda.$$

Example. Let f be the density of the normal distribution $\mathcal{N}(0,1)$, that is

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

Then its Fourier transform is given by

$$\begin{aligned}
\widehat{f}(\lambda) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(i\lambda x) \exp\left(-\frac{x^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(x^2 - 2i\lambda x + (i\lambda)^2\right) - \frac{1}{2}\lambda^2\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\lambda^2}{2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x - i\lambda)^2\, dx\right).
\end{aligned}$$

By the change $z = x - i\lambda$ we obtain

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x - i\lambda)^2\, dx\right) = \int_{\operatorname{Im} z = -\lambda} \exp\left(-\frac{z^2}{2}\right) dz$$

where the latter integral is understood in the sense of contour integration in the complex plane. For any $c > 0$, consider the straight-line segment

$$\alpha_c = \{x + iy : x \in [-c, c],\ y = -\lambda\}$$

so that

$$\int_{\operatorname{Im} z = -\lambda} \exp\left(-\frac{z^2}{2}\right) dz = \lim_{c \to +\infty} \int_{\alpha_c} \exp\left(-\frac{z^2}{2}\right) dz.$$

Consider also the following segments

$$\begin{aligned}\beta_c &= \{x+iy : x=c, y\in[-\lambda,0]\}\\ \gamma_c &= \{x+iy : x\in[-c,c]\,,\ y=0\}\\ \delta_c &= \{x+iy : x=-c,\ y\in[-\lambda,0]\}\end{aligned}$$

and choose their orientation so that $\alpha_c+\beta_c+\gamma_c+\delta_c$ is a closed contour (see Fig. 8.1).

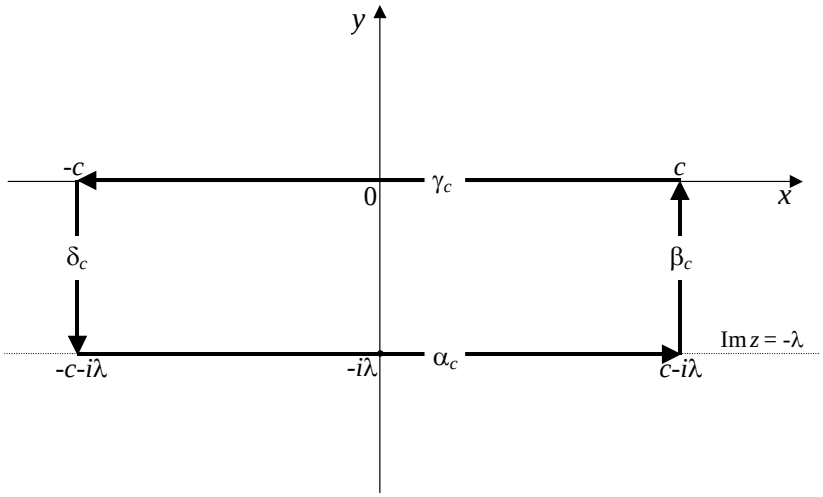


Figure 8.1: A closed contour $\alpha_c+\beta_c+\gamma_c+\delta_c$

Since the function $\exp(-z^2/2)$ is an holomorphic function on entire $\mathbb{C}$, we obtain by the Cauchy theorem

$$\int_{\alpha_c+\beta_c+\gamma_c+\delta_c}\exp\left(-\frac{z^2}{2}\right)dz=0$$

whence

$$\int_{\alpha_c}\exp\left(-\frac{z^2}{2}\right)dz=-\int_{\beta_c+\gamma_c+\delta_c}\exp\left(-\frac{z^2}{2}\right)dz$$

If $c\to\infty$ then

$$\int_{\beta_c}\exp\left(-\frac{z^2}{2}\right)dz\to 0\ \text{ and }\ \int_{\delta_c}\exp\left(-\frac{z^2}{2}\right)dz\to 0$$

because, for $x=\pm c$ and for $|y|\le|\lambda|$,

$$\left|\exp\left(-\frac{z^2}{2}\right)\right|=\left|\exp\left(-\frac{x^2}{2}-ixy+\frac{y^2}{2}\right)\right|\le\exp\left(-\frac{c^2}{2}+\frac{|\lambda|^2}{2}\right)\underset{c\to\infty}{\longrightarrow}0,$$

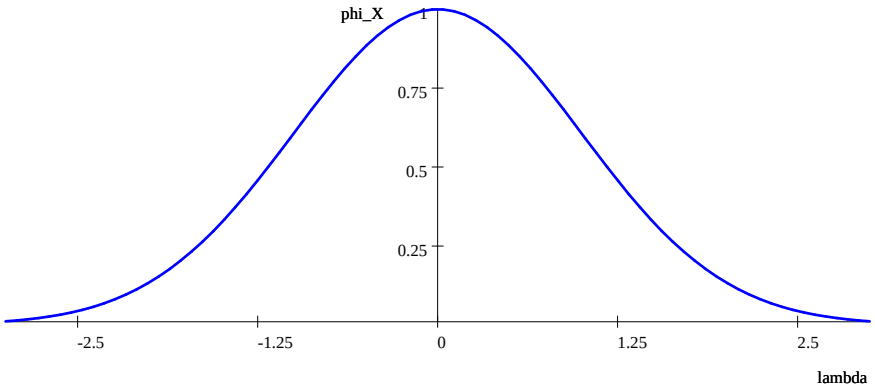


Figure 8.2: The characteristic function of $\mathcal{N}(0,1)$

while the lengths of β_c and δ_c are equal to $|\lambda|$ and, hence, remain bounded. It follows that

$$\begin{aligned}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x-i\lambda)^2\, dx\right) &= \lim_{c\to\infty}\int_{\alpha_c} e^{-z^2/2}dz \\ &= -\lim_{c\to\infty}\int_{\gamma_c} e^{-x^2/2}dx \\ &= \lim_{c\to\infty}\int_{-c}^{c} e^{-x^2/2}dx \\ &= \int_{-\infty}^{+\infty} e^{-x^2/2}dx = \sqrt{2\pi}.\end{aligned}$$

Finally, we conclude

$$\widehat{f}(\lambda) = \exp\left(-\frac{\lambda^2}{2}\right).$$

It follows that if $X \sim \mathcal{N}(0,1)$ then

$$\varphi_X(\lambda) = \exp\left(-\frac{\lambda^2}{2}\right)$$

(see Fig. 8.2).

Example. Let f be the density of the exponential distribution with parameter $a > 0$, that is

$$f(x) = \begin{cases} ae^{-ax}, & x > 0 \\ 0, & x \le 0 \end{cases}.$$

Then its Fourier transform can be computed as follows:

$$\widehat{f}\,(\lambda) = \int_0^{\infty} e^{i\lambda x} a e^{-ax}dx = \int_0^{\infty} e^{-(a-i\lambda)x} a dx = \frac{a}{a-i\lambda}. \tag{8.4}$$

Justification of the last identity in (8.4) can be done in two ways. Indeed, one can change $z = (a - \lambda i)\, x$ so that

$$\int_0^\infty e^{-(a-i\lambda)x} a dx = \frac{a}{a - i\lambda} \int_\gamma e^{-z} dz,$$

where γ is the image of the real half-axis $\{x \geq 0\}$ under the transformation $x \mapsto (a - i\lambda)\, x$, that is, the ray starting from 0 and going through the point $a - i\lambda$ towards ∞. Since $e^{-z} \to 0$ as $z \to \infty$ along γ, one obtains by the fundamental theorem of calculus that

$$\int_\gamma e^{-z} dz = \left[-e^{-z}\right]_0^\infty = 1.$$

Alternatively, one can use the fundamental theorem of calculus directly in (8.4) because the integrand in (8.4) has the primitive function $\frac{-ae^{-(a-i\lambda)x}}{(a-i\lambda)}$.

Example. Let f be the density of the Cauchy distribution, that is

$$f(x) = \frac{1}{\pi\,(1 + x^2)}.$$

To compute its Fourier transform

$$\widehat{f}(\lambda) = \frac{1}{\pi} \int_{-\infty}^\infty \frac{e^{i\lambda x} dx}{1 + x^2}\,,$$

denote

$$g(z) = \frac{1}{\pi} \frac{e^{i\lambda z}}{1 + z^2}$$

and consider, for a large $r > 0$, a segment

$$\alpha_r = \{x + iy : -r \leq x \leq r,\ y = 0\}$$

and a semi-circle

$$\beta_r = \left\{x + iy : x^2 + y^2 = r^2,\ y \geq 0\right\},$$

where α_r is oriented from the negative to positive x and β_r is oriented counter-clockwise so that $\alpha_r + \beta_r$ is closed (see Fig. 8.3).

Since $\alpha_r + \beta_r$ bounds a semi-disk containing the point $z = i$ that is a pole for g, we have by the Cauchy theorem

$$\int_{\alpha_r + \beta_r} g(z) dz = 2\pi i \operatorname*{res}_{z=i} [g(z)]\,.$$

To evaluate the residue, expand g into a power series in $w = z - i$:

$$\begin{aligned} g(z) &= \frac{\exp i\lambda z}{\pi\,(z - i)\,(z + i)} = \frac{\exp\left(-\lambda + i\lambda w\right)}{\pi w\,(2i + w)} \\ &= \frac{e^{-\lambda}}{2\pi i} \frac{1}{w} \frac{\exp\left(i\lambda w\right)}{1 + w/2i} = \frac{e^{-\lambda}}{2\pi i} \frac{1}{w} \left(1 + c_1 w + c_2 w^2 + ...\right), \end{aligned}$$

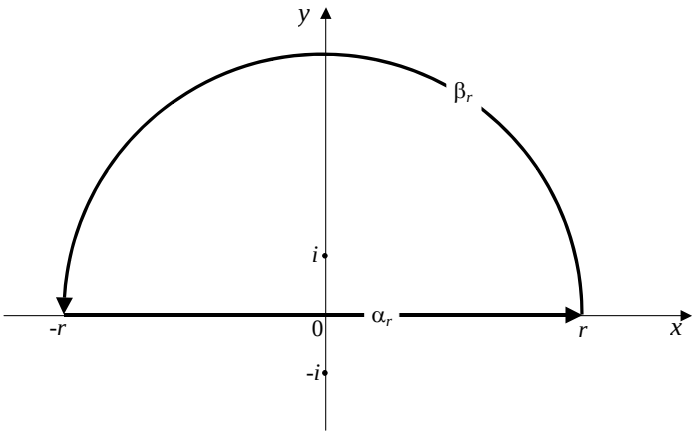


Figure 8.3: A closed contour $\alpha_r + \beta_r$

where $1 + c_1 w + c_2 w^2 + ...$ is the Taylor expansion of the function $\frac{\exp(i\lambda w)}{1+w/2i}$ that is holomorphic in a neighborhood of 0 and takes value 1 at 0. Since $\operatorname{res}_{z=i}[g(z)]$ is equal to the coefficient in front of $\frac{1}{z-i}$ in the Laurent series of $g(z)$, we obtain

$$\operatorname*{res}_{z=i}[g(z)] = \frac{e^{-\lambda}}{2\pi i}.$$

and, hence,

$$\int_{\alpha_r+\beta_r} g(z)dz = e^{-\lambda}.$$

If $\lambda > 0$ then

$$\left|\int_{\beta_r} g(z)dz\right| = \frac{1}{\pi}\left|\int_{\beta_r} \frac{e^{i\lambda x-\lambda y}}{1+z^2}dz\right| < \frac{1}{\pi}\int_{\beta_r} \frac{1}{r^2-1}ds \leq \frac{r}{r^2-1} \underset{r\to\infty}{\longrightarrow} 0,$$

where we have used that $\lambda y \geq 0$ on β_r and the length of β_r is πr. Therefore,

$$\int_{\alpha_r} g(z)dz \underset{r\to\infty}{\longrightarrow} e^{-\lambda}$$

and $\widehat{f}(\lambda) = e^{-\lambda}$. If $\lambda < 0$ then one obtains the same way $\widehat{f}(\lambda) = e^{\lambda}$ by considering a semi-circle in the negative part $y < 0$. Hence,

$$\widehat{f}(\lambda) = e^{-|\lambda|}.$$

It follows that if $X \sim$Cauchy(1) then

$$\varphi_X(\lambda) = e^{-|\lambda|}$$

(see Fig. 8.4).

Let us prove some general properties of characteristic functions.

Lecture 23
06.12.10

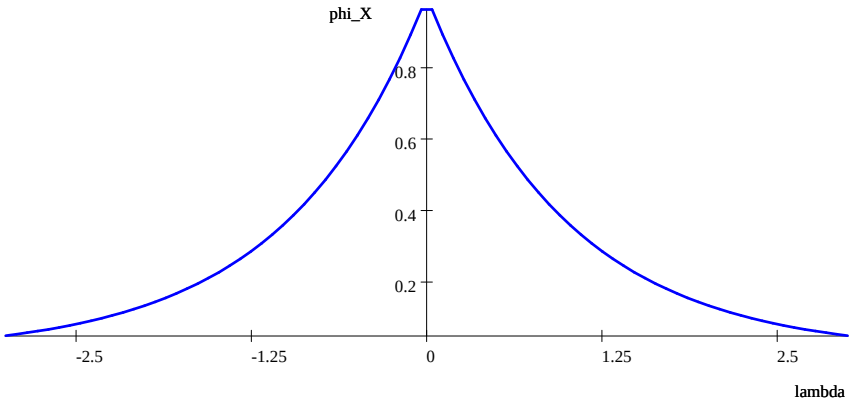


Figure 8.4: The characteristic function of *Cauchy* (1)

Lemma 8.2 *If Z, W two independent complex-valued random variables and both Z, W are integrable then*

$$\mathbb{E}(ZW) = \mathbb{E}Z\ \mathbb{E}W.$$

Recall that this property is true for real valued random variables by Theorem 5.19.

Proof. Let $Z = X + iY$ and $W = U + iV$, where X, Y, U, V are real-valued random variables. By Theorem 5.18, every component of Z is independent of every component of W. Using Theorem 5.19 we obtain

$$\begin{aligned}
\mathbb{E}(ZW) &= \mathbb{E}((X+iY)(U+iV)) \\
&= \mathbb{E}(XU - YV) + i\mathbb{E}(XV + YU) \\
&= \mathbb{E}X\ \mathbb{E}U - \mathbb{E}Y\ \mathbb{E}V + i\mathbb{E}X\ \mathbb{E}V + i\mathbb{E}Y\ \mathbb{E}U \\
&= (\mathbb{E}X + i\mathbb{E}Y)(\mathbb{E}U + i\mathbb{E}V) \\
&= \mathbb{E}Z\ \mathbb{E}W.
\end{aligned}$$

■

Theorem 8.3 *For all real-valued random variables X, Y the following is true:*

(a) *For any constant $c \neq 0$,*

$$\varphi_{cX}(\lambda) = \varphi_X(c\lambda).$$

(b) *If X and Y are independent then*

$$\varphi_{X+Y} = \varphi_X \varphi_Y \tag{8.5}$$

Proof. Indeed, we have

$$\varphi_{cX}(\lambda) = \mathbb{E}\left(e^{i\lambda(cX)}\right) = \mathbb{E}\left(e^{i(\lambda c)X}\right) = \varphi_X(c\lambda)$$

and

$$\varphi_{X+Y}(\lambda) = \mathbb{E}\left(e^{i\lambda(X+Y)}\right) = \mathbb{E}\left(e^{i\lambda X}e^{i\lambda Y}\right) = \mathbb{E}\left(e^{i\lambda X}\right)\mathbb{E}\left(e^{i\lambda Y}\right) = \varphi_X(\lambda)\varphi_Y(\lambda).$$

■

The introduction of characteristic functions is motivated by the property (8.5). As one sees from the proof, (8.5) is a consequence of the property of the exponential function

$$\exp(x+y) = \exp(x)\exp(y).$$

One could have defined the characteristic function by

$$\varphi_X(\lambda) = \mathbb{E}\exp(c\lambda X)$$

with a real-valued constant c, which would also satisfy (8.5). However, with a real-valued c the random variable $\exp(c\lambda X)$ is not necessarily integrable. The choice $c = i$ has an advantage that $\exp(i\lambda X)$ is always integrable.

The identity (8.5) makes it easy to compute the characteristic function for the sum of independent random variables. Using it inductively, one obtains that if $X_1, ..., X_n$ are independent random variables then

$$\varphi_{S_n} = \varphi_{X_1}\varphi_{X_2}...\varphi_{X_n} \tag{8.6}$$

where $S_n = X_1 + ... + X_n$.

Characteristic functions will be used to prove the following main theorem (see Section 8.8).

CENTRAL LIMIT THEOREM. *Let $\{X_n\}$ be a sequence of independent identically distributed random variables with a common finite mean a and a common finite variance b. Then*

$$\frac{S_n - an}{\sqrt{bn}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1) \quad \text{as } n \to \infty. \tag{8.7}$$

In other words, for large n the distribution of $\frac{S_n - an}{\sqrt{bn}}$ is approximately $\mathcal{N}(0,1)$, which means that P_{S_n} is approximately $\mathcal{N}(an, bn)$. That $\mathbb{E}S_n = an$ and $\operatorname{var} S_n = bn$ we have computed in the proof of the weak law of large numbers (Theorem 6.1). We also know that if $X_n \sim \mathcal{N}(a,b)$ then $S_n \sim \mathcal{N}(an, bn)$, that is, S_n is exactly normal. The main point of the central limit theorem is that S_n is asymptotically normal as $n \to \infty$ *regardless* of the common distribution of X_n. Hence, this theorem exhibits a distinguished role and universality of the normal distribution.

The formula (8.6) allows to compute the characteristic function of S_n via the common characteristic function of X_n. Then one proves that $\varphi_{S_n}(\lambda)$ converges to the characteristic function of a normal distribution as $n \to \infty$. Using that, one proves also the convergence (8.7) of distributions.

This approach to the proof of the central limit theorem requires first development of a theory of characteristic functions. In addition to Theorem 8.3, we have to establish the following properties of characteristic functions:

1. φ_Y determines uniquely its distribution measure P_Y (the uniqueness theorem)

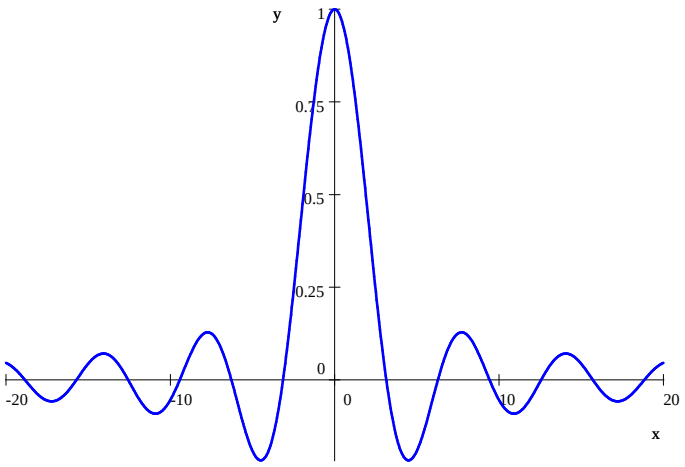


Figure 8.5: The characteristic function of the uniform distribution on $[-1,1]$

2. if $\varphi_{Y_n}(\lambda) \to \varphi_Y(\lambda)$ pointwise as $n \to \infty$ then $Y_n \xrightarrow{\mathrm{D}} Y$ (the continuity theorem).

After having proved these (and other) properties of characteristic functions, we will come back to the proof of the central limit theorem.

Example. Let $X \sim \mathcal{N}(a,b)$. Then $Y = \frac{X-a}{\sqrt{b}} \sim \mathcal{N}(0,1)$ whence

$$\varphi_X(\lambda) = \varphi_{a+\sqrt{b}Y}(\lambda) = \varphi_a(\lambda)\varphi_Y\left(\sqrt{b}\lambda\right) = \exp\left(ia\lambda - \frac{b\lambda^2}{2}\right).$$

Example. Let $X \sim B(n,p)$. To compute φ_X recall that X has the same distribution as $X_1 + ... + X_n$ where X_k are independent Bernoulli random variables $B(1,p)$. Since

$$\varphi_Y(X_k) = \mathbb{E}\left(e^{i\lambda X_k}\right) = e^{i\lambda}p + (1-p),$$

we obtain

$$\varphi_X(\lambda) = \varphi_{X_1}(\lambda)...\varphi_{X_n}(\lambda) = \left(e^{i\lambda}p + (1-p)\right)^n.$$

Example. Let $X \sim \mathcal{U}(-1,1)$. Then

$$\varphi_X(\lambda) = \mathbb{E}\left(e^{i\lambda X}\right) = \int_{\mathbb{R}} e^{i\lambda x} dP_X = \frac{1}{2}\int_{-1}^{1} e^{i\lambda x} dx = \frac{e^{i\lambda} - e^{-i\lambda}}{2i\lambda} = \frac{\sin\lambda}{\lambda}$$

(see Fig. 8.5).

Not every function can be a characteristic function. Consider some general properties of characteristic functions.

Theorem 8.4 *If φ is the characteristic function of a random variable X then the following is true.*

(a) $\varphi(0) = 1$ *and* $|\varphi(\lambda)| \leq 1$ *for all* $\lambda \in \mathbb{R}$.

(b) φ *is uniformly continuous on* $(-\infty, +\infty)$.

(c) φ *is non-negative definite in the following sense: for all reals* $\lambda_1, \lambda_2, ..., \lambda_n$ *and complex* $z_1, z_2, ..., z_n$,

$$\sum_{k,j=1}^{n} \varphi(\lambda_k - \lambda_j) z_k \overline{z_j} \geq 0. \tag{8.8}$$

(d) *If the distribution measure* $\mu = P_X$ *is even (that is, for any Borel set* E *on* $\mathbb{R}$, $\mu(E) = \mu(-E)$*), then* $\varphi(\lambda)$ *is real-valued.*

Remark. The condition (8.8) means that the matrix $(\varphi(\lambda_k - \lambda_j))_{j,k=1}^{n}$ is non-negative definite, for any choice of $\lambda_k \in \mathbb{R}$. By a theorem of Bochner, the conditions $(a), (b), (c)$ are not only necessary but also sufficient for φ to be a characteristic function.

Example. As we already know from the previous examples, the following functions are characteristic functions:

$$1,\ \exp(i\lambda),\ \cos\lambda,\ \exp\left(-\frac{\lambda^2}{2}\right),\ \exp(-|\lambda|),\ \frac{1}{1-i\lambda}, \frac{\sin\lambda}{\lambda}.$$

Let us show that the following functions are not characteristic functions:

$$\sin\lambda,\ \exp(\lambda),\ \cos\left(\lambda^2\right).$$

Indeed, $\sin\lambda$ is not a characteristic function because $\varphi(0) = 0 \neq 1$. Function e^{λ} is not a characteristic function because it is unbounded. Function $\cos\left(\lambda^2\right)$ is not a characteristic function because it is not uniformly continuous, which follows from the observation that its derivative $2\lambda\sin\left(\lambda^2\right)$ is unbounded.

Proof of Theorem 8.4. (a) By (8.1), we have

$$\varphi(0) = \mathbb{E}\exp(i0X) = \mathbb{E}1 = 1.$$

By Lemma 8.1 we have

$$|\varphi(\lambda)| = \left|\mathbb{E}e^{i\lambda X}\right| \leq \mathbb{E}\left(\left|e^{i\lambda X}\right|\right) = 1.$$

(b) We have

$$\begin{aligned} |\varphi(\lambda+h) - \varphi(\lambda)| &= \left|\mathbb{E}\left(e^{i(\lambda+h)X} - e^{i\lambda X}\right)\right| \\ &\leq \mathbb{E}\left|e^{i(\lambda+h)X} - e^{i\lambda X}\right| \\ &= \mathbb{E}\left|e^{i\lambda X}\left(e^{ihX} - 1\right)\right| \\ &= \mathbb{E}\left|e^{ihX} - 1\right|. \end{aligned}$$

Note that the right hand side here does not depend on λ, whence we obtain

$$\sup_{\lambda\in\mathbb{R}} |\varphi(\lambda+h)-\varphi(\lambda)| \leq \mathbb{E}\left|e^{ihX}-1\right|. \tag{8.9}$$

We claim that

$$\mathbb{E}\left|e^{ihX}-1\right| \to 0 \quad \text{as } h\to 0.$$

Clearly, $\left|e^{ihX}-1\right| \to 0$ pointwise as $h\to 0$. The family of random variables $\left|e^{ihX}-1\right|$ is uniformly bounded by 2. Therefore, by the bounded convergence theorem,

$$\lim_{h\to 0}\mathbb{E}\left|e^{ihX}-1\right| = \mathbb{E}\lim_{h\to 0}\left|e^{ihX}-1\right| = 0.$$

It follows from (8.9) that

$$\sup_{\lambda\in\mathbb{R}} |\varphi(\lambda+h)-\varphi(\lambda)| \to 0 \text{ as } h\to 0,$$

that is, φ is uniformly continuous.

(c) By (8.1), we have

$$\begin{aligned}
\sum_{k,j=1}^{n}\varphi(\lambda_k-\lambda_j)z_k\overline{z_j} &= \mathbb{E}\left[\sum_{k,j} e^{i\lambda_k X - i\lambda_j X} z_k\overline{z_j}\right] \\
&= \mathbb{E}\left[\sum_k e^{i\lambda_k X} z_k \sum_j e^{-i\lambda_j X}\overline{z_j}\right] \\
&= \mathbb{E}\left[\sum_k e^{i\lambda_k X} z_k \overline{\sum_k e^{i\lambda_k X} z_k}\right] \\
&= \mathbb{E}\left|\sum_k e^{i\lambda_k X} z_k\right|^2 \geq 0.
\end{aligned}$$

(d) If μ is symmetric then, for any odd μ-integrable function $g(x)$,

$$\int_{-\infty}^{\infty} g\,d\mu = 0.$$

Applying this to $g(x)=\sin\lambda x$, we obtain

$$\varphi(\lambda) = \int_{-\infty}^{\infty}(\cos\lambda x + i\sin\lambda x)\,d\mu = \int_{-\infty}^{\infty}\cos\lambda x\,d\mu,$$

which implies that $\varphi(\lambda)$ is real. ■

8.3 Inversion theorems

8.3.1 Inversion theorem for measures

Let μ be a finite measure on the σ-algebra $\mathcal{B}(\mathbb{R})$. Recall that the distribution function of μ is defined by

$$F_\mu(x) = \mu(-\infty, x]$$

and that by Theorem 2.10[2] measure μ is uniquely determined by F_μ. The Fourier transform of μ (= the characteristic function of μ) is defined by

$$\widehat{\mu}(\lambda) = \varphi_\mu(\lambda) = \int_{-\infty}^{\infty} e^{i\lambda x} d\mu(x)$$

(note that the function $x \mapsto e^{i\lambda x}$ is μ-integrable for any real λ). It turns out that μ can be recovered from $\widehat{\mu}$ as follows.

Theorem 8.5 (The inversion theorem for measures) *Let μ be a finite measure on $\mathcal{B}(\mathbb{R})$ and F_μ be the distribution function of μ. If a, b are the points of continuity of F_μ and $a < b$ then*

$$F_\mu(b) - F_\mu(a) = p.v. \int_{-\infty}^{+\infty} \frac{e^{-ia\lambda} - e^{-ib\lambda}}{2\pi i\lambda} \widehat{\mu}(\lambda) d\lambda. \tag{8.10}$$

Equivalently, for all real $a < b$ which are not atoms for μ,

$$\mu(a, b] = p.v. \int_{-\infty}^{+\infty} \frac{e^{-ia\lambda} - e^{-ib\lambda}}{2\pi i\lambda} \widehat{\mu}(\lambda) d\lambda. \tag{8.11}$$

Here $p.v.$ stands for the "principal value", that is

$$p.v. \int_{-\infty}^{\infty} ... = \lim_{N\to\infty} \int_{-N}^{N} ... .$$

A point a is an atom for measure μ if $\mu(\{a\}) > 0$, which is equivalent to a being a point of discontinuity of F_μ.

Before the proof of Theorem 8.5 let us obtain some consequences.

Corollary 8.6 (a) *Two finite Borel measures μ and ν coincide if and only if their characteristic functions coincide.*

(b) *Two random variables X and Y have the same distribution if and only if they have the same characteristic function.*

Proof. (a) If $\widehat{\mu} = \widehat{\nu}$ then (8.10) implies that for all points $a < b$ at which both F_μ and F_ν are continuous,

$$F_\mu(b) - F_\mu(a) = F_\nu(b) - F_\nu(a).$$

By letting $a \to -\infty$, we obtain

$$F_\mu(b) = F_\nu(b),$$

[2]More precisely, Theorem 2.10 was proved for probability measures, that is, when $\mu(\mathbb{R}) = 1$ where now we consider a more general class of finite measure when $\mu(\mathbb{R}) < \infty$. However, the statement (and the proof) of Theorem 2.10 remains valid with exception that the condition $F(+\infty) = 1$ should be replaced by $F(+\infty) < \infty$.

for all b at which both F_μ and F_ν are continuous. The set of points at which either F_μ or F_ν is *not* continuous, is at most countable. Hence, $F_\mu = F_\nu$ outside a countable set, whence it follows form the right continuity that $F_\mu = F_\nu$ at all points. By Theorem 2.10 we obtain $\mu = \nu$.

(b) Since φ_X is the Fourier transform of measure P_X, we obtain by (a) that $P_X = P_Y$ is equivalent to $\varphi_X = \varphi_Y$. ■

Corollary 8.7 *If $\widehat{\mu}$ is integrable with respect to the Lebesgue measure, that is,*

$$\int_{-\infty}^{\infty} |\widehat{\mu}(\lambda)|\, d\lambda < \infty,$$

then μ has the density function f as follows

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\lambda x} \widehat{\mu}(\lambda) d\lambda. \tag{8.12}$$

Proof. Define function f by (8.12) and prove that f is a density function of μ. For that, it suffice to show that, for all $a < b$,

$$\int_a^b f(x)dx = \mu(a, b]. \tag{8.13}$$

Let us first verify (8.13) provided a and b are not atoms of μ. Indeed, the summability of $\widehat{\mu}$ implies that the function

$$(x, \lambda) \mapsto e^{-i\lambda x} \widehat{\mu}(\lambda)$$

is integrable in $[a, b] \times \mathbb{R}$. Then Fubini's theorem gives

$$\begin{aligned} \int_a^b f(x)dx &= \int_a^b \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-i\lambda x} \widehat{\mu}(\lambda) d\lambda dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_a^b e^{-i\lambda x} \widehat{\mu}(\lambda) dx d\lambda = \int_{-\infty}^{\infty} \frac{e^{-ia\lambda} - e^{-ib\lambda}}{2\pi i \lambda} \widehat{\mu}(\lambda) d\lambda. \end{aligned}$$

By Theorem 8.5, the right hand side here is equal to $\mu(a, b]$.

Since the both sides of (8.13) are right continuous and coincide outside a countable set, they coincide everywhere. Since the left hand side in (8.13) is continuous in b, this implies a posteriori that μ has no atoms. ■

Remark. The right hand side of (8.12) is called the *inverse Fourier transform* of the function $\widehat{\mu}$. Hence, we have the following relation between the density function f (if it exists) and the characteristic function φ of a finite measure (or of a random variable):

$$\varphi(\lambda) = \int_{-\infty}^{\infty} e^{i\lambda x} f(x) dx \tag{8.14}$$

and

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\lambda x} \varphi(\lambda) d\lambda. \tag{8.15}$$

In other words, φ is obtained from f by the Fourier transform, and f is obtained from φ by the inverse Fourier transform. Formulas (8.14) and (8.15) are true in a slightly more general context provided both f and φ are integrable on $\mathbb{R}$ (see Theorem 8.8) below).

Lecture 24
07.12.10

Proof of Theorem 8.5. Assuming that a and b are not atoms of μ, we need to prove

$$\mu(a,b] = \lim_{N\to\infty} \int_{-N}^{+N} \frac{e^{-ia\lambda} - e^{-ib\lambda}}{2\pi i\lambda} \underbrace{\left(\int_{-\infty}^{\infty} e^{i\lambda x} d\mu(x) \right)}_{\widehat{\mu}(\lambda)} d\lambda. \tag{8.16}$$

Using Fubini's theorem, the right hand side of (8.16), before taking lim, is

$$\int_{-N}^{+N} \frac{e^{-ia\lambda} - e^{-ib\lambda}}{2\pi i\lambda} \left(\int_{-\infty}^{\infty} e^{i\lambda x} d\mu(x) \right) d\lambda = \int_{-\infty}^{\infty} \underbrace{\int_{-N}^{+N} \frac{e^{-ia\lambda} - e^{-ib\lambda}}{2\pi i\lambda} e^{i\lambda x} d\lambda}_{\Phi_N(x)} d\mu(x). \tag{8.17}$$

The use of Fubini's theorem is justified because the the integrand $\frac{e^{-ia\lambda}-e^{-ib\lambda}}{2\pi i\lambda} e^{i\lambda x}$ is a bounded function of λ, x that is hence integrable on $(-\infty,+\infty)\times[-N,N]$ with respect to the measure $d\mu \times d\lambda$.

Since

$$\frac{e^{-ia\lambda} - e^{-ib\lambda}}{2\pi i\lambda} = \frac{1}{2\pi} \int_a^b e^{-it\lambda} dt,$$

the internal integral in (8.17) is equal to

$$\begin{aligned}
\Phi_N(x) &= \frac{1}{2\pi} \int_{-N}^{N} \left(\int_a^b e^{-i\lambda t} dt \right) e^{i\lambda x} d\lambda \\
&= \frac{1}{2\pi} \int_a^b \left(\int_{-N}^{N} e^{i\lambda(x-t)} d\lambda \right) dt \\
&= \frac{1}{2\pi} \int_a^b \left(\frac{e^{iN(x-t)} - e^{-iN(x-t)}}{i\,(x-t)} \right) dt \\
&= \frac{1}{\pi} \int_a^b \frac{\sin N\,(x-t)}{x-t} dt.
\end{aligned}$$

Changing $u = N(t-x)$, we obtain

$$\Phi_N(x) = \frac{1}{\pi} \int_a^b \frac{\sin N\,(t-x)}{t-x} dt = \frac{1}{\pi} \int_{N(a-x)}^{N(b-x)} \frac{\sin u}{u} du. \tag{8.18}$$

If $x \in (a,b)$ then $N(b-x) \to +\infty$ and $N(a-x) \to -\infty$ as $N \to \infty$. Therefore, for such x,

$$\lim_{N\to\infty} \Phi_N(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin u}{u} du = 1.$$

If x is outside $[a,b]$ then $\Phi_N(x) \to 0$ since both $N(b-x)$ and $N(a-x)$ go to $+\infty$ or $-\infty$. If $x = a$ or $x = b$ then in the same way

$$\lim_{N\to\infty} \Phi_N(x) = \frac{1}{\pi} \int_0^{\infty} \frac{\sin u}{u} du = 1/2.$$

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

$$\lim_{N\to\infty} \Phi_N(x) = \mathbf{1}_{(a,b)} + \frac{\mathbf{1}_{\{a\}}}{2} + \frac{\mathbf{1}_{\{b\}}}{2}.$$

The function $\Phi_N(x)$ is uniformly bounded for all x and $N > 0$, which follows from (8.18) and the fact that the integral

$$\int_A^B \frac{\sin u}{u} du = \int_0^B \frac{\sin u}{u} du - \int_0^A \frac{\sin u}{u} du$$

is uniformly bounded for all $A, B \in \mathbb{R}$. The latter follows in turn from the continuity of the function

$$A \mapsto \int_0^A \frac{\sin u}{u} du$$

and the existence of its finite limits as $A \to \pm\infty$.

Therefore, the right hand side of (8.16) is

$$\begin{aligned}
\lim_{N\to\infty} \int_{-\infty}^{\infty} \Phi_N(x) d\mu(x) &= \int_{-\infty}^{\infty} \lim_{N\to\infty} \Phi_N(x) d\mu(x) \\
&= \int_{-\infty}^{\infty} \left[\mathbf{1}_{(a,b)} + \frac{\mathbf{1}_{\{a\}}}{2} + \frac{\mathbf{1}_{\{b\}}}{2} \right] d\mu(x) \\
&= \mu(a,b) + \frac{1}{2}\mu(a) + \frac{1}{2}\mu(b).
\end{aligned}$$

Here we have interchanged the integral and the limit by the bounded convergence theorem, using the fact that Φ_N is uniformly bounded and measure μ is finite. Finally, since a and b are not atoms, $\mu(a) = \mu(b) = 0$ whence the claim follows. ■

8.3.2 Inversion theorem for functions

The following theorem is a version of Theorem 8.5 for functions.

Theorem 8.8 (The inversion theorem for functions). *If f is a bounded continuous function on $\mathbb{R}$ and, in addition, f and $\widehat{f}$ are both integrable then*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\lambda x} \widehat{f}(\lambda) d\lambda. \tag{8.19}$$

Remark. Note that if f is integrable then $\widehat{f}$ is bounded and continuous (cf. the proof of Theorem 8.4). Similarly, if $\widehat{f}$ is integrable then by (8.19) f is bounded and continuous.

The hypothesis that $\widehat{f}$ is integrable may be not easy to verify in applications. The following lemma is helpful for this purpose. Denote by $C^2(\mathbb{R})$ the class of functions on $\mathbb{R}$ that are twice continuously differentiable.

Lemma 8.9 *If $f \in C^2(\mathbb{R})$ and the functions f, f', f'' are integrable then f satisfies all the hypotheses of* Theorem 8.8 *and, therefore, the inversion formula* (8.19) *holds for such a function.*

The proof of this lemma will be given later on in Corollary 8.14.

Proof of Theorem 8.8. We have

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\lambda x}\widehat{f}(\lambda)d\lambda = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\lambda x}\left(\int_{-\infty}^{\infty} e^{i\lambda y}f(y)dy\right)d\lambda.$$

We would like to interchange the order of integration. Since this is not possible, we first introduce a regularizing factor $g(\lambda) = \exp\left(-\lambda^2/2\right)$. Since $g(\varepsilon\lambda) \to 1$ as $\varepsilon \to 0$ and $\widehat{f}$ is integrable, we have, by the dominated convergence theorem,

$$\int_{-\infty}^{\infty} e^{-i\lambda x}\widehat{f}(\lambda)d\lambda = \lim_{\varepsilon\to 0}\int_{-\infty}^{\infty} e^{-i\lambda x}\widehat{f}(\lambda)g\left(\varepsilon\lambda\right)d\lambda.$$

Then we have

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{-i\lambda x}\widehat{f}(\lambda)g\left(\varepsilon\lambda\right)d\lambda &= \int_{-\infty}^{\infty} e^{-i\lambda x}g\left(\varepsilon\lambda\right)\left(\int_{-\infty}^{\infty} e^{i\lambda y}f(y)dy\right)d\lambda \\
&= \int_{-\infty}^{\infty} f(y)\left(\int_{-\infty}^{\infty} e^{i\lambda(y-x)}g\left(\varepsilon\lambda\right)d\lambda\right)dy \quad [\text{change } \lambda' = \varepsilon\lambda] \\
&= \int_{-\infty}^{\infty} f(y)\widehat{g}(\frac{y-x}{\varepsilon})\frac{1}{\varepsilon}dy \quad [\text{change } z = \frac{y-x}{\varepsilon}] \\
&= \int_{-\infty}^{\infty} f(x+\varepsilon z)\widehat{g}(z)dz.
\end{aligned}$$

As $\varepsilon \to 0$, we have $f(x+\varepsilon z) \to f(x)$ pointwise. Also, we know that

$$\widehat{g}(z) = \sqrt{2\pi}\exp\left(-z^2/2\right).$$

Since f is bounded, we obtain by the dominated convergence theorem

$$\lim_{\varepsilon\to\infty}\int_{-\infty}^{\infty} f(x+\varepsilon z)\widehat{g}(z)dz = \int_{-\infty}^{\infty} f(x)\widehat{g}(z)dz = \sqrt{2\pi}f(x)\int_{-\infty}^{\infty}\exp\left(-z^2/2\right)dz = 2\pi f(x).$$

We conclude that

$$\lim_{\varepsilon\to 0}\int_{-\infty}^{\infty} e^{-i\lambda x}\widehat{f}(\lambda)g\left(\varepsilon\lambda\right)d\lambda = 2\pi f(x),$$

whence the claim follows. ■

8.4 Plancherel formula

Denote the inverse Fourier transform by

$$\widetilde{f}(\lambda) := \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\lambda x}f(x)dx,$$

assuming that function f is integrable. Clearly,

$$\widetilde{f}(\lambda) = \frac{1}{2\pi}\widehat{f}(-\lambda). \tag{8.20}$$

In particular, $\widehat{f}$ is integrable if and only if $\widetilde{f}$ is integrable. By Theorem 8.8, we have the identity

$$\widetilde{\widehat{f}} = f$$

provided f is bounded and continuous, and both f and $\widehat{f}$ are integrable. This and (8.20) imply that

$$\widehat{\widetilde{f}} = f \tag{8.21}$$

assuming that f is bounded and continuous, and both f and $\widetilde{f}$ are integrable.

Theorem 8.10 (The Plancherel formula) *Let μ be a finite Borel measure. Suppose that f is a bounded continuous function such that f and $\widetilde{f}$ are integrable. Then the following identity holds*

$$\int_{-\infty}^{+\infty} f d\mu = \int_{-\infty}^{+\infty} \widetilde{f}\,\widehat{\mu} d\lambda. \tag{8.22}$$

Proof. Denote $h = \widetilde{f}$. Then by (8.21) $f = \widehat{h}$ so that (8.22) amounts to

$$\int_{-\infty}^{+\infty} \widehat{h} d\mu = \int_{-\infty}^{+\infty} h\,\widehat{\mu}\, d\lambda.$$

Using definitions of $\widehat{h}$ and $\widehat{\mu}$, we rewrite this as

$$\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} h(\lambda) e^{i\lambda x} dx \right) d\mu(x) = \int_{-\infty}^{\infty} h(\lambda) \left(\int_{-\infty}^{\infty} e^{i\lambda x} d\mu(x) \right) d\lambda,$$

which is true by Fubini because h is integrable and μ is finite. ■

Since $\widehat{\mu}$ coincides with the characteristic function φ_μ of measure μ, the identity (8.22) can be rewritten in the form

$$\int_{-\infty}^{+\infty} f d\mu = \int_{-\infty}^{+\infty} \widetilde{f}\,\varphi_\mu d\lambda. \tag{8.23}$$

If X is a random variable then applying this identity to $\mu = P_X$ and using $\varphi_X = \varphi_\mu$, we obtain the identity

$$\mathbb{E}\left(f(X)\right) = \int_{-\infty}^{\infty} \widetilde{f}\,(\lambda)\varphi_X(\lambda)\, d\lambda.$$

If measure μ has the density function g then (8.22) becomes

$$\int_{-\infty}^{+\infty} f g dx = \int_{-\infty}^{+\infty} \widetilde{f}\,\widehat{g} d\lambda.$$

This identity easily extends to any integrable function g.

8.5 The continuity theorem

The next theorem gives another characterization of weak convergence of probability measures. Namely, it turns out that weak convergence of probability measures is *equivalent* to a pointwise convergence of characteristic functions.

Theorem 8.11 *Let $\{\mu_n\}$ be a sequence of probability measures μ be another probability measure.*

(a) *If $\mu_n \Rightarrow \mu$ then $\varphi_{\mu_n}(\lambda) \to \varphi_\mu(\lambda)$ for all $\lambda \in \mathbb{R}$.*

(b) *If $\varphi_{\mu_n}(\lambda) \to \varphi_\mu(\lambda)$ for all $\lambda \in \mathbb{R}$, then $\mu_n \Rightarrow \mu$.*

Remark. The second statement can be strengthened as follows: if $\varphi_{\mu_n}(\lambda) \to \varphi(\lambda)$ for some function φ that is continuous at 0 then φ is the characteristic function of some probability measure μ, and $\mu_n \Rightarrow \mu$ (*Lévy's continuity theorem*). However, in our applications of Theorem 8.11, the limit function φ will obviously be a characteristic function, so we will not have to justify that.

Proof. (a) By definition, $\mu_n \Rightarrow \mu$ if, for any bounded continuous function f,

$$\int_{-\infty}^{\infty} f d\mu_n \to \int_{-\infty}^{\infty} f d\mu. \tag{8.24}$$

In this definition f is a real-valued function but (8.24) holds also for bounded continuous *complex-valued* function f because (8.24) holds separately for $\operatorname{Re} f$ and $\operatorname{Im} f$. Applying (8.24) for $f(x) = e^{i\lambda x}$, we obtain $\varphi_n(\lambda) \to \varphi(\lambda)$, for any $\lambda \in \mathbb{R}$.

(b) We need to verify (8.24) for all bounded continuous functions f. Assume first that f is in addition integrable, and so be $\widetilde{f}$. Using Plancherel's formula (8.23) and the hypothesis that $\varphi_{\mu_n} \to \varphi_\mu$ pointwise, we obtain

$$\int_{-\infty}^{\infty} f d\mu_n = \int_{-\infty}^{\infty} \widetilde{f}\varphi_{\mu_n} d\lambda \to \int_{-\infty}^{\infty} \widetilde{f}\varphi_\mu d\lambda = \int_{-\infty}^{\infty} f d\mu$$

where the convergence takes place by the dominated convergence theorem because all φ_{μ_n} are uniformly bounded and $\widetilde{f}$ is integrable.

Let f be an arbitrary bounded continuous function. We approximate f by a sequence $\{f_k\}$ of such functions that are in addition integrable, and $\widetilde{f_k}$ are also integrable. Indeed, for any $k > 0$ find a function $f_k \in C^2(\mathbb{R})$ with the following properties:

- $\sup_{[-k,k]} |f_k - f| < 2^{-k}$;
- $\operatorname{supp} f \subset [-(k+1), k+1]$;
- $\sup_{\mathbb{R}} |f_k| \le \sup_{\mathbb{R}} |f| + 1$.

Lecture 25
13.12.10

For example, by the Weierstrass approximation theorem (Theorem 6.3), f can be uniformly approximated on $[k,k]$ by a polynomial. Multiplying the polynomial by a cutoff function that is equal to 1 in $[-k,k]$ and vanishes outside $[-(k+\delta),k+\delta]$ for sufficiently small $\delta>0$, we obtain f_k.

Clearly, f_k is bounded, continuous and integrable, and the same is true for the derivatives f_k' and f_k''. Hence, $\widetilde{f_k}$ is also integrable by Lemma 8.9, and (8.24) holds for each function f_k.

For any such k, we have

$$\left|\int_{-\infty}^{\infty} f\,d\mu_n - \int_{-\infty}^{\infty} f\,d\mu\right| \le \int_{-\infty}^{\infty} |f-f_k|\,d\mu_n + \left|\int_{-\infty}^{\infty} f_k d\mu_n - \int_{-\infty}^{\infty} f_k d\mu\right| + \int_{-\infty}^{\infty} |f_k - f|\,d\mu. \tag{8.25}$$

The first (and the third term) on the right hand side is estimated as follows:

$$\begin{aligned}\int_{-\infty}^{\infty} |f-f_k|\,d\mu_n &= \int_{-k}^{k} |f-f_k|\,d\mu_n + \left(\int_k^{\infty} + \int_{-\infty}^{-k}\right)|f-f_k|\,d\mu_n \\ &\le 2^{-k} + C\,[1-\mu_n(-k,k)]\end{aligned}$$

where C is an upper bound constant for all $|f-f_k|$ (for example, can take $C = 2\sup|f|+1$). Hence, we obtain from (8.25)

$$\begin{aligned}\left|\int_{-\infty}^{\infty} f\,d\mu_n - \int_{-\infty}^{\infty} f\,d\mu\right| &\le 2^{-k+1} + C\,[1-\mu_n(-k,k)] + C\,[1-\mu(-k,k)] \\ &\quad + \left|\int_{-\infty}^{\infty} f_k d\mu_n - \int_{-\infty}^{\infty} f_k d\mu\right|.\end{aligned}$$

If $k\to\infty$ then $1-\mu(-k,k)\to 0$. We need an upper bound for $1-\mu_n(-k,k)$ which would be uniform in n. Such an estimate can be obtained using the following lemma that will be proved after we finish the proof of Theorem 8.11.

Lemma 8.12 *For any probability measure ν and for any $k>0$,*

$$1-\nu(-k,k) \le \frac{k}{2}\int_{-2/k}^{2/k} |1-\varphi_\nu(\lambda)|\,d\lambda. \tag{8.26}$$

Given any $\varepsilon>0$, we select k so large that

$$\left|1-\varphi_\mu(\lambda)\right| \le \varepsilon \quad \text{for all } \lambda\in[\frac{2}{k},-\frac{2}{k}],$$

which is possible to do because $\varphi_\mu(0)=0$ and φ_μ is continuous. Then

$$\frac{k}{2}\int_{-2/k}^{2/k} \left|1-\varphi_\mu(\lambda)\right| d\lambda \le \frac{k}{2}\varepsilon\frac{4}{k} = 2\varepsilon.$$

Since $\varphi_{\mu_n}(\lambda)\to\varphi_\mu(\lambda)$ and all functions φ_{μ_n} are uniformly bounded, we obtain that

$$\int_{-2/k}^{2/k} \left|1-\varphi_{\mu_n}(\lambda)\right| d\lambda \to \int_{-2/k}^{2/k} \left|1-\varphi_\mu(\lambda)\right| d\lambda$$

as $n \to \infty$. In particular, for all n large enough,

$$\frac{k}{2}\int_{-2/k}^{2/k} \left|1-\varphi_{\mu_n}(\lambda)\right| d\lambda < 3\varepsilon,$$

whence by (8.26)

$$1-\mu(-k,k) \le 2\varepsilon \quad \text{and} \quad 1-\mu_n(-k,k) < 3\varepsilon.$$

Therefore, for any $\varepsilon > 0$ there exists k such that, for all n large enough,

$$\left|\int_{-\infty}^{\infty} f\, d\mu_n - \int_{-\infty}^{\infty} f\, d\mu\right| \le 2^{-k+1} + 5C\varepsilon + \left|\int_{-\infty}^{\infty} f_k d\mu_n - \int_{-\infty}^{\infty} f_k d\mu\right|.$$

As $n \to \infty$, the last term goes to 0 as was proved above. Hence,

$$\limsup_{n\to\infty} \left|\int_{-\infty}^{\infty} f\, d\mu_n - \int_{-\infty}^{\infty} f\, d\mu\right| \le 2^{-k+1} + 5C\varepsilon,$$

whence the claim follows because ε can be made arbitrarily small and k can be made arbitrarily large. ■

Now let us prove Lemma 8.12 (whereas Lemma 8.9 will be proved in the next section).

Proof of Lemma 8.12. Denote for simplicity $u = 2/k$. Then we need to prove that

$$1-\nu(-k,k) \le \frac{1}{u}\int_{-u}^{u} [1-\varphi_\nu(\lambda)]\, d\lambda. \tag{8.27}$$

We do not write the modulus on the right hand of (8.27) as the integral in (8.27) will happen to be real and non-negative. Using the definition of φ_ν and $\nu(\mathbb{R}) = 1$, the right hand side of (8.27) can be represented as

$$\begin{aligned}
\frac{1}{u}\int_{-u}^{u}\left(\int_{-\infty}^{\infty}\left(1-e^{i\lambda x}\right) d\nu(x)\right) d\lambda &= \frac{1}{u}\int_{-\infty}^{\infty}\int_{-u}^{u}\left(1-e^{i\lambda x}\right) d\lambda d\nu(x) \\
&= 2\int_{-\infty}^{\infty}\left(1-\frac{\sin ux}{ux}\right) d\nu(x) \\
&= 2\left[\int_{(-k,k)} + \int_{(-\infty,-k]} + \int_{[k,\infty)}\right]\left(1-\frac{\sin ux}{ux}\right) d\nu(x) \\
&\ge \left[\int_{(-\infty,-k]} + \int_{[k,\infty)}\right] d\nu(x) \\
&= \nu(-\infty,k] + \nu[k,\infty) = 1-\nu(-k,k).
\end{aligned}$$

Here we used the fact that the integrand $\left(1-\frac{\sin ux}{ux}\right)$ is non-negative as well as

$$1-\frac{\sin ux}{ux} \ge 1-\frac{1}{u\,|x|} = 1-\frac{k}{2\,|x|} \ge \frac{1}{2} \quad \text{for} \quad |x| \ge k.$$

■

8.6 Fourier transform and differentiation

Recall that the Fourier transform $\widehat{f}(\lambda)$ is defined for any integrable function f on $\mathbb{R}$ and is a bounded continuous function of λ. Here we relate the differentiability properties of $\widehat{f}$ with the integrability properties of f and vice versa.

Theorem 8.13 *Let k be a positive integer.*

(a) *Let $f \in C^k(\mathbb{R})$, and assume that all functions $f, f', ..., f^{(k)}$ are integrable. Then*

$$\widehat{\frac{d^k f}{dx^k}} = (-i\lambda)^k \widehat{f}. \tag{8.28}$$

(b) *Let $f(x)$ be an integrable function on $\mathbb{R}$, and assume that $x^k f(x)$ is also integrable. Then $\widehat{f} \in C^k(\mathbb{R})$ and*

$$\frac{d^k \widehat{f}}{d\lambda^k} = \widehat{(ix)^k f}. \tag{8.29}$$

(c) *Let X be a random variable with a finite k-th moment, that is $\mathbb{E}\left(|X|^k\right) < \infty$. Then $\varphi_X \in C^k(\mathbb{R})$ and*

$$\frac{d^k \varphi_X}{d\lambda^k} = \mathbb{E}\left(e^{i\lambda X}(iX)^k\right). \tag{8.30}$$

In particular,

$$\frac{d^k \varphi_X}{d\lambda^k}(0) = i^k \mathbb{E}\left(X^k\right). \tag{8.31}$$

Remark. Denote by D the operation of differentiation, by F the Fourier transform and by M the operation of multiplication by ix (or $i\lambda$ if the argument is λ). Then the formulas (8.28) and (8.29) can be rewritten as

$$FD^k = (-M)^k F \quad \text{and} \quad D^k F = FM^k.$$

In particular, for the case $k = 1$ we have

$$FD = -MF \quad \text{and} \quad DF = FM.$$

In other words, in order to interchange D and F, one has to replace D by M or $-M$, respectively.

Remark. If distribution measure P_X has density f then (8.30) follows (8.29). Indeed, in this case $\varphi_X = \widehat{f}$ and

$$\mathbb{E}\left(e^{i\lambda X}(iX)^k\right) = \int_{\mathbb{R}} e^{i\lambda x}(ix)^k f(x)dx = \widehat{(ix)^k f}.$$

Remark. Formula (8.31) can be used to compute the moments of X via its characteristic functions. The important particular cases are

$$\begin{aligned} \varphi'_X(0) &= i\mathbb{E}X \\ \varphi''_X(0) &= -\mathbb{E}\left(X^2\right). \end{aligned}$$

As a consequence, we see that

$$\operatorname{var} X = -\varphi''_X(0) + \left[\varphi'_X(0)\right]^2.$$

Proof of Theorem 8.13. (a) It suffices to prove (8.28) for the case $k = 1$, that is,

$$\widehat{\frac{df}{dx}} = (-i\lambda)\,\widehat{f},$$

assuming that f and f' are integrable, which then implies (8.28) for all k by induction. For any reals $a < b$, we have

$$\int_a^b f'(x)\, e^{i\lambda x} dx = \int_a^b e^{i\lambda x} df(x) = \left[e^{i\lambda x} f(x)\right]_a^b - i\lambda \int_a^b f(x) e^{i\lambda x} dx.$$

so that

$$\widehat{f}'(\lambda) = \int_{-\infty}^{\infty} f'(x) e^{i\lambda x} dx = \lim_{\substack{a\to-\infty \\ b\to+\infty}} \left(e^{i\lambda x} f(x)\Big|_a^b - i\lambda \int_a^b f(x) e^{i\lambda x} dx \right).$$

Since $|f|$ is integrable, there is a sequence of $a_k \to -\infty$ and $b_k \to +\infty$ such that $|f(a_k)| \to 0$ and $|f(b_k)| \to 0$. Choosing $a = a_k$ and $b = b_k$ in the above computation and letting $k \to \infty$, we obtain

$$\widehat{f}'(\lambda) = -i\lambda \int_{-\infty}^{\infty} f(x) e^{i\lambda x} dx = (-i\lambda)\,\widehat{f}(\lambda),$$

which was to be proved.

(b) Note that if f and $x^k f$ are integrable then also $x^m f$ is integrable for any $m = 0, 1, 2, ..., k$ by the following inequality

$$|x|^m \le 1 + |x|^k.$$

Therefore, it suffices to prove (8.29) for $k = 1$ and then use induction in k. Assume that $f(x)$ and $xf(x)$ are integrable. Differentiating $\widehat{f}$, we obtain

$$\begin{aligned}
\frac{d}{d\lambda}\widehat{f} &= \lim_{h\to 0}\int_{-\infty}^{\infty}\frac{e^{i(\lambda+h)x}-e^{i\lambda x}}{h}f(x)dx \\
&= \int_{-\infty}^{\infty}\lim_{h\to 0}\frac{e^{i(\lambda+h)x}-e^{i\lambda x}}{h}f(x)dx \\
&= \int_{-\infty}^{\infty}\frac{d}{d\lambda}e^{i\lambda x}f(x)dx \\
&= \int_{-\infty}^{\infty}ixe^{i\lambda x}f(x)dx \\
&= \widehat{ixf}.
\end{aligned}$$

The interchange of the order of $\lim_{h\to 0}$ and integration is justified as follows. We have

$$\left|\frac{e^{i(\lambda+h)x}-e^{i\lambda x}}{h}\right| = \left|\frac{e^{ihx}-1}{hx}\right||x| \le |x|$$

where we have used the inequality

$$\left|\frac{e^{i\xi}-1}{\xi}\right| \le 1$$

which follows from $\left|\left(e^{i\xi}\right)'\right| \le 1$. Therefore,

$$\left|\frac{e^{i(\lambda+h)x}-e^{i\lambda x}}{h}f(x)\right| \le |xf(x)|.$$

Since function $|xf(x)|$ is integrable, the interchange of lim and integration follows from the dominated convergence theorem.

Hence, we have obtained that $\widehat{f}$ is differentiable at any real λ and

$$\frac{d}{d\lambda}\widehat{f} = \widehat{ixf}.$$

Since the right hand side is a continuous function as a Fourier transform of an integrable function, we conclude that $\widehat{f} \in C^1(\mathbb{R})$.

(c) The finiteness of k-th moment implies the finiteness of all m-th moments with $m < k$. Similarly to part (b) we have

$$\frac{d}{d\lambda}\varphi_X(\lambda) = \frac{d}{d\lambda}\int_{-\infty}^{\infty}e^{i\lambda x}dP_X = \int_{-\infty}^{\infty}\frac{d}{d\lambda}e^{i\lambda x}dP_X = \int_{-\infty}^{\infty}ixe^{i\lambda x}dP_X = \mathbb{E}\left(e^{i\lambda X}iX\right),$$

which proves (8.30) for $k = 1$. By induction we obtain (8.30) for all k. ■

Corollary 8.14 (=Lemma 8.9) *Let $f \in C^2(\mathbb{R})$, and assume that all functions f, f', f'' are integrable. Then f satisfies all the hypotheses of the inversion theorem 8.8.*

Proof. We need to verify that f if bounded, continuous, and functions f and $\widehat{f}$ are integrable. Since the continuity and integrability of f are given, it remains to show that f is bounded and that $\widehat{f}$ is integrable. Since

$$f(x) = f(0) + \int_0^x f'(t)\, dt$$

and

$$|f(x)| \leq |f(0)| + \int_{-\infty}^{+\infty} |f'(t)|\, dt,$$

the boundedness of f follows from the integrability of f'.

Since the function $\widehat{f}$ is continuous (as the Fourier transform of an integrable function), it is integrable on any bounded interval. By (8.28), we have

$$\widehat{f}(\lambda) = -\frac{1}{\lambda^2}\widehat{f''}(\lambda).$$

Since $\widehat{f''}$ is bounded (as the Fourier transform of an integrable function), we conclude that, for large λ,

$$\left|\widehat{f}(\lambda)\right| \leq \frac{C}{\lambda^2}$$

whence we see that the function $\widehat{f}$ is integrable at ∞ and, hence, on the entire $\mathbb{R}$. ■

8.7 A summary of the properties of characteristic functions

Let us now list all the properties of characteristic functions that have been established above and that will be used in the next section to prove the central limit theorem.

1. The characteristic function φ_X of a random variable X is a uniformly continuous function on $\mathbb{R}$, such that $\varphi_X(0) = 1$ and $|\varphi_X(\lambda)| \leq 1$ (Theorem 8.4).

2. If $X_1, X_2, ..., X_n$ are independent random variables then

$$\varphi_{X_1+...+X_n} = \varphi_{X_1}\varphi_{X_2}...\varphi_{X_n}$$

(Theorem 8.3).

3. φ_X determines uniquely its distribution measure P_X and, hence, its distribution function F_X (Corollary 8.6).

4. $X_n \xrightarrow{\mathrm{D}} X$ if and only if $\varphi_{X_n}(\lambda) \to \varphi_X(\lambda)$ for any $\lambda \in \mathbb{R}$ (Theorem 8.11).

5. If $\mathbb{E}|X| < \infty$ then $\varphi_X \in C^1$ and $\varphi'_X(0) = i\mathbb{E}X$.
If $\mathbb{E}|X|^2 < \infty$ then $\varphi_X \in C^2$ and $\varphi''_X(0) = -\mathbb{E}X^2$ (Theorem 8.13).

8.8 The central limit theorem

The following theorem is one of the main results of this course.

Theorem 8.15 *Let $\{X_n\}$ be a sequence of independent identically distributed random variables with a common finite expectation a and a common finite variance b. Let $S_n = X_1 + X_2 + ... + X_n$. Then*

$$\frac{S_n - an}{\sqrt{bn}} \xrightarrow{\text{D}} \mathcal{N}(0,1) \quad \text{as } n \to \infty. \tag{8.32}$$

In particular, if $a = 0$ then

$$\frac{S_n}{\sqrt{n}} \xrightarrow{\text{D}} \mathcal{N}(0,b) \quad \text{as} \quad n \to \infty.$$

Corollary 8.16 *For all real x,*

$$\mathbb{P}\left(S_n \le an + x\sqrt{bn}\right) \longrightarrow \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \quad \text{as } n \to \infty. \tag{8.33}$$

Also, for all real x and y such that $y < x$,

$$\mathbb{P}\left(an + y\sqrt{bn} < S_n \le an + x\sqrt{bn}\right) \longrightarrow \int_{y}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \quad \text{as } n \to \infty. \tag{8.34}$$

Proof. Indeed, (8.32) and Theorem 7.3 imply that the distribution function of $\frac{S_n - an}{\sqrt{bn}}$ converges to $F(x)$ - the distribution function of $\mathcal{N}(0,1)$ at the points of continuity of F, that is, at all points, whence (8.33) follows. Obviously, (8.34) follows from (8.33) by considering $F(x) - F(y)$. ■

The central limit theorem can be regarded as a fundamental theorem of probability theory. Apart from numerous practical applications, based on the fact that S_n has approximately the normal distribution $\mathcal{N}(an, bn)$ *regardless* of the distribution of X_n, the very fact of existence of the limit distribution has a far reaching consequences for other branches of mathematics and science. One says that the normal distribution is the *law of attraction* of sequences $\{X_n\}$ with finite second moment.

If the higher moments of X_n are finite then it is possible to estimate the rate of convergence in (8.32). We state the following theorem without proof.

THEOREM. *Let $\{X_n\}$ be a sequence of independent identically distributed random variables with a mean 0, a variance 1 and a finite third moment c^3. Let $S_n = X_1 + X_2 + ... + X_n$. Then*

$$\sup_{x \in \mathbb{R}} \left| F_{\frac{S_n}{\sqrt{n}}}(x) - F_{\mathcal{N}(0,1)}(x) \right| \le \frac{c^3}{\sqrt{n}}.$$

Lecture 26
14.12.10

Proof of Theorem 8.15. Renaming $\frac{X_n-a}{\sqrt{b}}$ by X_n, we may assume without loss of generality that $a=0$ and $b=1$. Let $\varphi(\lambda)$ be the common characteristic function of each X_n. By Theorem 8.3, the characteristic function of $\frac{S_n}{\sqrt{n}}$ is

$$\varphi_{\frac{S_n}{\sqrt{n}}}(\lambda) = \varphi^n(\frac{\lambda}{\sqrt{n}}).$$

By Theorem 8.11, to prove that $\frac{S_n}{\sqrt{n}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1)$, it suffices to show that $\varphi_{\frac{S_n}{\sqrt{n}}}(\lambda)$ converges to the characteristic function of $\mathcal{N}(0,1)$ as $n\to\infty$, that is, for any $\lambda\in\mathbb{R}$,

$$\lim_{n\to\infty} \varphi^n(\frac{\lambda}{\sqrt{n}}) = \exp\left(-\frac{\lambda^2}{2}\right).$$

Let us apply Taylor's formula to expand $\varphi(\lambda)$ near 0:

$$\varphi(\lambda) = 1 + \varphi'(0)\lambda + \frac{\varphi''(0)}{2}\lambda^2 + o(\lambda^2), \quad \text{as } \lambda\to 0. \tag{8.35}$$

Indeed, as follows from Theorem 8.13, $\varphi\in C^2$ because X_n has finite second moment. Moreover, (8.31) yields

$$\begin{aligned} \varphi'(0) &= i\mathbb{E}\,(X_n) = 0 \\ \varphi''(0) &= -\mathbb{E}\left(X_n^2\right) = -1, \end{aligned}$$

so that the expansion (8.35) amounts to

$$\varphi(\lambda) = 1 - \frac{\lambda^2}{2} + o(\lambda^2).$$

Replacing λ by $\lambda/\sqrt{n}$, we see that

$$\varphi(\frac{\lambda}{\sqrt{n}}) = 1 - \frac{\lambda^2}{2n} + o(\frac{1}{n}),$$

provided λ is fixed but $n\to\infty$. Finally, we obtain

$$\varphi^n(\frac{\lambda}{\sqrt{n}}) = \left[1 - \frac{\lambda^2}{2n} + o(\frac{1}{n})\right]^n \to \exp\left(-\frac{\lambda^2}{2}\right),$$

which was to be proved.

In the last line we have used the following fact from analysis: if $\{z_n\}$ is a sequence of complex numbers such that $z_n\to z$ then

$$\left(1+\frac{z_n}{n}\right)^n \to \exp(z) \text{ as } n\to\infty.$$

We have

$$\left(1+\frac{z}{n}\right)^n \to \exp(z) \quad \text{as } n\to\infty$$

and

$$\frac{1+\frac{z_n}{n}}{1+\frac{z}{n}} = 1 + w_n$$

where

$$w_n := \frac{z_n - z}{n+z} = o\left(\frac{1}{n}\right) \quad \text{as } n \to \infty. \tag{8.36}$$

Therefore, it suffices to prove that

$$(1+w_n)^n \to 1 \quad \text{as} \quad n \to \infty. \tag{8.37}$$

By the binomial formula we have

$$(1+w_n)^n - 1 = nw_n + \binom{n}{2} w_n^2 + ... + \binom{n}{n} w_n^n.$$

By (8.36), for any $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that $|w_n n| < \varepsilon$ for all $n \geq N$. Using the obvious estimate $\binom{n}{k} \leq n^k$, we obtain, for all $n \geq N$,

$$|(1+w_n)^n - 1| \leq nw_n + (nw_n)^2 + ... + (nw_n)^n \leq \sum_{k=1}^{\infty} \varepsilon^k = \frac{\varepsilon}{1-\varepsilon},$$

whence (8.37) follows. ■

The history of the central limit theorem started in 1733 when de Moivre discovered (8.32) for Bernoulli random variables X_n taking values 0 and 1 with probability $p = \frac{1}{2}$. This was later extended by Laplace to general p and nowadays is referred to as the Moivre-Laplace theorem. The original proof of that theorem was based on a difficult analysis of the binomial distribution (indeed, as we know $S_n \sim B(n,p)$), and some points of the proof were not quite rigorous. The modern rigorous proof that works for an arbitrary distribution of X_n was discovered by Lyapunov in 1901. For that proof, Lyapunov introduced the method of characteristic functions that has become since a very powerful tool for many other problems. Lyapunov's proof was presented above.

The adjective "central" refers both to the central role this theorem plays in probability theory, and to the following observation. Choose in (8.34) $x > 0$ and $y = -x$ so that (8.34) becomes

$$\mathbb{P}\left(|S_n - an| \leq x\sqrt{bn}\right) \to \int_{-x}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \tag{8.38}$$

as $n \to \infty$. If x is large enough then the integral in the right hand side of (8.38) is closed to 1. Then we obtain for large n that

$$\mathbb{P}\left(|S_n - an| \leq x\sqrt{bn}\right) \approx 1.$$

Hence, S_n concentrates near its *central* value an with the error of the order $x\sqrt{bn}$. If $a \neq 0$ then for large n

$$x\sqrt{bn} << an$$

so that with high probability S_n concentrates in a relatively small neighborhood of an.

Example. Consider a sequence of long computations performed by a computer. Let X_n be a rounding error at step n and $S_n = X_1 + ... + X_n$ – the error after n steps (this can be an absolute error by addition and relative error by multiplication). Assume that X_n are independent and uniformly distributed on $[-\varepsilon, \varepsilon]$ so that $\mathbb{E}X_n = 0$ and $b = \operatorname{var} X_n = \frac{1}{3}\varepsilon^2$. It follows from (8.38) that

$$\mathbb{P}\left(|S_n| \le x\sqrt{bn}\right) \approx \int_{-x}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$$

Taking $x = 3$ and noticing that

$$\int_{-3}^{3} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \approx 0.9973,$$

we obtain

$$\mathbb{P}\left(|S_n| \le \sqrt{3n}\varepsilon\right) \approx 0.9973.$$

Example. (A practical experiment). Suppose we are given a dice (see Fig. 8.6) displaying numbers 1, 2, 3, 4, 5, 6 on its faces, but the dice is unfair: after rolling, it shows the numbers with different (unknown) probabilities.

Figure 8.6: A dice

Let X_n be the rolled number after n-th trial. Let us try to determine the unknown values $a = \mathbb{E}X_n$ and $b = \operatorname{var} X_n$ by watching a long series of trial. Assuming that the trials are independent, the sum $S_n = X_1 + ... + X_n$ must satisfy the laws of large numbers and the central limit theorem. By the strong law,

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} a \quad \text{as } n \to \infty,$$

so that a can be determined experimentally as $a \approx \frac{S_n}{n}$ for large n. Determining of the variance b is more involved. Fix some large n and make a large number of sequences of dice rolling computing each time S_n. Let $\Phi_n(x)$ be the frequency function of S_n, that is, $\Phi_n(x)$ is proportional to the number of trials when $S_n = x$, and Φ_n is normalized by the condition

$$\int_{\mathbb{R}} \Phi_n(x)\, dx = 1.$$

By the central limit theorem, $\Phi_n(x)$ must be almost $\mathcal{N}(a, b)$. Knowing $\Phi_n(x)$, one determines the parameters a, b to provide the best match between Φ_n and $\mathcal{N}(a, b)$.

Using the method of the proof of Theorem 8.15, we can prove the following version of the weak law of large numbers.

Theorem 8.17 *If $\{X_n\}$ are independent identically distributed random variables with a common finite mean $\mathbb{E}X_n = a$ then*

$$\frac{S_n}{n} \xrightarrow{\text{P}} a.$$

The difference with Theorem 6.1 is that in Theorem 8.17 we do not assume the finiteness of the variance, but in return require that all X_n have identical distribution.

Proof. Due to Theorem 7.5, it suffices to show that

$$\frac{S_n}{n} \xrightarrow{\text{D}} a. \tag{8.39}$$

Let φ be the common characteristic function of X_n so that

$$\varphi_{S_n/n} = \varphi\left(\frac{\lambda}{n}\right)^n$$

Since X_n is integrable, the characteristic function φ is differentiable and

$$\varphi'(0) = i\mathbb{E}X_n = ia.$$

It follows that, for a fixed λ and $n \to \infty$,

$$\varphi\left(\frac{\lambda}{n}\right)^n = \left(1 + \varphi'(0)\frac{\lambda}{n} + o(\frac{1}{n})\right)^n = \left(1 + \frac{ia\lambda}{n} + o(\frac{1}{n})\right)^n \to \exp(ia\lambda).$$

Hence, $\varphi_{S_n/n}(\lambda) \to \varphi_a(\lambda)$ whence (8.39) follows by Theorem 8.11. ■

Example. Recall that gamma distribution $\Gamma(t, 1)$ with parameters $t > 0$ and 1 is given by the density function

$$f(x) = \frac{x^{t-1}e^{-x}}{\Gamma(t)},\ x > 0,$$

where

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-t}dx.$$

In particular, for $n \in \mathbb{N}$ we have $\Gamma(n) = (n-1)!$. The characteristic function of $\Gamma(t, 1)$ can be computed as follows:

$$\begin{aligned}
\varphi(\lambda) &= \int_{-\infty}^{\infty} e^{i\lambda x} f(x)\, dx \\
&= \frac{1}{\Gamma(t)} \int_0^\infty x^{t-1}e^{-x(1-i\lambda)}dx \quad [\text{change } z = (1-i\lambda)x] \\
&= \frac{1}{\Gamma(t)(1-i\lambda)^t} \int_\gamma z^{t-1}e^{-z}dz \\
&= (1-i\lambda)^{-t},
\end{aligned}$$

where γ is the ray that starts at 0 and goes to ∞ through the point $1-i\lambda$. We have used here that

$$\int_{\gamma} z^{t-1}e^{-z}dz = \Gamma(t),$$

which can be justified using the Cauchy formula.

It follows that the product of the characteristic functions of $\Gamma(t_1,1)$ and $\Gamma(t_2,1)$ is the characteristic function of $\Gamma(t_1+t_2,1)$. Therefore, the sum of two independent random variables with distributions $\Gamma(t_1,1)$ and $\Gamma(t_2,1)$ has the distribution $\Gamma(t_1+t_2,1)$.

Observe that $\Gamma(1,1)$ coincides with the exponential distribution with density

$$f(x) = e^{-x}, \quad x \geq 0.$$

Let $\{X_n\}$ be a sequence of independent identically distributed variables with $X_n \sim \Gamma(1,1)$. By the aforementioned property of the gamma distribution, we have $S_n \sim \Gamma(n,1)$, that is,

$$f_{S_n}(x) = \frac{x^{n-1}e^{-x}}{\Gamma(n)}, \ x>0,$$

and

$$\varphi_{S_n}(\lambda) = (1-i\lambda)^{-n}.$$

By Exercise 40, we have $\mathbb{E}X_n = \operatorname{var} X_n = 1$. For the normalized sum $Y_n = \frac{S_n - n}{\sqrt{n}}$ we have by Theorem 5.15

$$f_{Y_n}(x) = \frac{\sqrt{n}\,(n+\sqrt{n}x)^{n-1}\,e^{-\left(n+\sqrt{n}x\right)}}{\Gamma(n)}, \quad x > -\sqrt{n}, \tag{8.40}$$

and by Theorem 8.3

$$\varphi_{Y_n}(\lambda) = e^{-in\frac{\lambda}{\sqrt{n}}}\left(1 - i\frac{\lambda}{\sqrt{n}}\right)^{-n}.$$

By the (proof of the) central limit theorem, we obtain

$$\varphi_{Y_n}(\lambda) \to \exp\left(-\frac{\lambda^2}{2}\right) \quad \text{as } n \to \infty. \tag{8.41}$$

We claim that in the present setting we can integrate this convergence over $\mathbb{R}$ against the Lebesgue measure. Indeed, we have

$$\left|\varphi_{Y_n}(\lambda)\right| = \left|1 - i\frac{\lambda}{\sqrt{n}}\right|^{-n} = \left(1 + \frac{\lambda^2}{n}\right)^{-n/2}.$$

Using Bernoulli's inequality

$$\left(1 + \frac{\lambda^2}{n}\right)^{n/2} \geq 1 + \frac{\lambda^2}{2},$$

we obtain

$$\left|\varphi_{Y_n}(\lambda)\right| \leq \frac{1}{1+\frac{\lambda^2}{2}}.$$

Since the function $\frac{1}{1+\frac{\lambda^2}{2}}$ is integrable on $\mathbb{R}$ with respect to the Lebesgue measure, we conclude by the dominated convergence theorem that (8.41) implies, for any $x \in \mathbb{R}$,

$$\frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\lambda x}\varphi_{Y_n}(\lambda)\,d\lambda \underset{n\to\infty}{\longrightarrow} \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\lambda x}\exp\left(-\frac{\lambda^2}{2}\right)d\lambda.$$

On the both sides we have the inverse Fourier transform whence by the inversion theorem

$$f_{Y_n}(x) \to \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right). \tag{8.42}$$

(cf. Fig. 8.7).

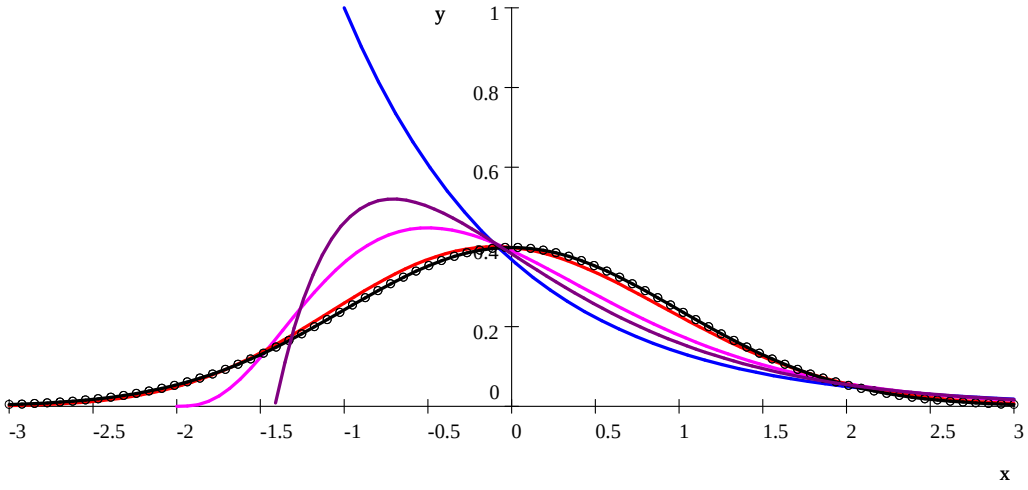


Figure 8.7: The graphs of f_{Y_n} for $n = 1, 2, 4, 100$ and that of $\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ (dotted)

The relation (8.42) is a particular case of a *local* central limit theorem, that asserts convergence of the density functions of Y_n to the normal density function under certain assumptions. Substituting f_{Y_n} from (8.40) and taking $x = 0$, we obtain

$$\frac{\sqrt{n}n^{n-1}e^{-n}}{\Gamma(n)} \to \frac{1}{\sqrt{2\pi}} \quad \text{as } n \to \infty,$$

that is nothing other than the Stirling formula

$$\Gamma(n) \sim \sqrt{2\pi}n^{n-1/2}e^{-n} \text{ as } n \to \infty.$$

Hence, the Stirling formula can be regarded as a manifestation of the local central limit theorem.

As we have seen above, the proofs of Theorem 8.15 and 8.17 make a strong use of the finiteness of the first and the second moments of X_n. Without that, other limiting behavior may take place, as one can see from examples below.

Example. If $X \sim Cauchy\,(c)$ that is X has the density

$$\frac{c}{\pi\left(x^2+c^2\right)}$$

then $Y = X/c \sim Cauchy\,(1)$ whence

$$\varphi_X(\lambda) = \varphi_{cY}(\lambda) = \varphi_Y(c\lambda) = \exp\left(-c\,|\lambda|\right).$$

If X and Y are two independent random variables such that $X \sim Cauchy\,(a)$ and $Y \sim Cauchy\,(b)$ then

$$\varphi_{X+Y}(\lambda) = \varphi_X(\lambda)\varphi_Y(\lambda) = \exp\left(-\left(a+b\right)|\lambda|\right)$$

so that $X + Y \sim Cauchy(a+b)$.

Let now $\{X_n\}$ be a sequence of independent identically distributed random variables such that $X_n \sim Cauchy\,(1)$. Then we have

$$S_n := X_1 + X_2 + ... + X_n \sim Cauchy(n)$$

which implies that, for all n,

$$\frac{S_n}{n} \sim Cauchy(1).$$

We see that the central limit theorem does not hold for this sequence. Under appropriate assumptions about the distribution of X_n one can show that

$$\frac{S_n}{n} \xrightarrow{\text{D}} Cauchy(1).$$

One says that the Cauchy distribution is the law of attraction for such sequences.

Example. It is possible to prove that the function

$$\varphi(\lambda) = \exp\left(-c\,|\lambda|^\alpha\right)$$

is a characteristic function of a certain probability measure, provided $\alpha \in (0,2]$ and $c > 0$. This measure is called a *symmetric α-stable* distribution with parameter c and is denoted by $\mathcal{S}_\alpha(c)$. For example, $Cauchy(c) = \mathcal{S}_1(c)$ and $\mathcal{N}(0,b) = \mathcal{S}_2(\frac{b}{2})$.

If X and Y are independent random variables such that $X \sim \mathcal{S}_\alpha(a)$ and $Y \sim \mathcal{S}_\alpha(b)$ then

$$\varphi_{X+Y}(\lambda) = \exp\left(-\left(a+b\right)|\lambda|^\alpha\right)$$

whence it follows that $X + Y \sim \mathcal{S}_\alpha(a+b)$. Note also that if $X \sim \mathcal{S}_\alpha\,(c)$ then $mX \sim \mathcal{S}_\alpha\left(|m|^\alpha\, c\right)$ because

$$\varphi_{mX}(\lambda) = \varphi_X\left(m\lambda\right) = \exp\left(-c\,|m\lambda|^\alpha\right) = \exp\left(-\,|m|^\alpha\, c\,|\lambda|^\alpha\right).$$

Let now $\{X_n\}$ be a sequence of independent identically distributed random variables such that $X_n \sim \mathcal{S}_\alpha(1)$. Then $S_n \sim \mathcal{S}_\alpha(n)$ and, consequently

$$\frac{S_n}{n^{1/\alpha}} \sim \mathcal{S}_\alpha\left(\left(n^{-1/\alpha}\right)^\alpha n\right) = \mathcal{S}_\alpha(1).$$

More generally, under appropriate assumptions about the distribution of X_n one can show that

$$\frac{S_n}{n^{1/\alpha}} \xrightarrow{\mathrm{D}} \mathcal{S}_\alpha(1).$$

This constitutes another type of the central limit theorem, where the law of attraction is $\mathcal{S}_\alpha(1)$. Note that for the case $\alpha = 1$ we obtain the previous example with the Cauchy distribution, while for $\alpha = 2$ – the classical central limit theorem where the law of attraction is the normal distribution.

8.9 Appendix: the list of useful distributions

name	*notation*	*density or stoch. sequence*	*characteristic function*	$\mathbb{E}X$	$\operatorname{var} X$
normal	$\mathcal{N}(a,b)$	$\frac{1}{\sqrt{2\pi b}}\exp\left(-\frac{(x-a)^2}{2b}\right)$	$\exp\left(ia\lambda - \frac{b\lambda^2}{2}\right)$	a	b
binomial	$B(n,p)$	$\binom{n}{k}p^k(1-p)^{n-k}$ $k=0,...,n$	$\left(e^{i\lambda}p+(1-p)\right)^n$	np	$np(1-p)$
Cauchy	$Cauchy(a)$	$\frac{1}{\pi}\frac{a}{a^2+x^2}$	$\exp(-a\lvert x\rvert)$	–	–
exponential	$Exp(a)$	ae^{-ax}, $x>0$	$\frac{a}{a-i\lambda}$	$1/a$	$1/a^2$
gamma	$\Gamma(a,1)$	$\frac{x^{a-1}e^{-x}}{\Gamma(a)}$, $x>0$	$(1-i\lambda)^{-a}$	a	a
Poisson	$Po(a)$	$e^{-a}\frac{a^k}{k!}$, $k=0,1,...$	$\exp\left(a\left(e^{i\lambda}-1\right)\right)$	a	a
uniform	$\mathcal{U}(a,b)$	$\frac{1}{b-a}$, $a<x<b$	$\frac{e^{i\lambda b}-e^{i\lambda a}}{i\lambda(b-a)}$	$\frac{a+b}{2}$	$\frac{1}{12}(b-a)^2$