Presentation of the lectures

- This is a one-semester class about Probability Theory. It is advisable to remember what you have seen last year about measure theory. We will quickly review the basics however.
- Contact me at `mmariani@hse.ru` for questions and comments. It is advisable to add TERVER in the email subject to skip spam filters, especially if you do not write from the university email.
- Website at `https://sites.google.com/view/terver2122`
- Feedback is very welcome.

Bibliography

- Resources (lectures, exercise sheets, assignments, extra material, references) on the website.
- A. Klenke, Probability Theory - A Comprehensive Course (2014).
- A.N. Shiryaev, Probability (ENG 1996, RUS 2011).
- P. Billingsley, Probability and measure (1995).

History in a nutshell

- What we usually call today *probability theory*, was born in the XVII century in France. Gamblers asked the assistance of mathematical-minded people. One of the oldest questions recorded concerned the following game. The gambler had to throw 24 times a pair of dice, and he would win twice his bet if he gets at least one pair of six. The game is unfavorable to the gambler: the probability of getting a pair of six is $1/36$. So the probability of not getting any pair in 24 tries is $(1 - 1/36)^{24}$. Therefore he wins with probability $1 - (35/36)^{24} \simeq 0.4914 < 1/2$. This was regarded as surprisingly at the time, and not everyone would agree with the computation.

History in a nutshell

History in a nutshell

- Probability theory was often considered as a minor business in Mathematics in the coming centuries. Somehow it concerned more the gambling and drinking which went on in the evening in infamous bars and casinos, than the sophisticated atmosphere of the European academia.
- Nevertheless progress was made by the likes of Pascal, Fermat, J.Bernoulli, deMoivre, Huygens, Gauss and Laplace, who first wrote about probability as a mathematical problem.
- The Russian school gave a strong contribution starting in the XIX century. We can mention Chebyshev, Markov and Kolmogorov, who made the modern foundation of Probability Theory in a book published in 1933.

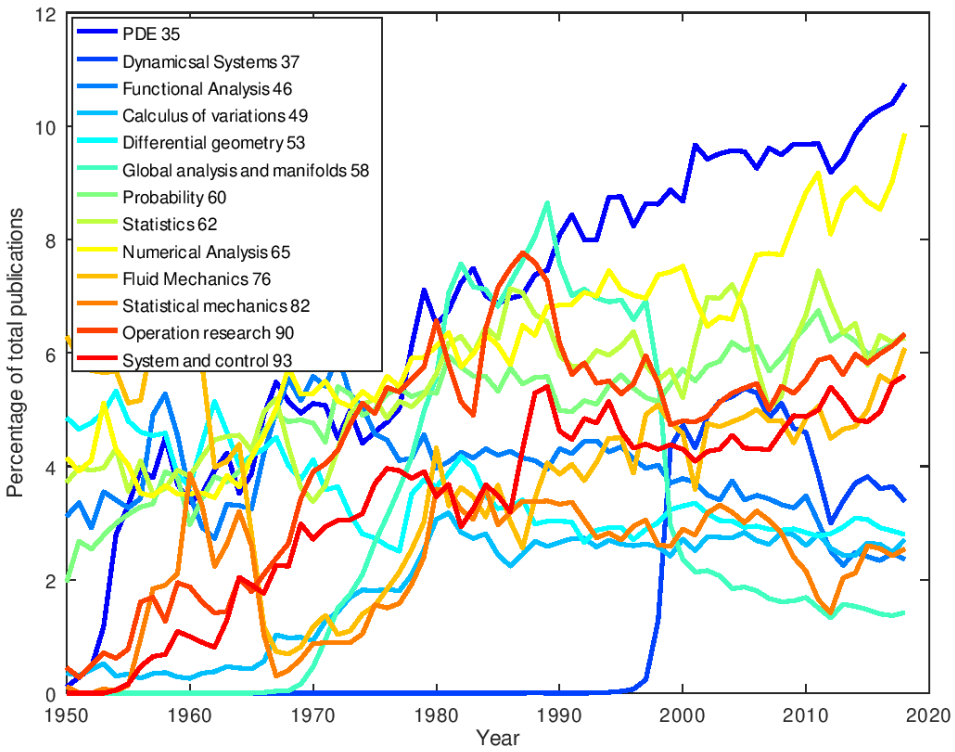
History in a nutshell

History in a nutshell

- A full academic recognition of Probability Theory is relatively recent. All the Fields medalists (Wendelin Werner (2006), Stanislav Smirnov (2010) and Martin Hairer (2014)) and Abel Prize winners (S.R.S. Varadhan (2007) and Yakov Sinai (2015)) who worked on probability, have been awarded in the XXI century.
- We live in a time of explosion of probabilistic applications in Pure and Applied Mathematics. It is one of the most popular and fastest growing mathematical fields both for undergraduate and graduate studies worldwide.
- No matter if you will end up studying Ergodic Theory, Number Theory, Theoretical Physics, Numerics, Mathematical Finance, Quantitative Biology, Data Science. You will need Probability.

Modern outlook

The graph shows the trend of mathematical publications. Probability and Statistics are the fastest growing, after Numerical Analysis of differential models. The bulk of the publications in applied probability (Finance, Data Science) are not counted here since they appear on non-mathematical journals.



What is probability?

What is probability? It is not a simple question, this is just a word in the language (in many languages indeed), which references to some more or less mathematical idea. Take the XVII century example of the dice before (24 pairs of dice etc). We cannot prove or disprove mathematically that the computation is right. We have to make a mathematical model for the game, and then compute within the model. But the computation is trivial, and the answer is already contained in the model itself, that we chose in first place. Thus it is a meta-mathematical problem, no surprise that smart people did not agree at the time.

Of course, modern Mathematics puts emphasis on the formal nature of objects, so rather then answering what is probability (an overwhelming task!), we in some sense just try to understand its properties and how to use it. Do not hope that, by the end of the class, you know what is probability however. For instance, I certainly do not know.

What is probability?

This is not just an empty discussion. As you will soon discover, the exercises of the first part of the course are somehow trivial from a mathematical point of view. The hardest part, is to translate our language, our intuition, in a mathematical model where we can simply compute everything. The math however will be very simple. These kind of classical exercises in Probability are extremely useful for mathematicians: in one step, we learn to formalize our intuition, but also to make effective models for simple real life phenomena.

(Almost) measure theory

You had an introduction to measure theory. As you may have heard, a probability is just a positive measure of total mass 1. True. But probability theory is not measure theory. It is rather the study of (some) measure-theoretic objects that are independent of the choice of the measure space. As Terence Tao put it

> *Probability theory is only allowed to study concepts and perform operations which are preserved with respect to extension of the underlying sample space. (This is analogous to how differential geometry is only allowed to study concepts and perform operations that are preserved with respect to coordinate change, or how graph theory is only allowed to study concepts and perform operations that are preserved with respect to relabeling of the vertices, etc).*

(Almost) measure theory

As long as one is adhering strictly to this dogma, one can insert as many new sources of randomness (or reorganise existing sources of randomness) as one pleases; but if one deviates from this dogma and uses specific properties of a single sample space, then one has left the category of probability theory and must now take care when doing any subsequent operation that could alter that sample space. This dogma is an important aspect of the probabilistic way of thinking, much as the insistence on studying concepts and performing operations that are invariant with respect to coordinate changes or other symmetries is an important aspect of the modern geometric way of thinking. With this probabilistic viewpoint, we shall soon see the sample space essentially disappear from view altogether, after a few foundational issues are dispensed with.

Naive Probability space

No matter what probability is, we certainly want to speak about probability of events and the likes. So certainly we want to have a nonempty set Ω, sometimes called the *sample space*, and a function $\mathbb{P}$ that associates to subsets A of Ω, their probability $\mathbb{P}(A)$, to be interpreted as the probability that the event A happens.
Of course, if A and B are two disjoint sets, we naturally expect that the probability that A or B happens, is just the sum of the probability of A plus the probability of B (e.g. the probability that you get 1 or 2 when tossing a die, is the sum of the probability of 1 plus that of 2).

Definition (THAT WE WILL NEVER USE)

A *naive probability space* is a couple $(\Omega, \mathbb{P})$ where Ω is a nonempty set and $\mathbb{P}\colon \mathbf{2}^{\Omega} \to [0,1]$ is a map from the power set of Ω to $[0,1]$ such that

- $\mathbb{P}(\Omega) = 1$.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$, for all $A, B \subset \Omega$ s.t. $A \cap B = \emptyset$.

Naive Probability space

Unfortunately this is a naive approach.

- On the one hand, in math we want often to discuss limits and continuous object, so we would like to replace the second condition with a countable (=счётное) union of sets and the right hand side with an infinite series (ряд).
- However, as you have maybe seen in measure theory, in this case the requirement that $\mathbb{P}$ can be evaluated on *any* subset of Ω is very strong, and not useful. See the Banach-Tarski paradox `https://bit.ly/2Z3L8px`

As you probably know from measure theory, a workaround (not the only one however) is to have $\mathbb{P}$ defined only on a sub-family of subsets of Ω, with some natural properties, a σ-algebra.
This is where our mathematical journey begins.

σ-algebras

Definition (σ-algebra)

If Ω is a non-empty set, a *σ-algebra* on Ω is a subset $\mathfrak{F}$ of the power set (power set =булеан) of Ω such that

(a) $\emptyset \in \mathfrak{F}$.

(b) If $A \in \mathfrak{F}$, its complement (=разность) is in $\mathfrak{F}$, namely $A \in \mathfrak{F}$ implies $A^c \in \mathfrak{F}$.

(c) If $(A_i)_{i \geq 0}$ is a countable family of sets in $\mathcal{F}$, then $\cup_{i \geq 0} A_i \in \mathfrak{F}$.

Definition (Measurable space)

A *measurable space* (=измеримое пространство) is a couple $(\Omega, \mathfrak{F})$ where Ω is a non-empty set and $\mathfrak{F}$ is a σ-algebra on Ω. In the context of probability and statistic, Ω is called the *sample space*, while elements of $\mathfrak{F}$ (measurable subsets of Ω) are called *events*.

Probability space

Definition (Probability space)

A *probability space* (=вероятностное пространство) is a triple $(\Omega, \mathfrak{F}; \mathbb{P})$ where $(\Omega, \mathfrak{F})$ is measurable space, and $\mathbb{P}\colon \mathfrak{F} \to [0,1]$ is such that

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

(b) If $(A_i)_{i\geq 0}$ is a countable family of events, that are pairwise disjoint, that is $A_i \cap A_j = \emptyset$ for $i \neq j$ then

$$\mathbb{P}(\cup_{i\geq 0} A_i) = \sum_{i\geq 0} \mathbb{P}(A_i) \tag{1}$$

Property (a) is called *normalization* while (b) σ-*additivity*.
Let us give some examples.

Examples

Example

If we want to model the throw of a die (=кубик), we can choose $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathfrak{F} = \mathbf{2}^{\Omega}$ (the power set of Ω), and $\mathbb{P}$ to be defined as $\mathbb{P}(A) = |A|/6$ where $|A|$ is the cardinality (number of elements) of A.
For instance if we want to compute the probability to get an even number when tossing a die, we can define $A = \{2, 4, 6\}$, and clearly $\mathbb{P}(A) = 3/6 = 1/2$.

Example

We measure the length of a table with a rules (линейка) that is graded by centimeters, so we see that the length is within 88 and 89 centimeters. We model the exact length of the table randomly. A reasonable model is to take $\Omega = [88, 89]$, $\mathfrak{F}$ to be the Borel (or Lebesgue if you prefer) σ-algebra of the interval $[88, 89]$ and $\mathbb{P}$ as the unique probability on this space satisfying $\mathbb{P}([a, b)] = b - a$ for all $88 \leq a \leq b \leq 89$.

Examples

Example

If Ω is a finite set, one can always define a probability on Ω, by taking $\mathfrak{F}$ the power set of Ω and $\mathbb{P}(A) = |A|/|\Omega|$. It is immediate to check that this satisfies (a) and (b) in the definition of a probability space. Of course, this is not the only probability that one can define on Ω (apart the trivial case where Ω consists of one point).

Example

We want to model the flip of two coins, each of them giving either head or tail (орёл или решка). Then we can take $\Omega = \{H, T\} \times \{H, T\}$ and again $\mathbb{P}(A) = |A|/4$. However, if we are only interested in the number of heads, we can also simply take $\Omega = \{0, 1, 2\}$ and $\mathbb{P}(A) = \sum_{i \in A} p_i$ where p_i is defined by $p_0 = 1/4$, $p_1 = 1/2$, $p_2 = 1/4$.

Elementary properties

Proposition (Elementary properties of probabilities)

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space. Then it holds

(a) $\mathbb{P}(\emptyset) = 0$.

(b) *If $(A_i)_{1 \leq i \leq n}$ is a finite family of pairwise disjoint events, then* $\mathbb{P}(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i)$.

(c) *For any event A, it holds* $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(d) *If $A \subset B$, then* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

(e) $\mathbb{P}(\cup_i A_i) \leq \sum_i \mathbb{P}(A_i)$.

Proof:

(a) Consider the σ-additivity property of $\mathbb{P}$ with all the A_i equal to the empty set. Then

$$1 \geq \mathbb{P}(\emptyset) = \mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i) = \sum_i \mathbb{P}(\emptyset) \tag{2}$$

Since the last series is finite, $\mathbb{P}(\emptyset) = 0$.

Elementary properties

(b) Let us define $A_i = \emptyset$ for $i \geq n+1$. Then we can apply the σ-additivity and (a) to get

$$\mathbb{P}(\cup_{i=1}^{n} A_i) = \mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i) + 0 \quad (3)$$

(c) In (b), take $n = 2$ and $A_1 = A$ and $A_2 = A^c$.

(d) From (b) we get $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$.

(e) Define $B_0 = A_0$ and $B_{i+1} = A_{i+1} \setminus (\cup_{j \leq i} A_i)$. Then the B_i are pairwise disjoint, and $\cup_i A_i = \cup_i B_i$. So that

$$\mathbb{P}(\cup_i A_i) = \mathbb{P}(\cup_i B_i) = \sum_i \mathbb{P}(B_i) \leq \sum_i \mathbb{P}(A_i) \quad (4)$$

where in the last line we used (d) and the fact that $B_i \subset A_i$.

Continuity of the probability

If $(A_i)_{i\geq 0}$ is a sequence of events in Ω, one can define their liminf and limsup as

$$\begin{aligned} \underline{\lim}_i A_i &= \bigcup_{i\geq 0} \left(\bigcap_{n\geq i} A_n \right) \\ \overline{\lim}_i A_i &= \bigcap_{i\geq 0} \left(\bigcup_{n\geq i} A_n \right) \end{aligned} \tag{5}$$

Notice that the liminf and limsup are also events. $\underline{\lim}_i A_i$ is the set of all points in Ω that are *definitely* in all of the A_i, while the limsup is the set of all points in Ω that are in infinitely many of the A_i's. It is easy to see that $\overline{\lim}_i A_i \supset \underline{\lim}_i A_i$. If they coincide, we say that the sequence (A_i) admits a limit, $\lim_i A_i = \underline{\lim}_i A_i = \overline{\lim}_i A_i$.
For instance monotone sequences admit a limit. In particular if the A_i are increasing, then $\lim_i A_i = \cup_i A_i$, while if they are decreasing, $\lim_i A_i = \cap_i A_i$.

Continuity of the probability

Proposition (Continuity of the probability)

Let $(A_i)_{i\geq 0}$ be a sequence of events of the probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. Then it holds

$$\mathbb{P}(\underline{\lim}_i A_i) \leq \underline{\lim}_i \mathbb{P}(A_i) \leq \overline{\lim}_i \mathbb{P}(A_i) \leq \mathbb{P}(\overline{\lim}_i A_i) \tag{6}$$

In particular if the sequence (A_i) admits a limit, then

$$\mathbb{P}(\lim_i A_i) = \lim_i \mathbb{P}(A_i) \tag{7}$$

Notice then in all of the above formulas, the right hand sides are limit of real numbers.

Proof: Define $B_i = \cap_{n\geq i} A_n$, $C_0 = B_0$ and and $C_{i+1} = B_{i+1} \setminus B_i$ for $i \geq 0$. Then the sets C_i are pairwise disjoint (=попарно непересекающиеся) and yet $\cup_{i\geq 0} C_i = \cup_{i\geq 0} B_i = \underline{\lim}_i A_i$. Therefore by σ-additivity

Continuity of the probability

$$\begin{aligned}\mathbb{P}(\underline{\lim}_i A_i) &= \mathbb{P}(\cup_{i\geq 0} C_i) = \sum_i \mathbb{P}(C_i) = \lim_n \sum_{i=0}^{n} \mathbb{P}(C_i) \\ &= \lim_n \mathbb{P}(\cup_{i=0}^{n} C_i) = \lim_n \mathbb{P}(B_n) = \sup_n \mathbb{P}(B_n) \\ &\leq \sup_n \inf_{i\geq n} \mathbb{P}(A_i) = \underline{\lim}_i \mathbb{P}(A_i)\end{aligned} \tag{8}$$

where in the inequality we used that $B_n \subset A_i$ for $i \geq n$. This proves the first inequality.
The second inequality is trivial. The third inequality is proved as the first one, or simply one can notice that $\overline{\lim}_i A_i = (\underline{\lim}_i A_i^c)^c$. So that by the inequality we just proved

$$\mathbb{P}(\overline{\lim_i} A_i) = 1 - \mathbb{P}(\underline{\lim_i} A_i^c) \geq 1 - \underline{\lim_i} \mathbb{P}(A_i^c) = \overline{\lim_i} \mathbb{P}(A_i) \tag{9}$$

□

Continuity of the probability

Corollary

If A_i is a decreasing sequence such that $\cap_i A_i = \emptyset$, then $\lim_i \mathbb{P}(A_i) = 0$.

Corollary

If $\Omega = \mathbb{R}$ and $\mathfrak{F}$ is the Borel σ-algebra, then any probability measure $\mathbb{P}$ on $\mathbb{R}$ satisfies $\lim_{L\to\infty} \mathbb{P}([L, \infty)) = \lim_{L\to\infty} \mathbb{P}((-\infty, -L]) = 0$.

Corollary (Borel-Cantelli Lemma)

If $(A_i)_{i\geq 0}$ is a sequence of events such that $\sum_{i\geq 0} \mathbb{P}(A_i) < \infty$, then it holds $\mathbb{P}(\overline{\lim}_i A_i) = 0$, namely the probability that infinitely many of the A_i happen simultaneously is 0.

Proof: By the continuity of probabilities, we have that

$$\mathbb{P}(\overline{\lim_i} A_i) = \lim_i \mathbb{P}(\cup_{n=i}^{\infty} A_n) \leq \lim_i \sum_{n=i}^{\infty} \mathbb{P}(A_n) = 0 \qquad (10)$$

which vanishes as the remainder of a converging series. □

Carathéodory extension theorem

A collection $\mathcal{S}$ of sets $A \subset \Omega$ is called a *semiring* if

- $\emptyset \in \mathcal{S}$.
- If $A, B \in \mathcal{S}$ then $A \cap B \in \mathcal{S}$.
- If $A, B \in \mathcal{S}$, then there exist $K_1, \ldots, K_n \in \mathcal{S}$ pairwise disjoint such that $A \setminus B = \cup_{i=1}^{n} K_i$.

For instance intervals of the form $[a, b)$ are a semiring in $\mathbb{R}$.

Theorem (Carathéodory, see Kenke Th.1.53)

Let $\mathcal{S}$ be a semiring and let $\tilde{\mu}\colon \mathcal{S} \to [0, \infty]$ be additive, σ-subadditive and σ-finite $\tilde{\mu}(\emptyset) = 0$. There there is a unique σ-finite measure $\mu\colon \sigma(\mathcal{S}) \to [0, \infty]$ s.t. $\mu(A) = \tilde{\mu}(A)$, for all $A \in \mathcal{S}$.

Discrete spaces

For discrete spaces, the complications of σ-algebras and measurability are largely simplified, and usually problems reduce to combinatorics.

Definition (Discrete density)

A map $p\colon \Omega \to [0,1]$ is called a *discrete density* (=дискретная плотность) on Ω if $\sum_{x\in\Omega} p(x) = 1$. Notice in particular that $p(x) > 0$ only for a finite or countable number of x.

Definition (Discrete probability space)

A probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ is *discrete* if there exists a finite or countable event $E \subset \Omega$ such that $\mathbb{P}(E) = 1$.

All the previous examples are discrete, except the one of the length of the table (where the probability of each point is 0).

Discrete spaces

There is a bijection (=биекция) between discrete densities and probability measures on a set Ω (equipped with a σ-algebra that contains singletons). The bijection is given by

$$\begin{aligned} p(x) &= \mathbb{P}(\{x\}) \\ \mathbb{P}(A) &= \sum_{x \in A} p(x) \end{aligned} \tag{11}$$

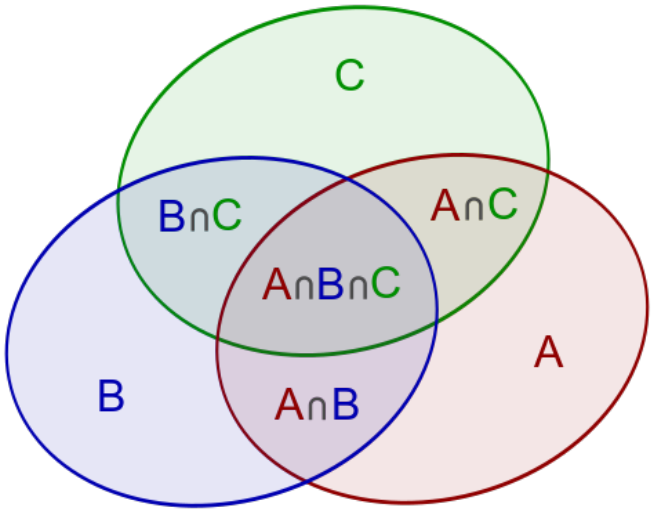
Probabilistic Inclusion-Exclusion

If we have two disjoint sets A and B, of course $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. If A and B are not disjoint, then $\mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)$ and $\mathbb{P}(B) = \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B)$. Therefore since $A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)$ we get

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \tag{12}$$

Similarly for three events

$$\begin{aligned} \mathbb{P}(A \cup B \cup C) =& \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) \\ & - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C) \end{aligned} \tag{13}$$



Probabilistic Inclusion-Exclusion

This generalizes to many sets a follows.

Proposition (Probabilistic Inclusion-Exclusion principle)

Given the events, $A_1, A_2, \ldots, A_n$*, it holds*

$$\mathbb{P}(\cup_{i=1}^{n} A_i) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{\substack{(i_1, i_2, \ldots, i_r): \\ 1 \leq i_1 < i_2 < \cdots < i_r \leq n}} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r}) \tag{14}$$

Inclusion-Exclusion

It is not hard to prove the proposition by induction, however here we prove a more general result, and then deduce the probabilistic Inclusion-Exclusion as a corollary.

Theorem (Inclusion-Exclusion)

Let S be a finite nonempty set. Let $f, g : 2^S \to \mathbb{R}$. Then it holds

$$f(E) = \sum_{F\,:\,F \subset E} g(F) \qquad \forall E \subset S \tag{15}$$

iff (=тогда и только тогда)

$$g(E) = \sum_{F\,:\,F \subset E} (-1)^{|E|-|F|} f(F) \qquad \forall E \subset F \tag{16}$$

Inclusion-Exclusion

Proof: For $E \subset S$ and $F \subset E$ denote

$$h(E,F) := \sum_{A\,:\,F\subset A\subset E} (-1)^{|E|-|A|}$$

If $|E| - |F| = |E \setminus F| = n \geq 1$, $E \setminus F$ has $\binom{n}{k}$ subsets of cardinality k. In the sum defining $h(E,F)$, there are thus $\binom{n}{k}$ terms written as $(-1)^k$. Or, for $n \geq 1$

$$h(E,F) = \sum_{k=0}^{n}(-1)^k\binom{n}{k} = \sum_{k=0}^{n}\binom{n}{k}1^{n-k}(-1)^k = (1+(-1))^n = 0$$

However for $E = F$ (that is $n = 0$), $h(E,F) = 1$.
We first show that (15) implies (16). For any $E \subset S$

$$\sum_{F\,:\,F\subset E} (-1)^{|E|-|F|}f(F) = \sum_{F\,:\,F\subset E} (-1)^{|E|-|F|} \sum_{A\,:\,A\subset F} g(A)$$

$$= \sum_{A\,:\,A\subset E} g(A) \sum_{F\,:\,A\subset F\subset E} (-1)^{|E|-|F|} = \sum_{A\,:\,A\subset E} g(A)h(E,A) = g(E)$$

Inclusion-Exclusion

Let us show that (16) implies (15). Per $E \subset S$, reasoning as above

$$\begin{aligned}\sum_{F\,:\,F\subset E} g(F) &= \sum_{F\,:\,F\subset E}\sum_{A\,:\,A\subset F} (-1)^{|F|-|A|} f(A)\\ &= \sum_{A\,:\,A\subset E} f(A) \sum_{F\,:\,A\subset F\subset E} (-1)^{|F|-|A|}\\ &= \sum_{A\,:\,A\subset E} f(A)h(E,A) = f(E)\end{aligned}$$

□

Probabilistic Inclusion-Exclusion

Proof of the p.I.E.: For $I \subset \{1, \ldots, n\}$, define

$$F_I := (\cap_{j \notin I} A_j) \bigcap (\cap_{i \in I} A_i^c)) \subset \Omega$$

Notice $F_{\{1,\ldots,n\}} = \emptyset$ and $F_I \cap F_{I'} = \emptyset$ for $I \neq I'$. Let then $f, g : \mathbf{2}^{\{1,2,\ldots,n\}} \to \mathbb{R}$ be defined as

$$g(I) := \mathbb{P}(F_I), \qquad f(I) := \sum_{J \subset I} g(J)$$

Since the F_I are pairwise disjoint

$$f(I) = \begin{cases} \mathbb{P}(\cap_{i \notin I} A_i) & \text{if } I \subsetneq \{1, \ldots, n\} \\ \mathbb{P}(\cup_{i=1}^n A_i) & \text{if } I = \{1, \ldots, n\} \end{cases}$$

Using (16)

$$\begin{aligned} 0 = & g(\{1, \ldots, n\}) = \sum_{I \subset \{1,\ldots,n\}} (-1)^{n-|I|} f(I) \\ = & \mathbb{P}(\cup_{i=1}^n A_i) - \sum_{I \subsetneq \{1,\ldots,n\}} (-1)^{n-|I|} \mathbb{P}(\cap_{i \notin I} A_i) \end{aligned} \tag{17}$$

Probabilistic Inclusion-Exclusion

Example

n people go to the Bolshoi theater, and leave their umbrellas at the wardrobe. When departing, the umbrellas are mixed up and each of them collects a random umbrella. What is the probability that at least one man gets his own umbrella? Calculate the limit of this probability as $n \to \infty$.

Let us number the man and the umbrellas from 1 to n, so that the assignment of the umbrellas is just a permutation on n elements. We take as probability space $\Omega = S_n$ the set permutations on n elements, $|\Omega| = n!$, with the discrete σ-algebra. We are thinking that each permutation is equally likely, so $\mathbb{P}$ is just uniform $\mathbb{P}(A) = |A|/n!$. We are asked the probability of the event $A := \{\pi \in \Omega : \exists i : \pi(i) = i\}$, (the set of permutations with at least one fixed point).

Probabilistic Inclusion-Exclusion

Define $A_i := \{\pi \in \Omega : \pi(i) = i\}$, the set of permutations that fix i. Then $\mathbb{P}(A_i) = (n-1)!/n! = 1/n$, $\mathbb{P}(A_i \cap A_j) = (n-2)!/n!$ and in general if we intersect r distinct sets A_i, we have a probability $(n-r)!/n!$. On the other hand $A = \cup_i A_i$ so that from (14)

$$\mathbb{P}(A) = \sum_{r=1}^{n} (-1)^{r+1} \binom{n}{r} \frac{(n-r)!}{n!} = \sum_{r=1}^{n} (-1)^{r+1}/r! \qquad (18)$$

This converges to e^{-1} as $n \to \infty$.

Observations

- We should imagine σ-algebra as the mathematical way to formalize the idea of observability and measurability. Let us consider a model for throwing a die and a coin. So our space is $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{0, 1\}$ with the discrete σ-algebra $\mathfrak{F} = \mathbf{2}^{\Omega}$. Suppose now that one person can only observe the result of the die, and a second person the result of the coin. The σ algebra associated to the first person is formed by sets of the form $A \times \{0, 1\}$ for any $A \subset \{1, 2, 3, 4, 5, 6\}$. The σ-algebra associated to the second person is formed by sets of the form $\{1, 2, 3, 4, 5, 6\} \times A$ with $A \subset \{0, 1\}$.
- In this case, we understand that the observing the result of the die, will not influence the probabilities of the possible results of the coin. Let us try to formalize this idea.

The meaning of it

Hereafter $(\Omega, \mathfrak{F}, \mathbb{P})$ is a probability space.

Definition

Two events A and B are *independent* (or more precisely $\mathbb{P}$-independent) if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

Definition

Two sub-σ-algebras $\mathfrak{F}_1, \mathfrak{F}_2 \subset \mathfrak{F}$ are called independent if every $A \in \mathfrak{F}_1$ and $B \in \mathfrak{F}_2$ are independent.

Remark

If A and B are independent, then A^c and B are also independent, as

$$\begin{aligned} \mathbb{P}(A^c \cap B) &= \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) \\ &= \mathbb{P}(B)(1 - \mathbb{P}(A)) = \mathbb{P}(A^c)\mathbb{P}(B) \end{aligned} \tag{19}$$

Independence

- What does it mean for two events to be independent? By the definition, it says that the likelihood of $A \cap B$ within B is the same of the likelihood of A in Ω. So it states that if for instance the event B happens, this does not influence the probability that A happens.
- If A is an event, the smallest σ-algebra containing A is $\mathfrak{F}_A = \{\emptyset, A, A^c, \Omega\}$. This corresponds to the observer that only knows whether A happened or not. Then, thanks to the previous remark, A and B are independent iff $\mathfrak{F}_A$ and $\mathfrak{F}_B$ are independent.
- Very often, asking for independence is a meta-mathematical choice within the model. If one says, let us throw two dice, one silently assumes their results to be independent. People betting on the lottery (weekly extraction of numbers) may disagree as they create strategies to bet on numbers depending on the results of the past. Of course, we cannot mathematically disprove them, we can just train to create better models.

Independence

Definition

An arbitrary family $(\mathfrak{F}_\alpha)_{\alpha\in\mathcal{A}}$ of σ-algebras $\mathfrak{F}_\alpha \subset \mathfrak{F}$ is called independent if for every $n \geq 1$ and every $\alpha_1, \ldots, \alpha_n \in \mathcal{A}$ it holds

$$\mathbb{P}(\cap_{i=1}^n A_i) = \prod_{i=1}^n \mathbb{P}(A_i) \qquad \text{for every } A_1 \in \mathfrak{F}_{\alpha_1} \ldots A_n \in \mathfrak{F}_{\alpha_n} \quad (20)$$

The events $(A_\alpha)_{\alpha\in\mathcal{A}}$ are independent if the associated σ-algebras are independent, namely if for every $n \geq 1$ and every $\alpha_1, \ldots, \alpha_n \in \mathcal{A}$

$$\mathbb{P}(\cap_{i=1}^n A_{\alpha_i}) = \prod_{i=1}^n \mathbb{P}(A_{\alpha_i}) \quad (21)$$

So (20) is the general definition of independence, all the others can be derived from that one.

Independence:example

Example

$n+1$ knights are drinking seated in a line. Each of them chooses (independently from the other) to drink either ale or mead with probability $p \in (0,1)$. For $i = 1, \ldots, n$ let A_i be the event where the knight i and the knight $i+1$ at his right side chose the same drink. We claim that the events $(A_i)_{i=1,\ldots,n}$ are independent iff $p = 1/2$. Indeed let us define B_i the event on which the knight i chose ale. The events B_i are independent as it is stated in the example (we assume that everyone made his choice without looking at their neighbors).
Then $A_i = (B_i \cap B_{i+1}) \cup (B_i^c \cap B_{i+1}^c)$. And $\mathbb{P}(A_i) = p^2 + (1-p)^2$. Similarly $\mathbb{P}(A_i \cap A_{i+1}) = p^3 + (1-p)^3 \geq (p^2 + (1-p)^2)^2$. The equality holds iff $p = 0, 1/2, 1$, (we excluded however the trivial cases 0 and 1). So certainly the events are dependent if $p \neq 1/2$.

Independence:example

Example (Continued)

For $p = 1/2$ we need to check that $P(A_{i_1} \cap \ldots \cap A_{i_k}) = 2^{-k}$ for every $1 \leq i_1 < i_2 < \ldots i_k \leq n$. Denote $A_i^+ = A_i$ and $A_i^- = A_i^c$. Then it is easily seen

$$\mathbb{P}(A_1^{\pm} \cap A_2^{\pm} \ldots \cap A_n^{\pm}) = 2^{-n} = \prod_{i=1}^{n} \mathbb{P}(A_i^{\pm}) \tag{22}$$

From this, we easily deduce independence as for instance

$$\begin{aligned} \mathbb{P}(A_1 \cap A_2 \ldots \cap A_{n-1}) &= \mathbb{P}(A_1 \cap A_2 \ldots \cap A_n) \\ &+ \mathbb{P}(A_1 \cap A_2 \ldots \cap A_{n-1} \cap A_n^c) = 2^{-n} + 2^{-n} = 2^{-n+1} \end{aligned} \tag{23}$$

Independence:example

Example

n knights are drinking seated at the round King's Arthur table. Each of them chooses (independently from the other) to drink either ale or mead with probability $p \in (0,1)$. For $i = 1, \ldots, n$ let A_i be the event where the knight i and the knight $i+1$ at his right side chose the same drink.
We claim that the events $(A_i)_{i=1,\ldots,n}$ are not independent no matter $p \in (0,1)$. Indeed, in this case, any sub-family of $n-1$ events will be independent reasoning as in the previous example, however

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = 2^{-n+1} \tag{24}$$

as this is just the probability that all the knights chose the same drink.

Conditioning

Definition

If B is an event with $\mathbb{P}(B) > 0$ we denote

$$\mathbb{P}(A|B) := \mathbb{P}(A \cap B)/\mathbb{P}(B) \tag{25}$$

the probability of an event A conditioned to B.

This is more than simple arithmetic , and like independence it will extend to σ-algebras and events B with $\mathbb{P}(B) = 0$. However, we will see it later in the lectures.

Proposition

For a given B with $\mathbb{P}(B) > 0$, define $\mathbb{P}_B : \mathfrak{F} \to [0,1]$ by $\mathbb{P}_B(A) = \mathbb{P}(A|B)$. Then $\mathbb{P}_B$ is a probability.

The proof is elementary (e.g. $\mathbb{P}_B(\Omega) = \mathbb{P}(\Omega \cap B)/\mathbb{P}(B) = 1$ etc). $\mathbb{P}(\cdot|B)$ clearly can be interpreted as the probability once we know that B happened.

Conditioning:example

Example

A player goes to a casino, to play the roulette. In this game, the player has a probability $p = 1/37$ to win at every turn. However, the player has a small suspicion, say $q = 1/100$, that the casino cheats so that the player never wins (however he think that with probability 99/100 the casino is fair).
After $n = 200$ turns, he never won. How can he know estimate the probability that the casino is fair?
Let us call A the event where the casino is fair, and B the event where the player never wins after $n = 200$ tries. We are looking for $\mathbb{P}(A|B)$. We know however that $\mathbb{P}(A) = (1 - q)$,
$\mathbb{P}(B|A) = (1 - p)^n$, and $\mathbb{P}(B|A^c) = 1$. Therefore, the first equality being immediate, we have

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A)+\mathbb{P}(B|A^c)\mathbb{P}(A^c)} = \frac{(1-p)^n(1-q)}{(1-p)^n(1-q)+q} \simeq 0.29 \quad (26)$$

This goes to 0 exponentially fast in n!

Bayes Formula

We can generalize the calculation in the previous example.

Theorem

Let $(A_i)_{i\in\mathcal{I}}$ be a finite or countable partition (разбиение) of Ω in events A_i. Then for every event B with $\mathbb{P}(B) > 0$ and $j \in \mathcal{I}$

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i)} \tag{27}$$

Доказательство.

Since the A_i are a partition, it holds that $B = \cup_{i\in I} B \cap A_i$, the union being disjoint. Therefore

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i)} \tag{28}$$

provided we understand $\mathbb{P}(B|A_i)\mathbb{P}(A_i) = 0$ whenever $\mathbb{P}(A_i) = 0$. □

Bayes Formula

Bayes formula is a basic tool in Statistics experimental Sciences.

Example

Anna tosses a green die. Then she tosses as many red dice as the result of the green die. We know that the sum of the results of the red dice is 5. What is the probability that the result of the green die was 1?

Let B be the event where the sum of the red dice is 5. Let A_i be the event where the green die is i. For instance we can take $\Omega = \{1,2,3,4,5,6\}^7$, with the first entry corresponding to the green die. Then $A_i : \{\omega : \omega_0 = i\}$ and $B = \{\omega : \sum_{k=1}^{\omega_0} \omega_k = 5\}$. The problem asks for $\mathbb{P}(A_1|B)$. We can calculate $\mathbb{P}(A_i) = 1/6$ and $\mathbb{P}(B|A_1) = 1/6$, $\mathbb{P}(B|A_2) = 4/6^2$, $\mathbb{P}(B|A_3) = 6/6^3$, $\mathbb{P}(B|A_4) = 4/6^4$, $\mathbb{P}(B|A_5) = 1/6^5$ and $\mathbb{P}(B|A_6) = 0$. Then

$$\mathbb{P}(A_1|B) = \frac{\frac{1}{6}\frac{1}{6}}{\frac{1}{6}(1/6 + 4/6^2 + 6/6^3 + 4/6^4 + 1/6^5)} \simeq 0.54 \qquad (29)$$

Products

So far we have been sloppy about existence of probability spaces, because we dealt with finite spaces. However, it is often useful in probability to define some random objects, and then add some more objects independent from the first one.

Definition

Let $(\Omega_\alpha, \mathfrak{F}_\alpha)$ be a family of measurable spaces indexed by $\alpha \in \mathcal{A}$ (any set $\mathcal{A}$). Let $\Omega := \prod_\alpha \Omega_\alpha$ and let $\pi_\alpha : \Omega \to \Omega_\alpha$ be the coordinate projection. Define $\mathfrak{F}$ as the smallest σ-algebra such that all the π_α are measurable. Then $(\Omega, \mathfrak{F})$ is called the product space of the $(\Omega_\alpha, \mathfrak{F}_\alpha)$.

If for every $\alpha \in \mathcal{A}$ we also have a probability measure $\mathbb{P}_\alpha$, we would also like to define a product probability $\mathbb{P} = \prod_\alpha \mathbb{P}_\alpha$. This is done as follows.

Products

First define for any n, any $\alpha_1, \ldots, \alpha_n \in \mathcal{A}$, and any $A_1 \in \mathfrak{F}_{\alpha_1}, \ldots, A_n \in \mathfrak{F}_{\alpha_n}$

$$\mathbb{P}(\cap_{i=1}^{n} \pi_{\alpha_i}^{-1}(A_i)) = \prod_{i=1}^{n} \mathbb{P}_{\alpha_i}(A_i) \tag{30}$$

Since sets of the form $\cap_{i=1}^{n} \pi_{\alpha_i}^{-1}(A_i)$ are a semi-ring, by Carathéodory extension theorem, $\mathbb{P}$ extends uniquely to a probability measure on $\mathfrak{F}$.
$(\Omega, \mathfrak{F}, \mathbb{P})$ is called the product of the $(\Omega_\alpha, \mathfrak{F}_\alpha)$.

Example

If we take $[0,1]$ equipped with the Borel σ-algebra $\mathfrak{B}$ and usual flat measure $\lambda = dx$, then the product of $([0,1], \mathfrak{B}, \lambda)$ with itself coincides with $[0,1]^2$ equipped with the Borel σ-algebra and the flat measure $dx\, dy$.

Lovasz lemma

Suppose that we have some bad event $A_1, \ldots, A_n$ that happen with some small probability, say $\mathbb{P}(A_i) \leq p < 1$. Of course $\mathbb{P}(\cup_i A_i) \leq np$, so that if $p < 1/n$ there is a positive probability that no bad event occurs as $\mathbb{P}(\cap_i A_i^c) \geq 1 - np > 0$.
On the other hand, if we also know that the A_i are independent, then of course $\mathbb{P}(\cap_i A_i^c) = (1-p)^n > 0$, regardless of $p < 1$.
One can say something nontrivial in the case where the A_i are not independent, but say each A_i depends on at most d of the other sets.

Theorem

Let $A_1, \ldots, A_n$ be events with $\mathbb{P}(A_i) \leq p$, and such that each A_i is independent of all the other events except for at most d of them. If

$$p < \begin{cases} 1/2 & \text{if } d = 1 \\ \frac{(d-1)^{d-1}}{d^d} & \text{if } d \geq 2 \end{cases} \tag{31}$$

Then $\mathbb{P}(\cap_{i=1}^n A_i^c) > 0$.

Lovasz lemma:proof

We prove a weaker statement assuming the slightly stronger condition $p < \frac{d^d}{(d+1)^{d+1}}$.

- Notice that

$$\mathbb{P}(\cap_i A_i^c) = \mathbb{P}(A_1^c | A_2^c \dots \cap A_n^c)\, \mathbb{P}(A_2^c | A_3^c \dots \cap A_n^c) \dots \mathbb{P}(A_n^c) \qquad (32)$$

- We claim that if $p < \frac{d^d}{(d+1)^{d+1}}$, then for every $I \subset \{1, \dots, n\}$ and every $j \notin I$, it holds

$$\mathbb{P}(A_j | \cap_{i \in I} A_i^c) \leq \frac{1}{d+1} \qquad (33)$$

  Once we prove (33) we conclude because all the terms in (32) are then strictly positive.

- Let us then prove (33) by induction on the cardinality of I. If I is empty, then $\mathbb{P}(A_j) \leq p < 1/(d+1)$.

Lovasz lemma:proof

- Let us prove the inductive step on the cardinality of I. Given j let I_j be the set of indexes i such that A_i is in the dependence cluster of A_j, so that $|I_j| \leq d$. (33) can be written as

$$\mathbb{P}(A_j | \cap_{i \in I} A_i^c) = \frac{\mathbb{P}(A_j \cap (\cap_{i \in I \cap I_j} A_i^c) | \cap_{i \in I \cap I_j^c} A_i^c))}{\mathbb{P}(\cap_{i \in I \cap I_j} A_i^c) | \cap_{i \in I \cap I_j^c} A_i^c))} \tag{34}$$

The numerator in (34) is bounded as

$$\begin{aligned} \mathbb{P}(A_j \cap (\cap_{i \in I \cap I_j} A_i^c) | \cap_{i \in I \cap I_j^c} A_i^c) &\leq \mathbb{P}(A_j | \cap_{i \in I \cap I_j^c} A_i^c)) \\ &= \mathbb{P}(A_j) \leq p \end{aligned} \tag{35}$$

where we used the fact that A_j is independent of the A_i for $i \notin I_j$.

Lovasz lemma:proof

As for the denominator, write $I \cap I_j = \{i_1, \ldots, i_r\}$. Then the denominator equals

$$\prod_{j=1}^{r} \mathbb{P}(A_{i_j}^c | (\cap_{k=j+1}^{r} A_{i_k}^c) \cap (\cap_{i \in I \cap I_j^c} A_i^c)) \geq \prod_{j=1}^{r} (1 - \tfrac{1}{d+1}) \geq (1 - \tfrac{1}{d+1})^d \tag{36}$$

where in the first inequality we used the induction hypotheses, and in the second the fact that $r = |I \cap I_j| \leq |I_j| \leq d$. Using (35) and (36), the probability in (33) is bounded as

$$\mathbb{P}(A_j | \cap_{i \in I} A_i^c) \leq \frac{p}{(1 - \frac{1}{d+1})^d} \leq \frac{1}{d+1} \tag{37}$$

where in the last inequality we used $p \leq \frac{d^d}{(d+1)^{d+1}}$.

Random Variables

Definition

If $(E, \mathfrak{E})$ is a measurable space, a measurable map $X : \Omega \to E$ is called an E-valued random variable. If $E = \mathbb{R}$ ad $\mathfrak{E}$ is the Borel σ-algebra, X is simply called a random variable, or real random variable.

Recall that *measurable* here means that $X^{-1}(A) \in \mathfrak{F}$ for every $A \in \mathfrak{E}$.

It is easy to check that if X is a random variable, then $\mu := \mathbb{P} \circ X^{-1}$ is a probability measure on $(E, \mathfrak{E})$. Indeed:

$$\begin{aligned} \mu(E) &= \mathbb{P}(X^{-1}(E)) = \mathbb{P}(\Omega) = 1 \\ \mu(\cup_i A_i) &= \mathbb{P}(X^{-1}(\cup_i A_i)) = \mathbb{P}(\cup_i X^{-1}(A_i)) \\ &= \sum_i \mathbb{P}(X^{-1}(A_i)) = \sum_i \mu(A_i) \end{aligned} \tag{38}$$

Random variables

Definition

If X an E-valued random variable, *the law of* X is the probability measure $\mu := \mathbb{P} \circ X^{-1}$. μ is also called the pushforward (образ) of $\mathbb{P}$ via μ.

Often one is interested in the law of random variables, rather than their definition on a probability space. To this aim, the following notation is always used

$$\mathbb{P}(X \in A) \quad \text{means} \quad \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}) \tag{39}$$

and more in general in the literature

$$\begin{aligned} &\mathbb{P}(\text{property (P) holds}) \qquad \text{means} \\ &\mathbb{P}(\{\omega \in \Omega : \text{property (P)}(\omega) \text{ holds}\}) \end{aligned} \tag{40}$$

Example

We flip a coin and we win one ruble if we get head. Let X be the random variable corresponding to our win. Then $\mathbb{P}(X = 1) = \mathbb{P}(X = 0) = 1/2$.

Random variables

This standard notation emphasizes that one rarely is interested in the exact probability space Ω.
An E-valued random variable is called discrete if its law is discrete, namely if the σ-algebra $\mathfrak{E}$ contains singletons and there exists a countable subset S of E such that $\mathbb{P}(X \in S) = 1$.
A real random variable is called continuous if its law is absolutely continuous w.r.t. to the Borel measure. This means that there exists $\varrho \in L_1(\mathbb{R}, dx)$, called the density of X, such that

$$\mathbb{P}(X \in A) = \int_A \varrho(x)\, dx \qquad \text{for all Borelian } A \subset \mathbb{R} \qquad (41)$$

(of course, it is enough to check (41) on intervals by Caratheodory extension theorem).
One can have a real random variable that is not discrete or continuous.

Random variables: examples

Usually names of random variables are just related to their law.

Example

A random variable taking values in a finite space E is called *uniform* (discrete) if

$$\mathbb{P}(X \in A) = |A|/|E| \tag{42}$$

Example

A (real) random variable is called *Bernoulli of parameter* $p \in [0,1]$ if

$$\mathbb{P}(X = 1) = p, \qquad \mathbb{P}(X = 0) = 1 - p \tag{43}$$

This is denoted $\mathrm{Bernoulli}(p)$.

Random variables: examples

Example

A random variable is called *binomial of parameters* $n \geq 1$ *and* $p \in [0,1]$ if

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, \ldots, n \tag{44}$$

This is denoted $\mathrm{Bin}(n, p)$.
Equivalently $X = \sum_{i=1}^{n} X_i$ where the X_i are in independent $\mathrm{Bernoulli}(p)$.

Random variables: examples

Example

A (real) random variable is called *uniform* (continuous) in the interval $[a, b]$ if its density ϱ is given by

$$\varrho(x) := \tfrac{1}{b-a} 1_{[a,b]}(x) \tag{45}$$

This is denoted $U(a, b)$.

Example

A (real) random variable is called *exponential* of parameter $\lambda > 0$ if its density ϱ is given by

$$\varrho(x) := \lambda \exp(-\lambda x) 1_{[0,\infty)}(t) \tag{46}$$

or equivalently if $\mathbb{P}(X > t) = \exp(-\lambda t)$ for all $t \geq 0$. This is denoted $\exp(\lambda)$.

Random variables: examples

Example

A (real) random variable is called *Gaussian* of parameters $\mu \in \mathbb{R}$ and $\sigma^2 \geq 0$ if its density ϱ is given by

$$\varrho(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{47}$$

however for $\sigma = 0$ we simply understand $\mathbb{P}(X = \mu) = 1$ (so the law of X is discrete in this case). This is denoted $\mathcal{N}(\mu, \sigma^2)$.

Random variables: independence

Given an E-valued random variable $X\colon \Omega \to E$, there is a σ-algebra $\mathfrak{F}_X$ associated to X, which is defined as the smallest σ-algebra such that X is measurable.

$$\begin{aligned} &\Omega \ni \omega \mapsto X(\omega) \in E \\ &\mathcal{P}(\Omega) \ni \mathbb{P} \mapsto \mathbb{P} \circ X^{-1} \in \mathcal{P}(E) \\ &\mathbf{2}^E \supset \mathfrak{E} \mapsto \mathfrak{F}_X \subset \mathbf{2}^\Omega \end{aligned} \tag{48}$$

Definition

A family of random variables is *independent* if their σ-algebras are independent. Namely the family $(X_\alpha)_{\alpha\in\mathcal{A}}$, where $X_\alpha\colon \Omega \to E_\alpha$, are independent if for every n, for every $\alpha_1, \ldots, \alpha_n$ and every $A_1, \ldots, A_n$ events in $E_{\alpha_1}, \ldots, E_{\alpha_m}$

$$\mathbb{P}(X_{\alpha_1} \in A_1, \ldots, X_{\alpha_n} \in A_n) = \mathbb{P}(X_{\alpha_1} \in A_1) \ldots \mathbb{P}(X_{\alpha_n} \in A_n) \tag{49}$$

Random variables: independence

Example

Suppose that $X_1, \ldots, X_n$ are independent Bernoulli random variables. Then the random variable $X = \sum_{i=1}^{n} X_i$ is a binomial random variable of parameters n and p.
Indeed $\mathbb{P}(X = k) = \mathbb{P}(|\{i \leq n : X_i = 1\}| = k) = \binom{n}{k} p^k (1-p)^{n-k}$

Notice that, given a probability measure μ on a space E, there is always a random variable with that law, since we can take $\Omega = E$, $\mathfrak{E} = \mathfrak{F}$, $\mathbb{P} = \mu$ and X the identity. In particular, since we built general product spaces, we can build arbitrary families of independent random variables.

Example

A countable infinite family $(X_i)_{i \geq 1}$ of independent Bernoulli(p) random variables is called a Bernoulli scheme. This can equivalently be regarded as a random variable X taking values in $\{0,1\}^{\mathbb{N}}$.

Random variables

Example

A continuous collection $(X_t)_{t\geq 0}$ of independent $\mathcal{N}(0,1)$ random variables is called a white noise.

There is a wide zoology of random variables, named after their law. One can study the theory in general, but of course relevant examples are the ingredient that enriches the theory. Moreover, one may be interest to classify random variables with certain properties etc. That is why particular examples are very popular on probability books.

BEWARE: One uses the law of random variables to give them a name. But of course one may have random variables with the same law that are not the same random variable.

If $X : \Omega \to E$ is a random variable, and $f : E \to F$ is measurable, then $Y = f(X)$ is of course an F-valued random variable. If μ is the law of X, then $\mu \circ f^{-1}$ is the law of Y, simply because $\mathbb{P} \circ Y^{-1} = \mu \circ f^{-1}$.

Random variables

Example

If $(X_i)_{i\geq 1}$ is a Bernoulli scheme of parameter $1/2$, then $Y = \sum_i 2^{-i} X_i$ is *defined with probability* 1 (a priori there may be a set of probability 0 where the X_i are not 0 or 1). The law of Y is the usual Borel measure on $[0,1]$. Indeed

$$\begin{aligned}\mathbb{P}(Y \in [0,1/2)) = \mathbb{P}(X_1 = 0) = \mathbb{P}(Y \in [1/2,1)) = \mathbb{P}(X_1 = 1) = 1/2\\ \mathbb{P}(X_1 = 0, X_2 = 0) = \mathbb{P}(Y \in [0,1/4)) = \mathbb{P}(Y \in [1/4,2/4))\\ = \mathbb{P}(Y \in [2/4,3/4)) = \mathbb{P}(Y \in [3/4,4/4)) = 1/4\end{aligned} \tag{50}$$

and more in general

$$\mathbb{P}(Y \in [k2^{-n}, (k+1)2^{-n})) = 2^{-n}, \qquad k = 0, \dots 2^{-n} - 1 \tag{51}$$

So since the law of Y coincides with the Borel measure on dyadic intervals, it coincides over all intervals and thus everywhere.

Expected value

If X is a random variable $X : \Omega \to [0, \infty]$ its *expected value* $\mathbb{E}[X]$ is nothing but the usual measure-theoretic integral

$$\mathbb{E}[X] := \int_\Omega X(\omega)\, d\mathbb{P}(\omega) \in [0, \infty] \tag{52}$$

while if $X : \Omega \to [-\infty, +\infty]$, $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$ provided these are not both infinite. There is however a more probabilistic way to construct the expected value, without a direct reference to the space Ω.

Definition

Let X be a discrete, nonnegative real random variable, and let S be a countable set such that $\mathbb{P}(X \in S) = 1$. Then define

$$\mathbb{E}[X] := \sum_{s \in S} s\, \mathbb{P}(X = s) \tag{53}$$

This corresponds to the idea of the barycenter of the law of X, and of course coincides with (52) for a discrete random variable.

Expected value

If X is a nonnegative random variable, define

$$\begin{aligned} \overline{\mathbb{E}}[X] &:= \inf_{Y \geq X, Y \text{ nonnegative, discrete}} \mathbb{E}[Y] \\ \underline{\mathbb{E}}[X] &:= \sup_{Z \leq X, Z \text{ nonnegative, discrete}} \mathbb{E}[Z] \end{aligned} \tag{54}$$

Since X is measurable, it is easy to see that these two definitions coincide. For instance the infimum and supremum are attained along the monotone sequences $Y_n = 2^{-n}[2^n X + 1]$, $Z_n := 2^{-n}[2^n X]$. However, by the definition of the expected value for discrete random variables, $0 \leq \mathbb{E}[Y_n] - \mathbb{E}[Z_n] \leq 2^{-n}$.

Definition

For a nonnegative random variable $X : \Omega \to [0, \infty]$, $\mathbb{E}[X]$ is defined as any of the formulas in (54). If $X : \Omega \to [-\infty, \infty]$, then $\mathbb{E}[X] := \mathbb{E}[X^+] - \mathbb{E}[X^-]$, provided $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ are not simultaneously $+\infty$.

Expected value

By the end, the expected value is just a notation for the integral, that is neutral w.r.t. the choice of the probability space Ω. As such, it enjoys all the usual properties of the integral: linearity, monotonicity (if $Y \geq X$ then $\mathbb{E}[Y] \geq \mathbb{E}[X]$). Moreover, since the measure is a probability, we have that if $\mathbb{P}(X = c) = 1$ for some constant $c \in \mathbb{R}$, then $\mathbb{E}[X] = c$.

Remark

The expected value does not depend on the probability space used, but only on the law of the random variable. This is evident for discrete random variables, and thus for any random variable by construction. In particular, if $X : \Omega \to \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ and $Y = f(X)$, then we can calculate $\mathbb{E}[Y]$ using the law $\mathbb{P} \circ f^{-1}$ of the random variable Y on $\mathbb{R}$, or the probability measure $\mathbb{P}$ on Ω obtaining

$$\mathbb{E}[f(X)] = \int f(X(\omega)) d\mathbb{P}(\omega) = \int y \, d(\mu \circ f^{-1})(y) = \mathbb{E}[Y] \quad (55)$$

which is just the change of variable formula.

Expected value

For a continuous random variable with density ϱ, the definition of expected value gives $\mathbb{E}[X] = \int x\, \varrho(x)\, dx$. So more in general we have

- for discrete r.v., $p_s := \mathbb{P}(X = s)$ and $\sum_s p_s = 1$

$$\mathbb{E}[f(X)] = \sum_{s \in S} f(s)\, p_s \tag{56}$$

- for continuous real r.v. with density ϱ

$$\mathbb{E}[f(X)] = \int f(x)\, \varrho(x)\, dx \tag{57}$$

Expected value

Example

If X is a $\mathrm{Bin}(n, p)$ random variable, then $\mathbb{E}[X] = n\,p$. Indeed

$$\sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = n\,p \tag{58}$$

as can be seen by induction over n (or summing directly).
However we could have calculated it much more easily recalling that the expected value is a property of the law of random variable. So recall that $X = \sum_{i=1}^{n} X_i$ where X_i are (independent) $\mathrm{Bernoulli}(p)$. So $\mathbb{E}[X_i] = p * 1 + (1-p) * 0 = p$ and $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np$.

Example

If X is $\exp(\lambda)$, then $\mathbb{E}[X] = \int_0^\infty x\,\lambda \exp(-\lambda x)\,dx = 1/\lambda$ (integrating by parts).

Random variables: recap

Hereafter $(\Omega, \mathfrak{F}, \mathbb{P})$ is a probability space, and $(E, \mathfrak{E})$ or $(E_\alpha, \mathfrak{E}_\alpha)$ are measurable spaces. Recall that an E-random variable is a measurable map $X : \Omega \to E$, and its distribution μ_X is the probability measure on E given by $\mu_X(A) := \mathbb{P}(X^{-1}(A))$.
If X is a real random variable we defined its expected value (when it exists) as

$$\mathbb{E}[X] := \int_\Omega X(\omega) d\mathbb{P}(\omega) = \int_{\mathbb{R}} x \, d\mu_X(x) \tag{59}$$

Definition

A property is said to hold almost surely (a.s. in English, p.s. in the French literature, почти наверное) or *with probability* 1 if it holds everywhere on Ω except on a set contained in an event of measure 0.

For instance on $\Omega = [0, 1]$ with the Borel measure, a number is irrational a.s..

L_p spaces

On the space of random variables there is an obvious equivalence relation: a.s. equality. X and Y are a.s. equal if $\mathbb{P}(X = Y) = 1$. The space of random variables $X : \mathbb{R}$ identified up to a.s. equivalence is called $L^0 \equiv L^0(\Omega, \mathfrak{F}, \mathbb{P})$. (Here $\mathbb{R}$ is equipped with the Lebesgue σ-algebra)[1].
For $p \geq 1$, the space of random variables $\in L^0$ such that $\mathbb{E}[|X|^p] < \infty$ is called L^p. $\|X\|_{L^p} := \mathbb{E}[|X|^p]^{1/p}$ is a Banach norm on L^p. Moreover one also defines L^∞ as the space of $X \in L^0$ such that

$$\|X\|_{L^\infty} := \sup_{p \geq 1} \mathbb{E}[|X|^p]^{1/p} < \infty \tag{60}$$

Notice that $X \in L^\infty$ iff there exists $c \geq 0$ such that $\mathbb{P}(|X| \leq c) = 1$ and actually $c = \|X\|_{L^\infty}$ is the minimal c for which this identity holds.

[1] In probability, sometimes it is better to use complete σ-algebras, sometimes it is not, and it is better to stick with Borel σ-algebras. I agree about measurability being mostly a boring issue, and we try to stay away from it.

Hölder inequality

Remark (Hölder inequality)

If $p, q \geq 1$ *are such that* $1/p + 1/q = 1$, $X \in L^p$ *and* $Y \in L^q$ *then*

$$\mathbb{E}[X\,Y] \leq \|X\|_{L^p}\,\|Y\|_{L^q} \tag{61}$$

The proof is elementary. For each $\lambda > 0$,
$a\,b \leq \lambda^p |a|^p/p + \lambda^{-q}|b|^q/q$. Therefore
$\mathbb{E}[XY] \leq \lambda^p \mathbb{E}[|X|^p] + \lambda^{-q}\mathbb{E}[|Y|^q]/q$. Optimizing over λ we get the result, which (trivially) holds also for $p = 1$ and $q = \infty$.
In particular for a r.v. Z, taking $X = |Z|^r$, $Y = 1$ and $p = s/r$ in the (61), it holds for $s \geq r \geq 1$

$$\|Z\|_{L^r} \leq \|Z\|_{L^s} \tag{62}$$

Hölder inequality

So the Hölder inequality can be iterated to get

Proposition (Hölder inequality II)

Let $r, p_1, \dots, p_n \in [1, \infty]$ *be such that* $\sum_i 1/p_i \leq 1/r$. *Then*

$$\|X_1 \cdots X_n\|_{L^r} \leq \|X_1\|_{L^{p_1}} \cdots \|X_1\|_{L^{p_n}} \tag{63}$$

Another generalization is the following (we will not use it). $\psi : [0, \infty) \to [0, \infty)$ is called a Young function if it is strictly convex, superlinear and $\psi(0) = 0$. For instance $\psi(x) = x^p/p$ is a Young function. Then if ψ and ϕ are Young functions such that $\psi(x) + \phi(y) \leq x\, y$

$$\mathbb{E}[XY] \leq \|X\|_{L^\psi} \|Y\|_{L^\phi} \tag{64}$$

where $\|X\|_{L^\psi} := \inf\{c > 0 : \mathbb{E}[\psi(|X|/c)] \leq 1\}$.

Random vectors and Jensen inequality

Notice that the definition of expected values also makes sense (componentwise) for $\mathbb{R}^d$-valued random variables (sometimes called random vectors). Simply $\mathbb{E}[X]_i = \mathbb{E}[X_i]$. Hölder inequality and L^p spaces easily generalize to this case. For instance
$\mathbb{E}[X \cdot Y] \leq \|X\|_{L^p} \|Y\|_{L^q}$
The following inequality is rather important.

Proposition (Jensen)

Let $\psi \colon \mathbb{R}^d \to \mathbb{R}$ be a convex function bounded from below and let X be an $\mathbb{R}^d$-valued random variable. Then

$$\mathbb{E}[\psi(X)] \geq \psi(\mathbb{E}[X]) \tag{65}$$

Доказательство.

First notice that (65) is well-posed since ψ is bounded from below.

Jensen inequality

Then recall that a convex function ψ is the supremum of the affine functions smaller than ψ:

$$\psi(x) = \sup\{a \cdot x + b \,:\, a \cdot y + b \leq \psi(y)\ \forall y\} \tag{66}$$

Therefore

$$\begin{aligned} \mathbb{E}[\psi(X)] &= \mathbb{E}[\sup_{a,b} a \cdot X + b] \geq \sup_{a,b} \mathbb{E}[a \cdot X + b] \\ &= \sup_{a,b} a \cdot \mathbb{E}[X] + b = \psi(\mathbb{E}[X]) \end{aligned} \tag{67}$$

Variance

The variance of an L^2 random variable X is defined as

$$\mathbb{D}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{68}$$

It quantifies how much a random variable differs from its expectation. Notice that $\mathbb{D}(X) \geq 0$ and $\mathbb{D}(X) = 0$ iff X is constant a.s.. Moreover $\mathbb{D}(\alpha X + \beta) = \alpha^2 \mathbb{D}(X)$ for $\alpha, \beta \in \mathbb{R}$.
If $X = (X_i)_{i=1}^d$ is an $\mathbb{R}^d$-r.v. in L^2, its *covariance* is the symmetric nonnegative matrix, $D_{i,j}(X) := \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$. For instance $D_{i,i}(X) = \mathbb{D}(X_i)$.
The *correlation* of X is the symmetric matrix $R_{i,j}(X) := D_{i,j}(X)/\sqrt{D_{i,i}(X) D_{j,j}}(X)$. Notice that $R_{i,j} \in [-1, 1]$. If $R_{i,j} = 0$, then X_i and X_j are called uncorrelated.

Markov and Chebyshev inequality

Proposition (Markov inequality)

Let $X \in L^1$. Then for $c > 0$

$$\mathbb{P}(|X| \geq c) \leq \mathbb{E}[|X|]/c \tag{69}$$

The proof is a single line

$$\mathbb{P}(|X| \geq c) = \mathbb{E}[1_{|X| \geq c}] \leq \mathbb{E}\Big[\frac{|X|}{c} 1_{|X| \geq c}\Big] \leq \mathbb{E}\Big[\frac{|X|}{c}\Big] \tag{70}$$

Proposition (Chebyshev)

For $X \in L^2$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq c) \leq \mathbb{D}[X]/c^2 \tag{71}$$

To prove it, just apply Markov inequality to the random variable $Y = |X - \mathbb{E}[X]|^2$.

Markov and Chebyshev inequality

The previous inequalities are rather crude, but they have an important interpretation. If $\mathbb{E}[f(X)] < \infty$ for some increasing f, then $\mathbb{P}(X \geq c)$ decreases at least as $1/f(c)$. For instance for $\lambda > 0$

$$\mathbb{P}(X \geq c) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda c}) \leq e^{-\lambda c}\mathbb{E}[e^{\lambda X}] \tag{72}$$

Expectation and independence

Proposition

Let $(X_i)_{i=1}^n$ be an independent family of real random variables $X_i \in L^1(\Omega, \mathfrak{F}, \mathbb{P})$. Then the product $X_1 \cdots X_n$ is in L^1 and it holds

$$\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n] \tag{73}$$

Iterating, it is enough to prove it for $n = 2$. If X_1 and X_2 are discrete random variables, then by definition

$$\mathbb{E}[X_1\, X_2] = \sum_{i,j} \mathbb{P}(X_1 = s_i, X_2 = r_j) s_i\, r_j \tag{74}$$

However by independence
$\mathbb{P}(X_1 = s_i, X_2 = r_j) = \mathbb{P}(X_1 = s_i)\mathbb{P}(X_2 = r_j)$.
So the previous quantity is also equal to

$$\sum_i \mathbb{P}(X_1 = s_i) s_i \sum_j \mathbb{P}(X_2 = r_j)\, r_j := \mathbb{E}[X_1]\mathbb{E}[X_2] \tag{75}$$

Expectation and independence

In the general (non-discrete) case, we can bound the product $X_1\,X_2$ with products of independent discrete r.v. in the definition of the expected value. So the statement follows.
A different proof: since X_1 and X_2 are independent, the distribution $\mu_{(X_1,X_2)}$ of the $\mathbb{R}^2$-random variable (X_1, X_2) is the product of μ_{X_1} and μ_{X_2}. Therefore using the formula of the expected value using the distribution, and Fubini's theorem

$$\begin{aligned}\mathbb{E}[X_1\,X_2] &= \int x_1\,x_2\,d\mu_{(X_1,X_2)}(x_1,x_2) = \int x_1\,x_2\,d\mu_{(X_1,X_2)}(x_1,x_2)\\ &= \int x_1\,x_2\,d\mu_{X_1}(x_1)d\mu_{X_2}(x_2) = \int x_1\,d\mu_{X_1}(x_1)\int x_2\,d\mu_{X_2}(x_2)\\ &= \mathbb{E}[X_1]\mathbb{E}[X_2]\end{aligned} \tag{76}$$

Expectation and independence

Random variables X, Y such that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ are called uncorrelated. So independent variables are uncorrelated. The inverse statement is not true.

Example

If $X \sim \mathcal{N}(0,1)$, and $Y = X^2$, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. But they are not independent. For instance $3 = \mathbb{E}[X^2\, Y] \neq \mathbb{E}[X^2]\mathbb{E}[Y] = 1$.

Remark

The random variables $X_1, \ldots X_n$, with $X_i : \Omega \to E_i$ are independent iff for each measurable bounded $f_i : E_i \to \mathbb{R}$ it holds

$$\mathbb{E}[f_1(X_1) \cdots f_n(X_n)] = \mathbb{E}[f_1(X_1)] \cdots \mathbb{E}[f_n(X_n)] \tag{77}$$

If we take the f_i to be the indicator function of arbitrary measurable sets in E_i, we get back to the definition of independence. On the other, if the X_i are independent so are the r.v. $f_i(X_i)$. So the reverse implication follows from the previous Proposition.

Variance and independence

Remark

It holds $\mathbb{D}(\sum_i X_i) = \sum_{i,j} D_{i,j}(X)$. *In particular if the random variables* $(X_i)_{i \leq n}$ *are pairwise uncorrelated (this is the case if for instance if they are pairwise independent), then*

$$\mathbb{D}(\sum_i X_i) = \sum_i \mathbb{D}(X_i) \tag{78}$$

Define $Y_i = X_i - \mathbb{E}[X_i]$. Then $\mathbb{D}(\sum_i X_i) = \mathbb{D}(\sum_i Y_i) = \sum_{i,j} \mathbb{E}[Y_i Y_j] = \sum_{i,j} D_{i,j}(Y) = \sum_{i,j} D_{i,j}(X)$.

Distribution of real random variables

Notice that expectation, variance, correlation, covariance etc only depend on the distribution μ_X of a random variable. Yet, the writing $\mathbb{E}[X]$ etc provide a useful intuitive suggestion.
The following definition is a basic tool in elementary probability.

Definition

Let μ be a probability measure on $\mathbb{R}$. The cumulative distribution function of μ is the function $F : \mathbb{R} \to [0, 1]$ defined as $F_\mu(x) := \mu((-\infty, x])$. If X is a random variable, its cumulative distribution is the cumulative distribution of its law, namely $F_X(x) \equiv F_{\mu_X}(x) = \mathbb{P}(X \leq x)$.

Notice that this only makes sense for real random variables (although one could generalize to $\mathbb{R}^d$ easily).

Distribution of real random variables

Remark

The cumulative distribution F_μ of a probability measure μ has the following properties

(a) *F_μ is non-decreasing.*

(b) $\lim_{x\to-\infty} F_\mu(x) = 0$ *and* $\lim_{x\to+\infty} F_\mu(x) = 1$.

(c) *F_μ is continuous from the right* $\lim_{x\downarrow y} F_\mu(x) = F_\mu(y)$.

Conversely given a function $F : \mathbb{R} \to [0,1]$ with the properties (a), (b), (c), there exists a probability μ such that $F = F_\mu$.

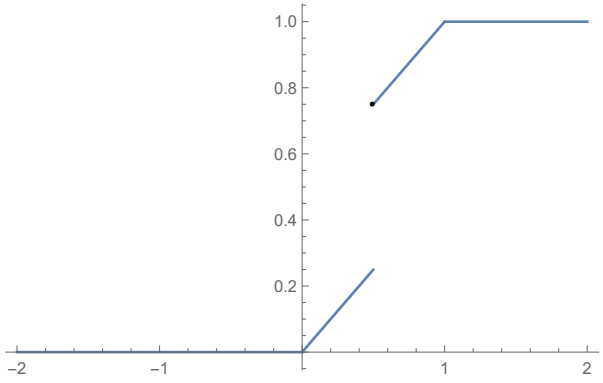
(a) is obvious. (b) follows from the continuity of the probability measures (1st lecture). Indeed $\emptyset = \cap_n(-\infty, -n]$. Therefore
$0 = \mu(\emptyset) = \mu(\lim_n(-\infty, -n]) = \lim_n \mu((-\infty, -n]) = \lim_n F_\mu(-n)$.
Since the limit $\lim_{x\to-\infty} F_\mu(x)$ exists by monotonicity, this limit is 0. Similarly
$\lim_{x\to+\infty} F_\mu(x) = \lim_n \mu((-\infty, n]) = \mu(\lim_n(-\infty, n]) = \mu(\mathbb{R}) = 1$.

Distribution of real random variables

The continuity from the right also follows from the monotonicity and the continuity of the probability since
$\lim_{x \downarrow y} F_\mu(x) = \lim_n F_\mu(y + 1/n) = \lim_n \mu((-\infty, x + 1/n]) =$
$\mu(\lim_n (-\infty, x + 1/n]) = \mu((-\infty, x]) = F_\mu(x)$.
The reverse statement follows from Caratheodory theorem. Given F satisfying (a), (b) and (c), define on intervals
$\mu((a, b]) = F(b) - F(a)$, provided $a \leq b$. This is non-negative from (a), satisfies $\mu(\emptyset) = 0$ and $\mu(\mathbb{R}) = 1$ from (b). From (c) we get that that if μ is σ-additive on intervals of the form $(a, b]$. Since these intervals form a semi-ring, we can apply Caratheory theorem.

Distribution of real random variables

Notice that $\lim_{x\uparrow y} F_\mu(x) = \mu((-\infty, x)) \leq F_\mu(x)$. In particular F_μ is continuous iff all the points have probability 0. For instance if X is a discrete random variable taking the values x_i, then F is an increasing function which is constant except on the x_i while $F(x_i) - F(x_i^-) = \mathbb{P}(X = x_i)$.
If μ admits a density ϱ then by definition $F_\mu(x) = \int_{-\infty}^{x} \varrho(t)dt$. So μ is continuous iff F_μ is absolutely continuous.
F can be used to built a variety of probability measures on $\mathbb{R}$. For instance the Cantor staircase satisfies (a), (b), (c). But the associated μ is not discrete or continuous.



Distribution of real random variables

BEWARE: there is no unique definition of *continuous random variable*. This is just used in elementary probability, and textbooks and literature are not consistent.

Distribution of real random variables

BEWARE: there is no unique definition of *continuous random variable*. This is just used in elementary probability, and textbooks and literature are not consistent.
It may mean

- The distribution of X admits a density. Equivalently, the distribution function F_X is absolutely continuous.
- For each $x \in \mathbb{R}$, $\mathbb{P}(X = x) = 0$. Equivalently the distribution function F_X is continuous.
- X takes value in $\mathbb{R}$ and is not discrete (!).

Calculating expectations

Proposition

For any $a \in \mathbb{R}$ and r.v. $X \in L^1$ it holds

$$\mathbb{E}[X] = a + \int_a^\infty \mathbb{P}(X \geq x)dx - \int_{-\infty}^a \mathbb{P}(X \leq x)dx \qquad (79)$$

We can prove it for $a = 0$, and then apply this result to the random variable $X - a$ for the general case $a \in \mathbb{R}$.
We have that $X^+ = \int_0^\infty 1_{X \geq x}\, dx$. So taking the expected value and using Fubini

$$\mathbb{E}[X^+] = \mathbb{E}\Big[\int_0^\infty 1_{X \geq x}\, dx\Big] = \int_0^\infty \mathbb{E}\big[1_{X \geq x}\big]\, dx = \int_0^\infty \mathbb{P}(X \geq x)\, dx \qquad (80)$$

Similarly we have $\mathbb{E}[X^-] = \int_{-\infty}^0 \mathbb{P}(X \leq x)$ from which (79) follows.

Calculating expectations

Proposition

Let X be a random variable and $f \in C^1(\mathbb{R})$ be non-decreasing and such that $f(X) \in L^1$. Then

$$\mathbb{E}[f(X)] = f(a) + \int_a^\infty f'(x)\mathbb{P}(X \geq x)\, dx - \int_{-\infty}^a f'(x)\mathbb{P}(X \geq x)\, dx \tag{81}$$

Characteristic function

If μ is a probability measure on $\mathbb{R}^d$, its characteristic function φ_μ is defined as

$$\begin{aligned} &\varphi_\mu \colon \mathbb{R}^d \to \mathbb{C} \\ &\varphi_\mu(u) := \int_{\mathbb{R}^d} e^{iu\cdot x} d\mu(x) \end{aligned} \tag{82}$$

Characteristic function

If μ is a probability measure on $\mathbb{R}^d$, its characteristic function φ_μ is defined as

$$\begin{aligned} &\varphi_\mu \colon \mathbb{R}^d \to \mathbb{C} \\ &\varphi_\mu(u) := \int_{\mathbb{R}^d} e^{iu\cdot x} d\mu(x) \end{aligned} \tag{82}$$

For a random variable X, its characteristic function φ_X is simply the characteristic function of its law, namely $\varphi_X \equiv \varphi_{\mu_X}(u) := \mathbb{E}[e^{iu\cdot X}]$.

Characteristic function

If μ is a probability measure on $\mathbb{R}^d$, its characteristic function φ_μ is defined as

$$\begin{aligned} &\varphi_\mu \colon \mathbb{R}^d \to \mathbb{C} \\ &\varphi_\mu(u) := \int_{\mathbb{R}^d} e^{iu\cdot x} d\mu(x) \end{aligned} \tag{82}$$

For a random variable X, its characteristic function φ_X is simply the characteristic function of its law, namely $\varphi_X \equiv \varphi_{\mu_X}(u) := \mathbb{E}[e^{iu\cdot X}]$. In particular, up to a constant, if μ admits a density ϱ, then φ_μ is simply the Fourier transform (Преобразование Фурье) of ϱ.

Characteristic function

Proposition

The characteristic function of a measure (or random variable) satisfies

(a) $|\varphi(u)| \leq \varphi(0) = 1$.

(b) φ *is uniformly continuous.*

Characteristic function

Proposition

The characteristic function of a measure (or random variable) satisfies

(a) $|\varphi(u)| \leq \varphi(0) = 1$.

(b) φ *is uniformly continuous.*

Let us use the notation of random variables. (a) is trivial: $|\varphi(u)| = |\mathbb{E}[e^{iu\cdot X}]| \leq \mathbb{E}[|e^{iu\cdot X}|] = \mathbb{E}[1] = 1$.

Characteristic function

Proposition

The characteristic function of a measure (or random variable) satisfies

(a) $|\varphi(u)| \leq \varphi(0) = 1$.

(b) φ *is uniformly continuous.*

Let us use the notation of random variables. (a) is trivial:
$|\varphi(u)| = |\mathbb{E}[e^{iu\cdot X}]| \leq \mathbb{E}[|e^{iu\cdot X}|] = \mathbb{E}[1] = 1$.
As for (b)

$$|\varphi(u+v) - \varphi(u)| = |\mathbb{E}[e^{i(u+v)\cdot X} - e^{iu\cdot X}]| = |\mathbb{E}[e^{iu\cdot X}(e^{iv\cdot X} - 1)]|$$
$$\leq \mathbb{E}[|e^{iu\cdot X}|\,|e^{iv\cdot X} - 1|] = \mathbb{E}[|e^{iv\cdot X} - 1|] \tag{83}$$

Since $|e^{iv\cdot x} - 1| \leq 2$ is uniformly integrable (равномерно интегрируемая), we have $\lim_{v\to 0} \sup_u |\varphi(u+v) - \varphi(u)| \leq \lim_{v\to 0} \mathbb{E}[|e^{iv\cdot X} - 1|] = \mathbb{E}[\lim_{v\to 0} |e^{iv\cdot X} - 1|] = 0$.

Characteristic function

Theorem

For two probability measures on $\mathbb{R}^d$ *it holds* $\mu = \nu$ *iff* $\varphi_\mu = \varphi_\nu$.

Characteristic function

Theorem

For two probability measures on $\mathbb{R}^d$ it holds $\mu = \nu$ iff $\varphi_\mu = \varphi_\nu$.

Lemma

For two probability measures on $\mathbb{R}^d$ it holds $\mu = \nu$ iff $\int f \, d\mu = \int f \, d\nu$ for all C^∞, compactly supported functions f.

Characteristic function

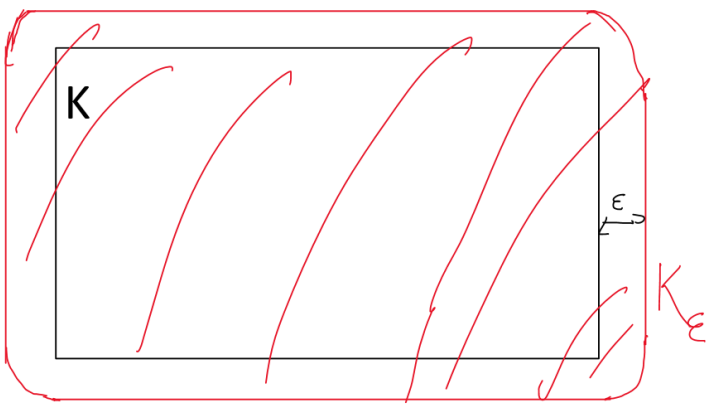
Theorem

For two probability measures on $\mathbb{R}^d$ it holds $\mu = \nu$ iff $\varphi_\mu = \varphi_\nu$.

Lemma

For two probability measures on $\mathbb{R}^d$ it holds $\mu = \nu$ iff $\int f \, d\mu = \int f \, d\nu$ for all C^∞, compactly supported functions f.

For a compact set $K \subset \mathbb{R}^d$ and $\varepsilon > 0$, let K_ε be the ε-enlargement of K.



Characteristic function

Take $f_{K,\varepsilon} \in C^\infty(\mathbb{R}^d; [0,1])$ a function such that $f_{K,\varepsilon}(x) = 1$ for $x \in K$ and $f_{K,\varepsilon}(x) = 0$ for $x \notin K_\varepsilon$.

Characteristic function

Take $f_{K,\varepsilon} \in C^\infty(\mathbb{R}^d; [0,1])$ a function such that $f_{K,\varepsilon}(x) = 1$ for $x \in K$ and $f_{K,\varepsilon}(x) = 0$ for $x \notin K_\varepsilon$.
Then it holds

$$\mu(K) \leq \int f_{K,\varepsilon} d\mu \leq \mu(K_\varepsilon)$$

(and the same for ν).

Characteristic function

Take $f_{K,\varepsilon} \in C^\infty(\mathbb{R}^d; [0,1])$ a function such that $f_{K,\varepsilon}(x) = 1$ for $x \in K$ and $f_{K,\varepsilon}(x) = 0$ for $x \notin K_\varepsilon$.
Then it holds

$$\mu(K) \leq \int f_{K,\varepsilon} d\mu \leq \mu(K_\varepsilon)$$

(and the same for ν).
Since $\cap_{\varepsilon>0} K_\varepsilon = K$, it holds by the continuity of probabilities

$$\mu(K) = \lim_{\varepsilon \to 0} \int f_{K,\varepsilon} d\mu = \lim_{\varepsilon \to 0} \int f_{K,\varepsilon} d\nu = \nu(K) \qquad (84)$$

Since $\mu(K) = \nu(K)$ for any compact set K, $\mu = \nu$.

Characteristic function

Proof of the theorem: Let μ and ν be such that $\varphi_\mu = \varphi_\nu$. Fix f continuous and compactly supported and $\varepsilon > 0$. We will prove

$$|\int f\, d\mu - \int f\, d\nu| \leq 2\varepsilon\,(1 + \max_x |f(x)| + \varepsilon) \tag{85}$$

In view of the previous Lemma and the fact that ε is arbitrary, this proves the statement.

Characteristic function

Proof of the theorem: Let μ and ν be such that $\varphi_\mu = \varphi_\nu$. Fix f continuous and compactly supported and $\varepsilon > 0$. We will prove

$$|\int f\, d\mu - \int f\, d\nu| \leq 2\varepsilon\,(1 + \max_x |f(x)| + \varepsilon) \tag{85}$$

In view of the previous Lemma and the fact that ε is arbitrary, this proves the statement.

By continuity of probability,
$\lim_{L\to\infty} \mu([-L, L]^d) = \nu([-L, L]^d) = 1$. So we can take $L \equiv L_\varepsilon$ such that $\mu([-L, L]^d), \nu([-L, L]^d) \geq 1 - \varepsilon$, and such that $f(x) = 0$ if $|x| \geq L$.

Characteristic function

Proof of the theorem: Let μ and ν be such that $\varphi_\mu = \varphi_\nu$. Fix f continuous and compactly supported and $\varepsilon > 0$. We will prove

$$|\int f\, d\mu - \int f\, d\nu| \leq 2\varepsilon\,(1 + \max_x |f(x)| + \varepsilon) \tag{85}$$

In view of the previous Lemma and the fact that ε is arbitrary, this proves the statement.

By continuity of probability,
$\lim_{L\to\infty} \mu([-L, L]^d) = \nu([-L, L]^d) = 1$. So we can take $L \equiv L_\varepsilon$ such that $\mu([-L, L]^d), \nu([-L, L]^d) \geq 1 - \varepsilon$, and such that $f(x) = 0$ if $|x| \geq L$.

By Fourier theorem one can find a finite combination of complex exponentials

$$f_\varepsilon(x) = \sum_{k=1}^{N} a_k \exp(iu_k \cdot x) \tag{86}$$

(here N, u_k, a_k depend on ε), such that $|f(x) - f_\varepsilon(x)| \leq \varepsilon$ for all $x \in [-L, L]^d$.

Characteristic function

Since $\max_x |f_\varepsilon(x)| \leq \varepsilon + \max_x |f(x)|$, it follows that

$$
\begin{aligned}
&|\int f\, d\mu - \sum_{k=1}^{N} a_k \varphi_\mu(u_k)| = |\int (f - f_\varepsilon)\, d\mu| \\
&\leq |\int_{[-L,L]^d} (f - f_\varepsilon)\, d\mu| + |\int_{|x|\geq L} f_\varepsilon d\mu| \leq \varepsilon + \varepsilon \max_x |f_\varepsilon(x)|
\end{aligned}
\tag{87}
$$

Characteristic function

Since $\max_x |f_\varepsilon(x)| \le \varepsilon + \max_x |f(x)|$, it follows that

$$
\begin{aligned}
&|\int f\, d\mu - \sum_{k=1}^{N} a_k \varphi_\mu(u_k)| = |\int (f - f_\varepsilon)\, d\mu| \\
&\le |\int_{[-L,L]^d} (f - f_\varepsilon)\, d\mu| + |\int_{|x|\ge L} f_\varepsilon d\mu| \le \varepsilon + \varepsilon \max_x |f_\varepsilon(x)|
\end{aligned}
\tag{87}
$$

Applying the same for ν we get (85).

Zoology of random variables

We use the notation of random variables. But in reality all of the following, just depends on their distributions.

Zoology of random variables

We use the notation of random variables. But in reality all of the following, just depends on their distributions.
Constant: If $\mathbb{P}(X = a) = 1$ for some a, then $\mu_X = \delta_a$, $\varphi_X(u) = e^{iu \cdot a}$, $\mathbb{E}[X] = A$, $\mathbb{D}(X) = 0$.

Zoology of random variables

We use the notation of random variables. But in reality all of the following, just depends on their distributions.

Constant: If $\mathbb{P}(X = a) = 1$ for some a, then $\mu_X = \delta_a$, $\varphi_X(u) = e^{iu \cdot a}$, $\mathbb{E}[X] = A$, $\mathbb{D}(X) = 0$.

Bernoulli: If $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$, then $\mu_X = p\delta_1 + (1 - p)\delta_0$, $\varphi_X(u) = pe^{iu} + (1 - p)$, $\mathbb{E}[X] = p$, $\mathbb{D}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p(1 - p)$.

Zoology of random variables

We use the notation of random variables. But in reality all of the following, just depends on their distributions.

Constant: If $\mathbb{P}(X = a) = 1$ for some a, then $\mu_X = \delta_a$, $\varphi_X(u) = e^{iu \cdot a}$, $\mathbb{E}[X] = A$, $\mathbb{D}(X) = 0$.

Bernoulli: If $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$, then $\mu_X = p\delta_1 + (1-p)\delta_0$, $\varphi_X(u) = pe^{iu} + (1-p)$, $\mathbb{E}[X] = p$, $\mathbb{D}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p(1-p)$.

Binomial: If $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, for $k = 0, \ldots, n$, then recall that $X = \sum_{k=1}^{n} X_k$ where the X_k are independent $\mathrm{Bernoulli}(p)$. Therefore $\varphi_X(u) = \mathbb{E}[\prod_k e^{iuX_k}] = \prod_k \mathbb{E}[e^{iuX_k}] = (pe^{iu} + (1-p))^n$. $\mathbb{E}[X] = \sum_k \mathbb{E}[X_k] = np(1-p)$ and $\mathbb{D}(X) = \sum_k \mathbb{D}(X_k) = np(1-p)$.

Zoology of random variables

Geometric: X is called geometric of parameter p if $\mathbb{P}(X = k) = (1-p)^{k-1}p$ for $k \in \mathbb{N}$. This represents the number of failures till the first success, in a procedure of independent tries, where the probability of success of each try is p.

Zoology of random variables

Geometric: X is called geometric of parameter p if $\mathbb{P}(X = k) = (1-p)^{k-1}p$ for $k \in \mathbb{N}$. This represents the number of failures till the first success, in a procedure of independent tries, where the probability of success of each try is p.

$$\varphi_X(u) = \sum_k (1-p)^{k-1} p e^{iuk} = \frac{p}{1-(1-p)e^{iu}}$$

$$\mathbb{E}[X] = \sum_k k(1-p)^{k-1}p = (1-p)/p$$

$$\mathbb{D}(X) = \sum_k k^2(1-p)^{k-1}p - p^{-2} = (1-p)/p^2$$

Zoology of random variables

Poisson: X is called Poisson of parameter λ if $\mathbb{P}(X = k) = e^{-\lambda}\lambda^k/k!$ for $k \in \mathbb{N}$. For instance, if an event has constant rate of happening equal to r, the number of times X it happens in a time period t is a Poisson random variable with parameter $\lambda = rt$ (see the exercises).

Zoology of random variables

Poisson: X is called Poisson of parameter λ if $\mathbb{P}(X = k) = e^{-\lambda}\lambda^k/k!$ for $k \in \mathbb{N}$. For instance, if an event has constant rate of happening equal to r, the number of times X it happens in a time period t is a Poisson random variable with parameter $\lambda = rt$ (see the exercises).

$$\varphi_X(u) = e^{\lambda(e^{iu}-1)}$$

$$\mathbb{E}[X] = \lambda$$

$$\mathbb{D}(X) = \lambda$$

Zoology of random variables

Uniform: If the law of X admits density $\varrho(x) = \frac{1}{b-a} 1_{[a,b]}(x)$ then

$$\varphi_X(u) = \frac{1}{b-a} \int_a^b e^{iux} dx = \frac{e^{iub} - e^{iua}}{iu(b-a)}$$

$$\mathbb{E}[X] = (a+b)/2$$

$$\mathbb{D}(X) = \int_a^b x^2 dx - (a+b)^2/4 = (b-a)^2/12$$

Exponential: If the law of X admits density $\varrho(x) = e^{-\lambda x} 1_{[0,\infty)}(x)$ then

$$\varphi_X(u) = \int_0^\infty e^{iux} e^{-\lambda x} dx = \frac{\lambda}{\lambda - iu}$$

$$\mathbb{E}[X] = 1/\lambda$$

$$\mathbb{D}(X) = 1/\lambda^2$$

Characteristic functions

Recall that X is a $\mathbb{R}^d$-random variable, then

$$\begin{aligned} \varphi_X &: \mathbb{R}^d \to \mathbb{C} \\ \varphi_X(t) &:= \mathbb{E}[e^{it\cdot X}] \end{aligned} \tag{88}$$

is the characteristic functions. We proved that φ is a uniformly continuous function such that $|\varphi(t)| \leq \varphi(0) = 1$.

Characteristic functions

Recall that X is a $\mathbb{R}^d$-random variable, then

$$\begin{aligned} \varphi_X &: \mathbb{R}^d \to \mathbb{C} \\ \varphi_X(t) &:= \mathbb{E}[e^{it\cdot X}] \end{aligned} \tag{88}$$

is the characteristic functions. We proved that φ is a uniformly continuous function such that $|\varphi(t)| \leq \varphi(0) = 1$.
We have also seen that φ_X only depends on the distribution of X, and it characterizes it. So if two random variables have same distribution iff they have the same characteristic function.

Characteristic functions

Recall that X is a $\mathbb{R}^d$-random variable, then

$$\begin{aligned} \varphi_X &: \mathbb{R}^d \to \mathbb{C} \\ \varphi_X(t) &:= \mathbb{E}[e^{it\cdot X}] \end{aligned} \tag{88}$$

is the characteristic functions. We proved that φ is a uniformly continuous function such that $|\varphi(t)| \leq \varphi(0) = 1$.
We have also seen that φ_X only depends on the distribution of X, and it characterizes it. So if two random variables have same distribution iff they have the same characteristic function.
Notice however that it is quite a special properties for a function to be a the characteristic function of some probability distribution (see Bochner's Theorem).

Characteristic functions

Recall that X is a $\mathbb{R}^d$-random variable, then

$$\begin{aligned} \varphi_X &: \mathbb{R}^d \to \mathbb{C} \\ \varphi_X(t) &:= \mathbb{E}[e^{it\cdot X}] \end{aligned} \tag{88}$$

is the characteristic functions. We proved that φ is a uniformly continuous function such that $|\varphi(t)| \leq \varphi(0) = 1$.
We have also seen that φ_X only depends on the distribution of X, and it characterizes it. So if two random variables have same distribution iff they have the same characteristic function.
Notice however that it is quite a special properties for a function to be a the characteristic function of some probability distribution (see Bochner's Theorem).
Therefore we can write certain quantities, such as the expected value of a random variable, using its characteristic function.

Characteristic functions

Proposition

If $\mathbb{E}[|X|^k] < \infty$, *then* φ_X *is* C^k *and*

$$\varphi^{(k)}(t) = i^k \mathbb{E}[e^{itX} X^k] \tag{89}$$

where $\varphi^{(k)}$ *is the k-derivative of* φ. *Moreover* $\varphi^{(k)}(t)$ *is bounded and uniformly continuous.*
In particular

$$\mathbb{E}[X^k] = i^k \varphi^{(k)}(0)$$

Characteristic functions

Proposition

If $\mathbb{E}[|X|^k] < \infty$, *then* φ_X *is* C^k *and*

$$\varphi^{(k)}(t) = i^k \mathbb{E}[e^{itX} X^k] \tag{89}$$

where $\varphi^{(k)}$ *is the k-derivative of* φ. *Moreover* $\varphi^{(k)}(t)$ *is bounded and uniformly continuous.*
In particular

$$\mathbb{E}[X^k] = i^k \varphi^{(k)}(0)$$

To prove the statement, one may proceed by induction over k. To ease the notation simple, let us do it for $k = 1$.

Characteristic functions

Proposition

If $\mathbb{E}[|X|^k] < \infty$, *then* φ_X *is* C^k *and*

$$\varphi^{(k)}(t) = i^k \mathbb{E}[e^{itX} X^k] \tag{89}$$

where $\varphi^{(k)}$ *is the k-derivative of* φ. *Moreover* $\varphi^{(k)}(t)$ *is bounded and uniformly continuous.*
In particular

$$\mathbb{E}[X^k] = i^k \varphi^{(k)}(0)$$

To prove the statement, one may proceed by induction over k. To ease the notation simple, let us do it for $k = 1$.

$$\lim_{h \to 0} \frac{\varphi(t+h) - \varphi(t)}{h} = \lim_{h \to 0} \mathbb{E}[e^{itX} \frac{e^{ihX} - 1}{h}] \tag{90}$$

Characteristic functions

Since $|e^{itX}\frac{e^{ihX}-1}{h}| = |\frac{e^{ihX}-1}{h}| \leq |X|$ and $\mathbb{E}[|X|] < \infty$, we can pass to the limit inside the expected value by uniform integrability:

$$\lim_{h\to 0} \frac{\varphi(t+h)-\varphi(t)}{h} = \mathbb{E}[e^{itX} iX] \tag{91}$$

Let us check that $\varphi^{(1)}$ is uniformly continuous.

$$|\varphi^{(1)}(t+h) - \varphi^{(1)}(t)| = |\mathbb{E}[e^{itX} X(e^{ihX}-1)]| \tag{92}$$

so that

$$\overline{\lim_{h\to 0}} \sup_{t\in\mathbb{R}} |\varphi^{(1)}(t+h) - \varphi^{(1)}(t)| \leq \overline{\lim_{h\to 0}} \mathbb{E}[|e^{ihX}-1||X|] = 0 \tag{93}$$

where in the last line we used that the integrand is bounded by $2|X|$ and thus uniformly integrable.

Characteristic functions

Since $|e^{itX}\frac{e^{ihX}-1}{h}| = |\frac{e^{ihX}-1}{h}| \leq |X|$ and $\mathbb{E}[|X|] < \infty$, we can pass to the limit inside the expected value by uniform integrability:

$$\lim_{h\to 0} \frac{\varphi(t+h)-\varphi(t)}{h} = \mathbb{E}[e^{itX} iX] \tag{91}$$

Let us check that $\varphi^{(1)}$ is uniformly continuous.

$$|\varphi^{(1)}(t+h) - \varphi^{(1)}(t)| = |\mathbb{E}[e^{itX}X(e^{ihX}-1)]| \tag{92}$$

so that

$$\overline{\lim_{h\to 0}} \sup_{t\in\mathbb{R}} |\varphi^{(1)}(t+h) - \varphi^{(1)}(t)| \leq \overline{\lim_{h\to 0}} \mathbb{E}[|e^{ihX}-1||X|] = 0 \tag{93}$$

where in the last line we used that the integrand is bounded by $2|X|$ and thus uniformly integrable.

So we have

- φ_X is uniformly continuous: any X.
- $\varphi_X \in C^k$, $\varphi(k)$ uniformly continuous: if $\mathbb{E}[|X|^k] < \infty$.
- φ_X is analytic in $|t| < R$: if $\mathbb{E}[e^{R|X|}] < \infty$.

Gaussian variables

So far we did not dedicate much attention to a fundamental class of random variables: Gaussian random variables.

Gaussian variables

So far we did not dedicate much attention to a fundamental class of random variables: Gaussian random variables.

Definition

An $\mathbb{R}^d$-random variable is called Gaussian if its characteristic function is the exponential of a polynomial of degree at most 2.

If we recall that $|\varphi_X| \leq \varphi(0) = 1$, it necessarily follows that there exists a nonnegative definite, symmetric $d \times d$ matrix S and a $m \in \mathbb{R}^d$ such that

$$\mathbb{E}[e^{it\cdot X}] = e^{-\frac{1}{2}St\cdot t + it\cdot m} \tag{94}$$

Gaussian variables

So far we did not dedicate much attention to a fundamental class of random variables: Gaussian random variables.

Definition

An $\mathbb{R}^d$-random variable is called Gaussian if its characteristic function is the exponential of a polynomial of degree at most 2.

If we recall that $|\varphi_X| \leq \varphi(0) = 1$, it necessarily follows that there exists a nonnegative definite, symmetric $d \times d$ matrix S and a $m \in \mathbb{R}^d$ such that

$$\mathbb{E}[e^{it\cdot X}] = e^{-\frac{1}{2}St\cdot t + it\cdot m} \tag{94}$$

Indeed, since $\varphi(0) = 1$, we have that the polynomial must have vanishing coefficient of t^0. On the other hand, if S has a strictly negative eigenvalue, then taking t proportional to the corresponding eigenvector one see that φ would be unbounded, which is impossible.

Gaussian variables

So far we did not dedicate much attention to a fundamental class of random variables: Gaussian random variables.

Definition

An $\mathbb{R}^d$-random variable is called Gaussian if its characteristic function is the exponential of a polynomial of degree at most 2.

If we recall that $|\varphi_X| \leq \varphi(0) = 1$, it necessarily follows that there exists a nonnegative definite, symmetric $d \times d$ matrix S and a $m \in \mathbb{R}^d$ such that

$$\mathbb{E}[e^{it\cdot X}] = e^{-\frac{1}{2}St\cdot t + it\cdot m} \tag{94}$$

Indeed, since $\varphi(0) = 1$, we have that the polynomial must have vanishing coefficient of t^0. On the other hand, if S has a strictly negative eigenvalue, then taking t proportional to the corresponding eigenvector one see that φ would be unbounded, which is impossible.

In this case one notes $X \sim \mathcal{N}(m, S)$.

Gaussian variables

Actually, it would be enough to define a scalar Gaussian variable as before (so for $d = 1$), and then define for any (real) topological vector space E:

Definition

An E-random variable is Gaussian if for any element if for any continuous linear $\ell\colon E \to \mathbb{R}$, $\ell(X)$ is Gaussian.

Gaussian variables

Actually, it would be enough to define a scalar Gaussian variable as before (so for $d = 1$), and then define for any (real) topological vector space E:

Definition

An E-random variable is Gaussian if for any element if for any continuous linear $\ell\colon E \to \mathbb{R}$, $\ell(X)$ is Gaussian.

It is easy to see that, if $E = \mathbb{R}^d$, this is equivalent to the previous definition. Indeed, if $E = \mathbb{R}^d$, we can identify any linear continuous map $\ell\mathbb{R}^d \to \mathbb{R}$ as a scalar product with an element of $\mathbb{R}^d$, so $\ell(x) = \ell \cdot x$. Since for each $\ell \in \mathbb{R}^d$ $\ell \cdot X$ is Gaussian we have

$$\mathbb{E}[e^{it\ell\cdot X}] = e^{-t^2\sigma_\ell^2 + itm_\ell} \qquad \forall t \in \mathbb{R} \tag{95}$$

for some $\sigma_\ell^2 \geq 0$ and $m_\ell \in \mathbb{R}$. However, writing the same formula with ℓ changed to $\alpha\ell_1 + \ell_2$ we see that σ_ℓ is a (nonnegative) quadratic form and m_ℓ linear in ℓ. Namely $\sigma_\ell^2 = S\ell \cdot \ell$ and $m_\ell = m \cdot \ell$ for some symmetric non-negative definite matrix S and $m \in \mathbb{R}^d$. So for $t = 1$ we get back to the previous definition.

Gaussian variables

In particular, at least for $\mathbb{R}^d$, the equivalence of these definitions can be stated as follows

A random vector $X = (X_1, \ldots, X_d)$ is Gaussian iff any linear combination $\sum_i \ell_i X_i$ with deterministic coefficients, is Gaussian.

Gaussian variables

In particular, at least for $\mathbb{R}^d$, the equivalence of these definitions can be stated as follows

A random vector $X = (X_1, \dots, X_d)$ is Gaussian iff any linear combination $\sum_i \ell_i X_i$ with deterministic coefficients, is Gaussian.

Notice that if $d = 1$ and $X \sim \mathcal{N}(m, s)$, then necessarily $s \geq 0$. If $s = 0$, then $\mathbb{P}(X = m) = 1$, as we have already seen that e^{itm} is the characteristic function of a constant (which is indeed Gaussian by our definition). On the other hand, if $s =: \sigma^2 > 0$, we claim that the distribution of X admits density

$$\varrho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{96}$$

Gaussian variables

In particular, at least for $\mathbb{R}^d$, the equivalence of these definitions can be stated as follows

A random vector $X = (X_1, \ldots, X_d)$ is Gaussian iff any linear combination $\sum_i \ell_i X_i$ with deterministic coefficients, is Gaussian.

Notice that if $d = 1$ and $X \sim \mathcal{N}(m, s)$, then necessarily $s \geq 0$. If $s = 0$, then $\mathbb{P}(X = m) = 1$, as we have already seen that e^{itm} is the characteristic function of a constant (which is indeed Gaussian by our definition). On the other hand, if $s =: \sigma^2 > 0$, we claim that the distribution of X admits density

$$\varrho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{96}$$

Indeed (this computation is enough since the characteristic functions characterizes the distribution)

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} e^{itx} dx = e^{-t^2\sigma^2 + itm} \tag{97}$$

Gaussian variables

In particular, at least for $\mathbb{R}^d$, the equivalence of these definitions can be stated as follows

A random vector $X = (X_1, \ldots, X_d)$ is Gaussian iff any linear combination $\sum_i \ell_i X_i$ with deterministic coefficients, is Gaussian.

Notice that if $d = 1$ and $X \sim \mathcal{N}(m, s)$, then necessarily $s \geq 0$. If $s = 0$, then $\mathbb{P}(X = m) = 1$, as we have already seen that e^{itm} is the characteristic function of a constant (which is indeed Gaussian by our definition). On the other hand, if $s =: \sigma^2 > 0$, we claim that the distribution of X admits density

$$\varrho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{96}$$

Indeed (this computation is enough since the characteristic functions characterizes the distribution)

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} e^{itx} dx = e^{-t^2\sigma^2 + itm} \tag{97}$$

Gaussian variables

An integration by parts shows that

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} x\, e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = m \tag{98}$$

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} (x-m)^2\, e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = \sigma^2 \tag{99}$$

So the parameters m, σ^2 have a clear meaning: $m = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{D}(X)$.

Gaussian variables

An integration by parts shows that

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} x\, e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = m \tag{98}$$

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} (x-m)^2\, e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = \sigma^2 \tag{99}$$

So the parameters m, σ^2 have a clear meaning: $m = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{D}(X)$.

Let us see what happens in higher dimension

Gaussian variables

An integration by parts shows that

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} x\, e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = m \tag{98}$$

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} (x-m)^2\, e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = \sigma^2 \tag{99}$$

So the parameters m, σ^2 have a clear meaning: $m = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{D}(X)$.

Let us see what happens in higher dimension

Proposition

Let $X \sim \mathcal{N}(m, S)$ be an $\mathbb{R}^d$ Gaussian random variable. Then $m = \mathbb{E}[X]$ and $S_{i,j} = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$.

Gaussian variables

An integration by parts shows that

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} x\, e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = m \tag{98}$$

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} (x-m)^2\, e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = \sigma^2 \tag{99}$$

So the parameters m, σ^2 have a clear meaning: $m = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{D}(X)$.

Let us see what happens in higher dimension

Proposition

Let $X \sim \mathcal{N}(m, S)$ be an $\mathbb{R}^d$ Gaussian random variable. Then $m = \mathbb{E}[X]$ and $S_{i,j} = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$.

To prove this statement, we can consider the case $m = 0$ and then apply the theorem to the variable $Y = X - m$ for the general case.

Gaussian variables

An integration by parts shows that

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} x\, e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = m \tag{98}$$

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} (x-m)^2\, e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = \sigma^2 \tag{99}$$

So the parameters m, σ^2 have a clear meaning: $m = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{D}(X)$.

Let us see what happens in higher dimension

Proposition

Let $X \sim \mathcal{N}(m, S)$ be an $\mathbb{R}^d$ Gaussian random variable. Then $m = \mathbb{E}[X]$ and $S_{i,j} = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$.

To prove this statement, we can consider the case $m = 0$ and then apply the theorem to the variable $Y = X - m$ for the general case. So assume $m = 0$. Then for $\ell \in \mathbb{R}^d$ and $t > 0$ it holds $\mathbb{E}[e^{it\ell \cdot X}] = \exp(-t^2 S\ell \cdot \ell/2)$. So $S\ell \cdot \ell = \mathbb{E}[(\ell \cdot X)^2]$ by the result in dimension one. So by polarization $S_{i,j} = \mathbb{E}[X_i X_j]$.

Gaussian variables

Notice that $X_1, \ldots, X_n$ are independent iff S is diagonal, so we have the simple statement A Gaussian vector is independent iff it is decorrelated $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j]$.

Gaussian variables

Notice that $X_1, \ldots, X_n$ are independent iff S is diagonal, so we have the simple statement A Gaussian vector is independent iff it is decorrelated $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j]$.
Let $X \sim \mathcal{N}(m, S)$, and $Y = AX + r$ for some matrices A and vector r. Since any linear combination of Y is an affine function of X, Y is still Gaussian. Moreover $\mathbb{E}[Y] = Am + r$ by linearity and

$$\begin{aligned} &\mathbb{E}[(Y_i - (Am + r)_i)(Y_j - (Am + r)_j)] \\ &\quad = \mathbb{E}[(A(X - m))_i (A(X - m)_j)] = (ASA^\dagger)_{i,j} \end{aligned} \tag{100}$$

Therefore $Y \sim \mathcal{N}(Am + r, ASA^\dagger)$.

Gaussian variables

Let $X \sim \mathcal{N}(m, S)$. By the spectral theorem, there exists an orthogonal matrix that diagonalizes S. If we choose $Y = AX - Am$, we have that (up to reordering the Y_i):

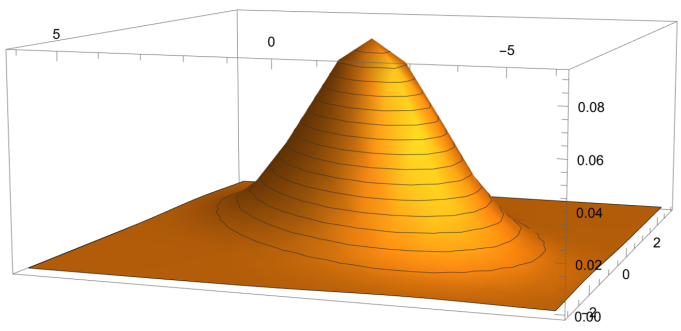
$$Y \sim \mathcal{N}(0, D) \tag{101}$$

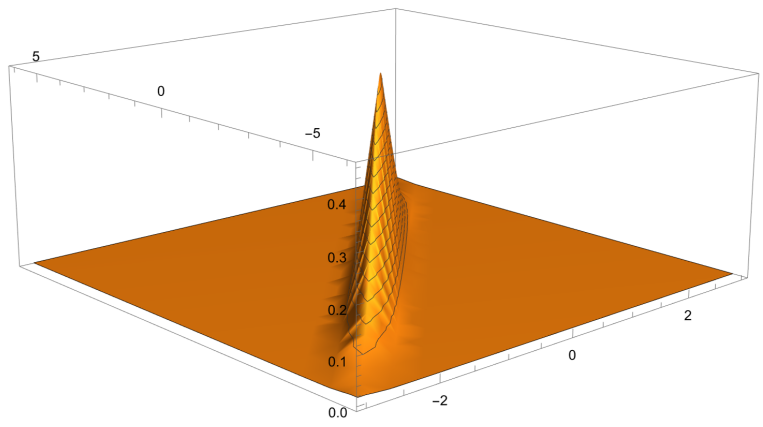
where $D = \mathrm{diag}(0, 0, \ldots, 0, \lambda_1, \ldots, \lambda_{d-k})$. Here k is the dimension of the kernel of S and $\lambda_i > 0$.
Thus $Y = (0, \ldots, 0, Y_1, \ldots, Y_{d-k})$ have all independent entries, the first k being 0, while $Y_i \sim \mathcal{N}(0, \lambda_i)$.
In particular, transforming the density back to X, we see that if S is invertible, X admits density on $\mathbb{R}^d$

$$\varrho(x) = \frac{1}{\sqrt{2\pi \det(S)}} \exp(-\tfrac{1}{2} S^{-1}(x - m) \cdot (x - m)) \tag{102}$$

Gaussian variables

Gaussian density in $d = 2$. On the left, the correlation matrix has equal eigenvalues. On the right, as one eigenvalue vanishes, the density concentrates on a linear subspace.





Gaussian variables

If $X \sim \mathcal{N}(0,1)$, integrating the density by parts, we see that if f is a smooth function with $f(X) \in L_1$

$$\mathbb{E}[f'(X)] = \mathbb{E}[Xf(X)] \tag{103}$$

If $X \sim \mathcal{N}(m, \sigma^2)$, this generalizes to

$$\mathbb{E}[f'(X)] = \frac{1}{\sigma^2}\mathbb{E}[(X-m)f(X)] \tag{104}$$

Measure theory:Radon Nikodym

Recall that a measure ν is absolutely continuous w.r.t. μ if $\nu(A) = 0$ whenever $\mu(A) = 0$. One writes $\nu \ll \mu$.

Measure theory:Radon Nikodym

Recall that a measure ν is absolutely continuous w.r.t. μ if $\nu(A) = 0$ whenever $\mu(A) = 0$. One writes $\nu \ll \mu$.
For instance, if $\varrho \geq 0$ is a measurable function and μ is a given measure, then we can define a new measure ν setting

$$\nu(E) = \int_E \varrho d\mu \qquad \text{for all measurable } E \tag{105}$$

Clearly $\nu \ll \mu$. In this case one uses the notation $\nu = \varrho\,\mu$ or sometimes $d\nu = \varrho\, d\mu$.

Measure theory:Radon Nikodym

Recall that a measure ν is absolutely continuous w.r.t. μ if $\nu(A) = 0$ whenever $\mu(A) = 0$. One writes $\nu \ll \mu$.
For instance, if $\varrho \geq 0$ is a measurable function and μ is a given measure, then we can define a new measure ν setting

$$\nu(E) = \int_E \varrho d\mu \qquad \text{for all measurable } E \tag{105}$$

Clearly $\nu \ll \mu$. In this case one uses the notation $\nu = \varrho\,\mu$ or sometimes $d\nu = \varrho\, d\mu$.
Radon Nikodym theorem states that all absolutely continuous measures have this form for some density ϱ.

Theorem (Radon-Nikodym)

Let μ be a σ-finite measure on a measurable space $(\Omega, \mathfrak{F})$. If $\nu \ll \mu$, then there exists a measurable function $\varrho\colon \Omega \to [0, \infty]$ such that $\nu = \varrho\mu$. ϱ is unique up to a.e. equivalence. If ν is finite, then $\varrho \in L_1(\mu)$.

Measure theory:Radon Nikodym

Given two measures λ, ν, we say that $\lambda \geq \mu$ if $\lambda(A) \geq \nu(A)$ for all measurable A.

Measure theory:Radon Nikodym

Given two measures λ, ν, we say that $\lambda \geq \mu$ if $\lambda(A) \geq \nu(A)$ for all measurable A.

Existence: First assume that μ and ν are finite. Define

$$\mathcal{S} := \{f \text{ measurable s.t. } f\mu \leq \nu\} \tag{106}$$

If $f, g \in \mathcal{S}$, then $\max(f, g) \in \mathcal{S}$, indeed:

$$\begin{aligned} \int_A \max(f, g)\, d\mu &= \int_{A \cap \{f \geq g\}} f\, d\mu + \int_{A \cap \{f < g\}} g\, d\mu \\ &\leq \nu(A \cap \{f \geq g\}) + \nu(A \cap \{f < g\}) = \nu(A) \end{aligned} \tag{107}$$

Notice that

$$\sup_{f \in \mathcal{S}} \int f\, d\mu \leq \nu(\Omega) < \infty \tag{108}$$

Since supremum of subsets of $\mathbb{R}$ are achieved along sequences, we can take a sequence $f_n \in \mathcal{S}$ such that

$$\sup_{f \in \mathcal{S}} \int f\, d\mu = \lim_n \int f_n\, d\mu \tag{109}$$

Measure theory:Radon Nikodym

With no loss of generality, we can assume f_n to be increasing (otherwise change f_n to $\max_{i\leq n} f_i$) which is increasing, satisfies (109) and is in $\mathcal{S}$ in view of (107)). Define then

$$\varrho = \lim_n f_n = \sup_n f_n \tag{110}$$

ϱ is non-negative, measurable. We claim that $\varrho \in \mathcal{S}$, indeed by monotone convergence

$$\int_A \varrho \, d\mu = \lim_n \int_A f_n \, d\mu \leq \nu(A) \qquad \forall A \in \mathfrak{F} \tag{111}$$

Moreover ϱ is maximal in $\mathcal{S}$ in the following sense: if $f \in \mathcal{S}$, then $f \leq \varrho$ a.e., as

$$\int_\Omega \varrho \, d\mu = \sup_{g\in S} \int_\Omega g \, d\mu \geq \int_\Omega \max(f, \varrho) \, d\mu \tag{112}$$

Measure theory:Radon Nikodym

Define the measure $\lambda = \varrho\,\mu$ as in (105). We want to show that $\lambda = \nu$

Since $\varrho \in \mathcal{S}$, $\lambda \leq \nu$.

Measure theory:Radon Nikodym

Define the measure $\lambda = \varrho\,\mu$ as in (105). We want to show that $\lambda = \nu$
Since $\varrho \in \mathcal{S}$, $\lambda \leq \nu$.
Define now (if $\mu(\Omega) = 0$ the theorem is trivial)

$$\varepsilon := \frac{\nu(\Omega) - \lambda(\Omega)}{2\mu(\Omega)} \geq 0 \tag{113}$$

To prove the opposite inequality $\lambda \geq \nu$, it is enough to check that $\varepsilon = 0$. Indeed, since we just proved $\nu \geq \lambda$ we have

$$0 \leq \sup_{A \in \mathfrak{F}} \nu(A) - \lambda(A) = \nu(\Omega) - \lambda(\Omega) = 2\varepsilon\mu(\Omega) \tag{114}$$

Measure theory:Radon Nikodym

Let $B \subset \Omega$ be a set of full measure for the positive part of the measure $\nu - \lambda - \varepsilon\mu$. In particular
$\nu(A \cap B) \geq \lambda(A \cap B) + \varepsilon\mu(A \cap B)$ for all measurable A. Therefore $1_B\,\nu \geq (\varrho\, 1_B + \varepsilon 1_B)\mu$, so that $(\varrho + \varepsilon)1_B \in \mathcal{S}$. So, since ϱ is maximal, we have $\varrho 1_B \leq \varrho$ a.e.. This yields $\varepsilon = 0$ or $\mu(B) = 0$. In the first case, we are done. In the second case, $\mu(B) = 0$, the measure $\nu - \lambda - \varepsilon\mu$ is negative, but then $\nu(\Omega) \leq \lambda(\Omega) + \varepsilon\mu(\Omega)$ which implies $\varepsilon = 0$ together with (113).

Measure theory:Radon Nikodym

Let $B \subset \Omega$ be a set of full measure for the positive part of the measure $\nu - \lambda - \varepsilon\mu$. In particular
$\nu(A \cap B) \geq \lambda(A \cap B) + \varepsilon\mu(A \cap B)$ for all measurable A. Therefore $1_B\,\nu \geq (\varrho\, 1_B + \varepsilon 1_B)\mu$, so that $(\varrho + \varepsilon)1_B \in \mathcal{S}$. So, since ϱ is maximal, we have $\varrho 1_B \leq \varrho$ a.e.. This yields $\varepsilon = 0$ or $\mu(B) = 0$. In the first case, we are done. In the second case, $\mu(B) = 0$, the measure $\nu - \lambda - \varepsilon\mu$ is negative, but then $\nu(\Omega) \leq \lambda(\Omega) + \varepsilon\mu(\Omega)$ which implies $\varepsilon = 0$ together with (113).
If the measures are both σ-finite, we can write Ω as a countable disjoint union $\Omega = \cup_n \Omega_n$ with $\mu(\Omega_n), \nu(\Omega_n) < \infty$, and use the previous proof to consistently define ϱ on each of the Ω_n. We skip the case where only μ is σ-finite.

Measure theory:Radon Nikodym

Let $B \subset \Omega$ be a set of full measure for the positive part of the measure $\nu - \lambda - \varepsilon\mu$. In particular
$\nu(A \cap B) \geq \lambda(A \cap B) + \varepsilon\mu(A \cap B)$ for all measurable A. Therefore $1_B\,\nu \geq (\varrho\,1_B + \varepsilon 1_B)\mu$, so that $(\varrho + \varepsilon)1_B \in \mathcal{S}$. So, since ϱ is maximal, we have $\varrho 1_B \leq \varrho$ a.e.. This yields $\varepsilon = 0$ or $\mu(B) = 0$. In the first case, we are done. In the second case, $\mu(B) = 0$, the measure $\nu - \lambda - \varepsilon\mu$ is negative, but then $\nu(\Omega) \leq \lambda(\Omega) + \varepsilon\mu(\Omega)$ which implies $\varepsilon = 0$ together with (113).
If the measures are both σ-finite, we can write Ω as a countable disjoint union $\Omega = \cup_n \Omega_n$ with $\mu(\Omega_n), \nu(\Omega_n) < \infty$, and use the previous proof to consistently define ϱ on each of the Ω_n. We skip the case where only μ is σ-finite.
Uniqueness If $\nu = \varrho_1\mu = \varrho_2\mu$ then $\int_E (\varrho_1 - \varrho_2)\,d\mu = 0$, for every measurable E, which yields $\varrho_1 = \varrho_2$ a.e..

Measure theory:Radon Nikodym

If $\nu = \varrho\mu$ one writes

$$\varrho = \frac{d\nu}{d\mu} \tag{115}$$

This is called the density of ν, or the Radon Nikodym derivative of ν w.r.t. μ. This is a useful notation since it is easily seen that

- $\int f d\nu = \int f \frac{d\nu}{d\mu} d\mu$ for all $f \in L_1(\mu)$.

Measure theory:Radon Nikodym

If $\nu = \varrho\mu$ one writes

$$\varrho = \frac{d\nu}{d\mu} \tag{115}$$

This is called the density of ν, or the Radon Nikodym derivative of ν w.r.t. μ. This is a useful notation since it is easily seen that

- $\int f d\nu = \int f \frac{d\nu}{d\mu} d\mu$ for all $f \in L_1(\mu)$.
- $\frac{d\lambda}{d\nu}\frac{d\nu}{d\mu} = \frac{d\lambda}{d\mu}$, in particular $1/(\frac{d\mu}{d\nu}) = \frac{d\nu}{d\mu}$.

Measure theory:Radon Nikodym

If $\nu = \varrho\mu$ one writes

$$\varrho = \frac{d\nu}{d\mu} \tag{115}$$

This is called the density of ν, or the Radon Nikodym derivative of ν w.r.t. μ. This is a useful notation since it is easily seen that

- $\int f d\nu = \int f \frac{d\nu}{d\mu} d\mu$ for all $f \in L_1(\mu)$.
- $\frac{d\lambda}{d\nu}\frac{d\nu}{d\mu} = \frac{d\lambda}{d\mu}$, in particular $1/(\frac{d\mu}{d\nu}) = \frac{d\nu}{d\mu}$.
- $\frac{d(a\nu+b\lambda)}{d\mu} = a\frac{d\nu}{d\mu} + b\frac{d\lambda}{d\mu}$

Measure theory:Radon Nikodym

If $\nu = \varrho\mu$ one writes

$$\varrho = \frac{d\nu}{d\mu} \tag{115}$$

This is called the density of ν, or the Radon Nikodym derivative of ν w.r.t. μ. This is a useful notation since it is easily seen that

- $\int f d\nu = \int f \frac{d\nu}{d\mu} d\mu$ for all $f \in L_1(\mu)$.
- $\frac{d\lambda}{d\nu}\frac{d\nu}{d\mu} = \frac{d\lambda}{d\mu}$, in particular $1/(\frac{d\mu}{d\nu}) = \frac{d\nu}{d\mu}$.
- $\frac{d(a\nu+b\lambda)}{d\mu} = a\frac{d\nu}{d\mu} + b\frac{d\lambda}{d\mu}$

Notice that if Ω is also a complete metric space with $\mathfrak{F}$ the Borel σ-algebra, and if ϱ is almost everywhere continuous, then

$$\varrho(x) = \lim_{\varepsilon \downarrow 0} \frac{\nu(B_\varepsilon(x))}{\mu(B_\varepsilon(x))} \tag{116}$$

Conditional expectation

We have defined $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$. It is however very natural to condition w.r.t. sets of probability 0. Also, we have defined independence for generic σ-algebras (including sets of probability 0). So we would also like to define conditional probability and expectations given any σ-algebra.

Conditional expectation

We have defined $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$. It is however very natural to condition w.r.t. sets of probability 0. Also, we have defined independence for generic σ-algebras (including sets of probability 0). So we would also like to define conditional probability and expectations given any σ-algebra.
Giving a σ-algebra is the mathematical way to describe how much information is available, what one can and cannot measure.
Suppose that we are interested in the temperature of the air. Right now, the temperature is a random variable Y, in one hour it will be X. We can use the present information to better predict X. For instance, at present we compute the expected value $\mathbb{E}[X]$; but if we measure Y, we can improve this, by calculating the expected value of X knowing Y. This is denoted $\mathbb{E}[X|Y]$. If the temperature is for instance a continuous random variable, every possible value of Y will have probability 0. Yet it is quite clear that $\mathbb{E}[X|Y]$ should mean something.

Conditional expectation

What properties should we want for $\mathbb{E}[X|Y]$?

- It should be a function of Y. If I know Y, I should know $\mathbb{E}[X|Y]$. So $\mathbb{E}[X|Y] = Z$ is random variable, as Y is indeed itself random, even if at some point we will observe it.
- Its average should still be $\mathbb{E}[X]$, namely $\mathbb{E}[Z] = \mathbb{E}[X]$.
- More generally, it is natural to ask $\mathbb{E}[Z\,1_{Y\in A}] = \mathbb{E}[X\,1_{Y\in A}]$.

Conditional expectation

What properties should we want for $\mathbb{E}[X|Y]$?

- It should be a function of Y. If I know Y, I should know $\mathbb{E}[X|Y]$. So $\mathbb{E}[X|Y] = Z$ is random variable, as Y is indeed itself random, even if at some point we will observe it.
- Its average should still be $\mathbb{E}[X]$, namely $\mathbb{E}[Z] = \mathbb{E}[X]$.
- More generally, it is natural to ask $\mathbb{E}[Z\,1_{Y\in A}] = \mathbb{E}[X\,1_{Y\in A}]$.

We can then give a mathematical definition of $\mathbb{E}[X|\mathfrak{G}]$ when $\mathfrak{G}$ is a σ-algebra, and which satisfies these requirements when $\mathfrak{G}$ is the σ-algebra generated by Y (the smallest one such that Y is measurable).

Conditional expectation

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space, and $\mathfrak{G}$ a sub σ-algebra of $\mathfrak{F}$. Let $X \in L_1(\mathbb{P})$. There is a unique, up to a.e. equivalence, random variable $Z \in L_1(\mathbb{P})$ such that

(a) Z is $\mathfrak{G}$-measurable.

(b) $\mathbb{E}[SZ] = \mathbb{E}[SX]$ for any $\mathfrak{G}$-measurable bounded random variable S.

Indeed, on $(\Omega, \mathfrak{G})$ consider the following two finite measures

$$\begin{aligned} &\mu \equiv \mathbb{P}_{|\mathfrak{G}} \text{ defined by } \mu(A) = \mathbb{P}(A) \qquad A \in \mathfrak{G} \\ &\nu \text{ defined by } \nu(A) = \mathbb{E}[X 1_A] \qquad A \in \mathfrak{G} \end{aligned} \tag{117}$$

Conditional expectation

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space, and $\mathfrak{G}$ a sub σ-algebra of $\mathfrak{F}$. Let $X \in L_1(\mathbb{P})$. There is a unique, up to a.e. equivalence, random variable $Z \in L_1(\mathbb{P})$ such that

(a) Z is $\mathfrak{G}$-measurable.

(b) $\mathbb{E}[SZ] = \mathbb{E}[SX]$ for any $\mathfrak{G}$-measurable bounded random variable S.

Indeed, on $(\Omega, \mathfrak{G})$ consider the following two finite measures

$$\begin{aligned} &\mu \equiv \mathbb{P}_{|\mathfrak{G}} \text{ defined by } \mu(A) = \mathbb{P}(A) \qquad A \in \mathfrak{G} \\ &\nu \text{ defined by } \nu(A) = \mathbb{E}[X1_A] \qquad A \in \mathfrak{G} \end{aligned} \tag{117}$$

Then $\nu \ll \mu$ and we set $Z = \frac{d\nu}{d\mu}$. (a) holds by construction. While (b) coincides with the definition of density of ν, so that it is unique up to a.e. equivalence.

Conditional expectation

The random variable Z constructed as above is denoted

$$Z = \mathbb{E}[X|\mathfrak{G}] \tag{118}$$

In view of (a) and (b), this corresponds to the expected value of X once we have the information $\mathfrak{G}$.

Example

Let $\mathfrak{G}$ be the σ-algebra generated by a finite partition $A_1, \ldots, A_n$ of Ω. Then $Z = \mathbb{E}[X|\mathfrak{G}]$ is random variable which is constant on each A_i and

$$\mathbb{E}[X|\mathfrak{G}](\omega) = \frac{\mathbb{E}[X1_{A_i}]}{\mathbb{P}(A_i)} \qquad \text{if } \omega \in A_i \tag{119}$$

Conditional expectation

The random variable Z constructed as above is denoted

$$Z = \mathbb{E}[X|\mathfrak{G}] \tag{118}$$

In view of (a) and (b), this corresponds to the expected value of X once we have the information $\mathfrak{G}$.

Example

Let $\mathfrak{G}$ be the σ-algebra generated by a finite partition $A_1, \ldots, A_n$ of Ω. Then $Z = \mathbb{E}[X|\mathfrak{G}]$ is random variable which is constant on each A_i and

$$\mathbb{E}[X|\mathfrak{G}](\omega) = \frac{\mathbb{E}[X1_{A_i}]}{\mathbb{P}(A_i)} \qquad \text{if } \omega \in A_i \tag{119}$$

At least when X is nice (for instance continuous on a topological Ω), (116) tell us that we can think $\mathbb{E}[X|\mathfrak{G}]$ as a limit of the previous example when we partition Ω in small sets.

Conditional probability

If $X = 1_A$, then one denotes

$$\mathbb{E}[1_A|\mathfrak{G}] = \mathbb{P}(A|\mathfrak{G}) \tag{120}$$

Conditional probability

If $X = 1_A$, then one denotes

$$\mathbb{E}[1_A|\mathfrak{G}] = \mathbb{P}(A|\mathfrak{G}) \tag{120}$$

For each measurable $A \in \mathfrak{F}$, $\mathbb{P}(A|\mathfrak{G})$ is a $\mathfrak{G}$-measurable function on Ω. However, for each fixed ω in Ω, it is easy to check that $A \mapsto \mathbb{P}(A|\mathfrak{G})(\omega)$ is a probability. So given a σ-algebra $\mathfrak{G}$, it remains defined a (measurable) map

$$\Omega \ni \omega \mapsto \mathbb{P}(\cdot|\mathfrak{G})(\omega) \in \mathcal{P}(\Omega; \mathfrak{F}) \tag{121}$$

Conditional probability

If $X = 1_A$, then one denotes

$$\mathbb{E}[1_A|\mathfrak{G}] = \mathbb{P}(A|\mathfrak{G}) \tag{120}$$

For each measurable $A \in \mathfrak{F}$, $\mathbb{P}(A|\mathfrak{G})$ is a $\mathfrak{G}$-measurable function on Ω. However, for each fixed ω in Ω, it is easy to check that $A \mapsto \mathbb{P}(A|\mathfrak{G})(\omega)$ is a probability. So given a σ-algebra $\mathfrak{G}$, it remains defined a (measurable) map

$$\Omega \ni \omega \mapsto \mathbb{P}(\cdot|\mathfrak{G})(\omega) \in \mathcal{P}(\Omega; \mathfrak{F}) \tag{121}$$

If $\mathfrak{G} = \{\emptyset, \Omega, B, B^c\}$, for some B with $\mathbb{P}(B) \in (0, 1)$, we get back to the old elementary conditional probability:

$$P(A|\mathfrak{G})(\omega) = \begin{cases} P(A|B) & \text{if } \omega \in B \\ P(A|B^c) & \text{if } \omega \in B^c \end{cases} \tag{122}$$

Conditional expectation

Proposition

It holds

(a) *Conditional expectation is linear:* $\mathbb{E}[aX + bY|\mathfrak{G}] = a\mathbb{E}[X|\mathfrak{G}] + b\mathbb{E}[Y|\mathfrak{G}]$, *and monotone: if* $X \geq Y$, $\mathbb{E}[X|\mathfrak{G}] \geq \mathbb{E}[Y|\mathfrak{G}]$.

(b) $\mathbb{E}[\mathbb{E}[X|\mathfrak{G}]] = \mathbb{E}[X]$.

(c) *If* $\phi\colon \mathbb{R} \to \mathbb{R}$ *is a convex function such that* $\phi(X) \in L_1$, *then* $\mathbb{E}[\phi(X)|\mathfrak{G}] \geq \phi(\mathbb{E}[X|\mathfrak{G}])$. *In particular* $|\mathbb{E}[X|\mathfrak{G}]| \leq \mathbb{E}[|X||\mathfrak{G}]$.

(d) *If* X *is* $\mathfrak{G}$*-measurable, then* $\mathbb{E}[X|\mathfrak{G}] = X$. *More generally, if* X *is* $\mathfrak{G}$*-measurable* $\mathbb{E}[XY|\mathfrak{G}] = X\mathbb{E}[Y|\mathfrak{G}]$.

(e) *If* X *is independent of* $\mathfrak{G}$, *then* $\mathbb{E}[X|\mathfrak{G}] = \mathbb{E}[X]$.

(f) *If* $\mathfrak{H} \subset \mathfrak{G}$, *then* $\mathbb{E}[\mathbb{E}[X|\mathfrak{G}]\mathfrak{H}] = \mathbb{E}[X|\mathfrak{H}]$. *In particular for* $\mathfrak{H}$ *the trivial* σ*-algebra*

$$\mathbb{E}[\mathbb{E}[X|\mathfrak{G}]] = \mathbb{E}[X] \tag{123}$$

Conditional expectation

(a) Use the definition as the Radon-Nikodym derivavite.

(b) Use $S = 1$ in the definition (b) of conditional expectation.

Conditional expectation

(a) Use the definition as the Radon-Nikodym derivavite.
(b) Use $S = 1$ in the definition (b) of conditional expectation.
(c) As in the proof of Jensen inequality. Since ϕ is convex, $\phi(x) = \sup_{a,b} ax + b$, where the supremum runs over all the a, b such that $ay + b \leq \phi(y)$ for every y. Then

$$\begin{aligned} \mathbb{E}[\phi(X)|\mathfrak{G}] &= \mathbb{E}[\sup_{a,b} aX + b|\mathfrak{G}] \geq sup_{a,b}\mathbb{E}[aX + b|\mathfrak{G}] \\ &= \sup_{a,b} a\mathbb{E}[X|\mathfrak{G}] + b = \phi(\mathbb{E}[X|\mathfrak{G}]) \end{aligned} \tag{124}$$

Conditional expectation

(a) Use the definition as the Radon-Nikodym derivavite.
(b) Use $S = 1$ in the definition (b) of conditional expectation.
(c) As in the proof of Jensen inequality. Since ϕ is convex, $\phi(x) = \sup_{a,b} ax + b$, where the supremum runs over all the a, b such that $ay + b \leq \phi(y)$ for every y. Then

$$\begin{aligned} \mathbb{E}[\phi(X)|\mathfrak{G}] &= \mathbb{E}[\sup_{a,b} aX + b|\mathfrak{G}] \geq sup_{a,b}\mathbb{E}[aX + b|\mathfrak{G}] \\ &= \sup_{a,b} a\mathbb{E}[X|\mathfrak{G}] + b = \phi(\mathbb{E}[X|\mathfrak{G}]) \end{aligned} \tag{124}$$

(d) Check that $X\mathbb{E}[Y|\mathfrak{G}]$ satisfies properties (a) an (b) in the definition of conditional expectation.

Conditional expectation

(a) Use the definition as the Radon-Nikodym derivavite.
(b) Use $S = 1$ in the definition (b) of conditional expectation.
(c) As in the proof of Jensen inequality. Since ϕ is convex, $\phi(x) = \sup_{a,b} ax + b$, where the supremum runs over all the a, b such that $ay + b \leq \phi(y)$ for every y. Then

$$\begin{aligned} \mathbb{E}[\phi(X)|\mathfrak{G}] &= \mathbb{E}[\sup_{a,b} aX + b|\mathfrak{G}] \geq sup_{a,b}\mathbb{E}[aX + b|\mathfrak{G}] \\ &= \sup_{a,b} a\mathbb{E}[X|\mathfrak{G}] + b = \phi(\mathbb{E}[X|\mathfrak{G}]) \end{aligned} \tag{124}$$

(d) Check that $X\mathbb{E}[Y|\mathfrak{G}]$ satisfies properties (a) an (b) in the definition of conditional expectation.
(e) Independence implies that for every $\mathfrak{G}$-measurable $Y \in L_1$, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. So $\mathbb{E}[X]$ satisfies (a) an (b) in the definition of conditional expectation.

Conditional expectation

(a) Use the definition as the Radon-Nikodym derivavite.
(b) Use $S = 1$ in the definition (b) of conditional expectation.
(c) As in the proof of Jensen inequality. Since ϕ is convex, $\phi(x) = \sup_{a,b} ax + b$, where the supremum runs over all the a, b such that $ay + b \leq \phi(y)$ for every y. Then

$$\begin{aligned} \mathbb{E}[\phi(X)|\mathfrak{G}] &= \mathbb{E}[\sup_{a,b} aX + b|\mathfrak{G}] \geq sup_{a,b}\mathbb{E}[aX + b|\mathfrak{G}] \\ &= \sup_{a,b} a\mathbb{E}[X|\mathfrak{G}] + b = \phi(\mathbb{E}[X|\mathfrak{G}]) \end{aligned} \tag{124}$$

(d) Check that $X\mathbb{E}[Y|\mathfrak{G}]$ satisfies properties (a) an (b) in the definition of conditional expectation.
(e) Independence implies that for every $\mathfrak{G}$-measurable $Y \in L_1$, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. So $\mathbb{E}[X]$ satisfies (a) an (b) in the definition of conditional expectation.
(f) Let $Z = \mathbb{E}[X|\mathfrak{G}]$. $\mathbb{E}[Z|\mathfrak{H}]$ is $\mathfrak{H}$-measurable, so we need to check $\mathbb{E}[\mathbb{E}[Z|\mathfrak{H}]Y] = \mathbb{E}[XY]$ for all Y $\mathfrak{H}$-measurable. This is immediate from (d) and the fact that Y is $\mathfrak{G}$-measurable.

Condtional expectation

For Y be a random variable, let $\mathfrak{G}_Y$ be the smallest σ-algebra w.r.t. which Y is measurable. Notice that if Z is $\mathfrak{G}_Y$-measurable random variable, then $Z = f(Y)$ for some measurable f.
Then one usually writes $\mathbb{E}[X|Y]$ to mean $\mathbb{E}[X|\mathfrak{G}_Y]$. Notice that this is a function of Y since it is $\mathfrak{G}_Y$ measurable. So

$$\mathbb{E}[X|Y] = f(Y) \tag{125}$$

for some function f. This allows the informal notation

$$\mathbb{E}[X|Y = y] = f(y) \tag{126}$$

which has the powerful interpretation *expected value of X knowing that $Y = y$*.

Condtional expectation

For Y be a random variable, let $\mathfrak{G}_Y$ be the smallest σ-algebra w.r.t. which Y is measurable. Notice that if Z is $\mathfrak{G}_Y$-measurable random variable, then $Z = f(Y)$ for some measurable f.
Then one usually writes $\mathbb{E}[X|Y]$ to mean $\mathbb{E}[X|\mathfrak{G}_Y]$. Notice that this is a function of Y since it is $\mathfrak{G}_Y$ measurable. So

$$\mathbb{E}[X|Y] = f(Y) \tag{125}$$

for some function f. This allows the informal notation

$$\mathbb{E}[X|Y = y] = f(y) \tag{126}$$

which has the powerful interpretation *expected value of X knowing that $Y = y$*.
Notice that in this definition X needs to be a real random variable, but Y can take value in any measurable set.

Convergence of random variables

Let X be random variable and (X_n) be a sequence of random variables, all defined on the same probability space $X, X_n \colon \Omega \to E$. We think that E is a nice metric space (complete, separable), but you may well think $E = \mathbb{R}^n$ or even $E = \mathbb{R}$ if the general framework is too abstract.

Convergence of random variables

Let X be random variable and (X_n) be a sequence of random variables, all defined on the same probability space $X, X_n \colon \Omega \to E$. We think that E is a nice metric space (complete, separable), but you may well think $E = \mathbb{R}^n$ or even $E = \mathbb{R}$ if the general framework is too abstract.
We say that

- X_n converges to X *almost surely* or *with probability* 1 if $\mathbb{P}(\lim_n X_n = X)$.

Convergence of random variables

Let X be random variable and (X_n) be a sequence of random variables, all defined on the same probability space $X, X_n \colon \Omega \to E$. We think that E is a nice metric space (complete, separable), but you may well think $E = \mathbb{R}^n$ or even $E = \mathbb{R}$ if the general framework is too abstract.
We say that

- X_n converges to X *almost surely* or *with probability* 1 if $\mathbb{P}(\lim_n X_n = X)$.
- X_n converges to X *in probability* if $\lim_n \mathbb{P}(\mathrm{distance}(X_n, X) > \varepsilon) = 0$, for every $\varepsilon > 0$.

Convergence of random variables

Let X be random variable and (X_n) be a sequence of random variables, all defined on the same probability space $X, X_n \colon \Omega \to E$. We think that E is a nice metric space (complete, separable), but you may well think $E = \mathbb{R}^n$ or even $E = \mathbb{R}$ if the general framework is too abstract.
We say that

- X_n converges to X *almost surely* or *with probability* 1 if $\mathbb{P}(\lim_n X_n = X)$.
- X_n converges to X *in probability* if $\lim_n \mathbb{P}(\mathrm{distance}(X_n, X) > \varepsilon) = 0$, for every $\varepsilon > 0$.
- In L^p, with $p \geq 1$, if $\lim_n \mathbb{E}(\mathrm{distance}(X_n, X)^p) = 0$.

Convergence of random variables

Let X be random variable and (X_n) be a sequence of random variables, all defined on the same probability space $X, X_n : \Omega \to E$. We think that E is a nice metric space (complete, separable), but you may well think $E = \mathbb{R}^n$ or even $E = \mathbb{R}$ if the general framework is too abstract.
We say that

- X_n converges to X *almost surely* or *with probability* 1 if $\mathbb{P}(\lim_n X_n = X)$.
- X_n converges to X *in probability* if $\lim_n \mathbb{P}(\mathrm{distance}(X_n, X) > \varepsilon) = 0$, for every $\varepsilon > 0$.
- In L^p, with $p \geq 1$, if $\lim_n \mathbb{E}(\mathrm{distance}(X_n, X)^p) = 0$.

This are just probabilistic notations for the 'almost everywhere', 'in measure' and 'in L^p' convergence, that you have already encountered in measure theory.

Convergence of random variables

The following relations hold between these definitions.

Convergence of random variables

The following relations hold between these definitions.

Proposition

Let X_n, X as above. Then

(a) *If $X_n \to X$ a.s., then $X_n \to X$ in probability.*

(b) *If $\sum_n \mathbb{P}(\mathrm{distance}(X_n, X) > \varepsilon) < \infty$ for every $\varepsilon > 0$, then $X_n \to X$ a.s.. In particular, if $X_n \to X$ in probability, then there exists a subsequence such that $X_{n_k} \to X$ a.s..*

(c) *If $X_n \to X$ in L^p, then $X_n \to X$ in probability.*

(d) *If $X_n \to X$ a.s. and X_n is p-uniformly integrable, then $X_n \to X$ in L^p. Here p-uniformly integrable means that for some $x \in E$*

$$\lim_{M \to \infty} \sup_n \mathbb{E}[\mathrm{distance}(X_n, x)^p 1_{\mathrm{distance}(X_n,x)^p > M}] = 0 \qquad (127)$$

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.
We also proved $\mathbb{P}(\underline{\lim}_n A_n) \leq \underline{\lim}_n \mathbb{P}(A_n)$ and the Borel-Cantelli lemma: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\overline{\lim}_n A_n) = 0$.

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.

We also proved $\mathbb{P}(\underline{\lim}_n A_n) \leq \underline{\lim}_n \mathbb{P}(A_n)$ and the Borel-Cantelli lemma: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\overline{\lim}_n A_n) = 0$.

Proof: Write $d(x, y)$ for the distance of E, and let $A_{n,k} := \{d(X_n, X) > 1/k\}$. Notice $\{\lim_n X_n = X\} = \cap_k \underline{\lim}_n A_{n,k}^c$.

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.

We also proved $\mathbb{P}(\underline{\lim}_n A_n) \leq \underline{\lim}_n \mathbb{P}(A_n)$ and the Borel-Cantelli lemma: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\overline{\lim}_n A_n) = 0$.

Proof: Write $d(x, y)$ for the distance of E, and let $A_{n,k} := \{d(X_n, X) > 1/k\}$. Notice $\{\lim_n X_n = X\} = \cap_k \underline{\lim}_n A_{n,k}^c$.

(a) If $\mathbb{P}(\lim_n X_n = X) = 1$, then for every k we have $1 = \mathbb{P}(\underline{\lim}_n A_{n,k}^c) \leq \underline{\lim}_n \mathbb{P}(A_{n,k}^c)$. So $\overline{\lim}_n \mathbb{P}(A_{n,k}) = 0$ which means convergence in probability.

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.

We also proved $\mathbb{P}(\underline{\lim}_n A_n) \leq \underline{\lim}_n \mathbb{P}(A_n)$ and the Borel-Cantelli lemma: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\overline{\lim}_n A_n) = 0$.

Proof: Write $d(x, y)$ for the distance of E, and let $A_{n,k} := \{d(X_n, X) > 1/k\}$. Notice $\{\lim_n X_n = X\} = \cap_k \underline{\lim}_n A_{n,k}^c$.

(a) If $\mathbb{P}(\lim_n X_n = X) = 1$, then for every k we have $1 = \mathbb{P}(\underline{\lim}_n A_{n,k}^c) \leq \underline{\lim}_n \mathbb{P}(A_{n,k}^c)$. So $\overline{\lim}_n \mathbb{P}(A_{n,k}) = 0$ which means convergence in probability.

(b) By Borel-Cantelli lemma, we have that $\mathbb{P}(\overline{\lim}_n A_{n,k}) = 0$ for every k. So $\mathbb{P}(\cup_k \overline{\lim}_n A_{n,k}^c) = 0$, which means convergence a.s..

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.
We also proved $\mathbb{P}(\underline{\lim}_n A_n) \leq \underline{\lim}_n \mathbb{P}(A_n)$ and the Borel-Cantelli lemma: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\overline{\lim}_n A_n) = 0$.
Proof: Write $d(x, y)$ for the distance of E, and let $A_{n,k} := \{d(X_n, X) > 1/k\}$. Notice $\{\lim_n X_n = X\} = \cap_k \underline{\lim}_n A_{n,k}^c$.
(a) If $\mathbb{P}(\lim_n X_n = X) = 1$, then for every k we have $1 = \mathbb{P}(\underline{\lim}_n A_{n,k}^c) \leq \underline{\lim}_n \mathbb{P}(A_{n,k}^c)$. So $\overline{\lim}_n \mathbb{P}(A_{n,k}) = 0$ which means convergence in probability.
(b) By Borel-Cantelli lemma, we have that $\mathbb{P}(\overline{\lim}_n A_{n,k}) = 0$ for every k. So $\mathbb{P}(\cup_k \overline{\lim}_n A_{n,k}^c) = 0$, which means convergence a.s.. To deduce that convergence in probability implies convergence a.s. along subsequences, notice that we can take

$$n_j := \inf\{m : \mathbb{P}(A_{r,j}) \leq 2^{-j} \, \forall r \geq m\} \tag{128}$$

This n_j exists since $\mathbb{P}(A_{n,k}) \to 0$ for every k. Moreover it is immediate to see the series $\sum_j \mathbb{P}(A_{n_j,k})$ converges.

Convergence of random variables

(c) This is just Markov inequality
$\mathbb{P}(d(X_n, X) \geq \varepsilon) \leq \mathbb{E}[d(X_n, X)^p]\varepsilon^{-p}$.

Convergence of random variables

(c) This is just Markov inequality
$\mathbb{P}(d(X_n, X) \geq \varepsilon) \leq \mathbb{E}[d(X_n, X)^p]\varepsilon^{-p}$.
(d) By triangular inequality,
$d(X_n, X)^p \leq c_p(d(X_n, x)^p + d(X, x)^p)$. So it is also uniformly integrable and we conclude by Lebesgue convergence.

Convergence of random variables

Proposition (Law of large numbers)

Let X_i be an i.i.d. sequence of real random variables with finite expectation $m = \mathbb{E}[X_i]$. Define

$$S_n := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{129}$$

Suppose that $\mathbb{E}[X_i^2] < \infty$. Then $S_n \to m$ in probability.
Suppose that $\mathbb{E}[X_i^4] < \infty$. Then $S_n \to m$ a.s.

Convergence of random variables

Proposition (Law of large numbers)

Let X_i be an i.i.d. sequence of real random variables with finite expectation $m = \mathbb{E}[X_i]$. Define

$$S_n := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{129}$$

Suppose that $\mathbb{E}[X_i^2] < \infty$. Then $S_n \to m$ in probability.
Suppose that $\mathbb{E}[X_i^4] < \infty$. Then $S_n \to m$ a.s.

Actually a much stronger result holds, due to Kolmogorov: *Suppose that $\mathbb{E}[|X_i|] < \infty$. Then $S_n \to m$ a.s.*.

Convergence of random variables

Proposition (Law of large numbers)

Let X_i be an i.i.d. sequence of real random variables with finite expectation $m = \mathbb{E}[X_i]$. Define

$$S_n := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{129}$$

Suppose that $\mathbb{E}[X_i^2] < \infty$. Then $S_n \to m$ in probability.
Suppose that $\mathbb{E}[X_i^4] < \infty$. Then $S_n \to m$ a.s.

Actually a much stronger result holds, due to Kolmogorov: *Suppose that $\mathbb{E}[|X_i|] < \infty$. Then $S_n \to m$ a.s.*.
We however prove the weaker version. For the first statement, note that $\mathbb{E}[S_n] = m$ and $\mathbb{D}(S_n) = \mathbb{D}(X_i)/n$. So

$$\mathbb{P}(|S_n - m| \geq \varepsilon) \leq \frac{\mathbb{D}(S_n)}{\varepsilon^2} = \frac{\mathbb{D}(X_i)}{n\varepsilon^2} \to 0 \tag{130}$$

namely convergence in probability.

Convergence of random variables

If now X_i has four moments, using the point (b) of the previous Proposition, we need to show that

$$\sum_n \mathbb{P}(|S_n - m| \geq \varepsilon) < \infty \tag{131}$$

Indeed $\mathbb{P}(|S_n - m| \geq \varepsilon) \leq \mathbb{E}[|S_n - m|^4]\varepsilon^{-4}$. But

$$\begin{aligned}\mathbb{E}[|S_n - m|^4] &= \tfrac{1}{n^4}\sum_i \mathbb{E}[(X_i - m)^4] + \tfrac{1}{n^4}\sum_{i\neq j} \mathbb{E}[(X_i - m)^2]\mathbb{E}[(X_j - m)^2] \\ &= \tfrac{1}{n^3}\mathbb{E}[(X_1 - m)^4] + \frac{n(n-1)}{n^4}\mathbb{E}[(X_1 - m)^2]^2\end{aligned} \tag{132}$$

which is summable.

Convergence of random variables

If now X_i has four moments, using the point (b) of the previous Proposition, we need to show that

$$\sum_n \mathbb{P}(|S_n - m| \geq \varepsilon) < \infty \tag{131}$$

Indeed $\mathbb{P}(|S_n - m| \geq \varepsilon) \leq \mathbb{E}[|S_n - m|^4]\varepsilon^{-4}$. But

$$\begin{aligned}\mathbb{E}[|S_n - m|^4] &= \tfrac{1}{n^4}\sum_i \mathbb{E}[(X_i - m)^4] + \tfrac{1}{n^4}\sum_{i\neq j} \mathbb{E}[(X_i - m)^2]\mathbb{E}[(X_j - m)^2] \\ &= \tfrac{1}{n^3}\mathbb{E}[(X_1 - m)^4] + \frac{n(n-1)}{n^4}\mathbb{E}[(X_1 - m)^2]^2\end{aligned} \tag{132}$$

which is summable. This Proposition tells us that as we keep playing head and tail, the average number of heads will converge to $1/2$. It is very intuitive.

Convergence of distributions

We have stressed many times that often one is not interested in a random variable, but just in its distribution. So we are going to define a most useful notion of convergence. We take $E = \mathbb{R}^d$ here, since dealing with non-locally compact spaces requires a slightly more delicate definition.

Convergence of distributions

We have stressed many times that often one is not interested in a random variable, but just in its distribution. So we are going to define a most useful notion of convergence. We take $E = \mathbb{R}^d$ here, since dealing with non-locally compact spaces requires a slightly more delicate definition.

E comes equippes with the Borel σ-algebra. We then consider the following topology on $\mathcal{P}(E)$ (the space of probability measures on E): the weakest topology such that the map $\mu \mapsto \int f \, d\mu$ is continuous for every continuous, compactly supported function $f : E \to \mathbb{R}$.

Convergence of distributions

We have stressed many times that often one is not interested in a random variable, but just in its distribution. So we are going to define a most useful notion of convergence. We take $E = \mathbb{R}^d$ here, since dealing with non-locally compact spaces requires a slightly more delicate definition.

E comes equippes with the Borel σ-algebra. We then consider the following topology on $\mathcal{P}(E)$ (the space of probability measures on E): the weakest topology such that the map $\mu \mapsto \int f \, d\mu$ is continuous for every continuous, compactly supported function $f : E \to \mathbb{R}$.

Then we will say that $X_n \to X$ *in law* or in distribution if the distribution of X_n converges to the distribution of X.

Convergence of distributions

We have stressed many times that often one is not interested in a random variable, but just in its distribution. So we are going to define a most useful notion of convergence. We take $E = \mathbb{R}^d$ here, since dealing with non-locally compact spaces requires a slightly more delicate definition.

E comes equippes with the Borel σ-algebra. We then consider the following topology on $\mathcal{P}(E)$ (the space of probability measures on E): the weakest topology such that the map $\mu \mapsto \int f \, d\mu$ is continuous for every continuous, compactly supported function $f : E \to \mathbb{R}$.

Then we will say that $X_n \to X$ *in law* or in distribution if the distribution of X_n converges to the distribution of X.

To avoid a too abstract formulation, let us give a less intrinsic but more operative definition.

Convergence of distributions

Proposition

Let X_n, X be $\mathbb{R}^d$-random variables (possibly each defined on a different probability space). The following are equivalent

(a) $\lim_n \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ *for every C^∞, compactly supported function $f : \mathbb{R}^d \to \mathbb{R}$.*

(b) $\lim_n \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ *for every continuous, bounded function $f : \mathbb{R}^d \to \mathbb{R}$*

(c) $\lim_n \varphi_{X_n}(t) = \varphi_X(t)$ *for every $t \in \mathbb{R}^d$ (pointwise convergence of the characteristic function).*

(d) $\underline{\lim}_n \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$ *for every lower semicontinuous function $f : \mathbb{R}^d \to \mathbb{R}$.*

(e) $\underline{\lim}_n \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$ *for every upper semicontinuous function $f : \mathbb{R}^d \to \mathbb{R}$.*

Convergence of distributions

(f) $\underline{\lim}_n \mathbb{E}[1_O] \geq \mathbb{E}[1_O]$ for every open set O.

(g) $\overline{\lim}_n \mathbb{E}[1_C] \leq \mathbb{E}[1_C]$ for every open set C.

(h) In $d = 1$, the cumulative distribution function $F_n(x) := \mathbb{P}(X_n \leq x)$ converges (pointwise) to $F(x) := \mathbb{P}(X \leq x)$ at every point x where F is continuous.

Convergence of distributions

(f) $\underline{\lim}_n \mathbb{E}[1_O] \geq \mathbb{E}[1_O]$ for every open set O.

(g) $\overline{\lim}_n \mathbb{E}[1_C] \leq \mathbb{E}[1_C]$ for every open set C.

(h) In $d = 1$, the cumulative distribution function $F_n(x) := \mathbb{P}(X_n \leq x)$ converges (pointwise) to $F(x) := \mathbb{P}(X \leq x)$ at every point x where F is continuous.

Proof: (a) $\Rightarrow$ (b) Assume (a). For $L > 0$, let χ_L be a C^∞ function $0 \leq \chi_L \leq 1$ such that that $\chi_L(x) = 1$ for $|x| \leq L$ and $\chi_L(x) = 0$ for $|x| \geq L$. If f is continuous and bounded, $f\chi_L$ is also compactly supported, and for every $L > 0$ and $\varepsilon > 0$ we can find a C^∞, compactly supported $f_{L,\varepsilon}$ such that $|f\chi_L - f_{L,\varepsilon}| \leq \varepsilon$. Then

$$
\begin{aligned}
|\mathbb{E}[f(X_n) - f(X)]| &\leq |\mathbb{E}[f(X_n)(1-\chi_L)(X_n)]| + |\mathbb{E}[(f\chi_L)(X_n) - f_{L,\varepsilon}(X_n)]| \\
&\quad + |\mathbb{E}[f_{L,\varepsilon}(X_n) - f_{L,\varepsilon}(X)]| + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]| \\
&\leq \|f\|(1 - \mathbb{E}[\chi_L(X_n)]) + \varepsilon \\
&\quad + |\mathbb{E}[f_{L,\varepsilon}(X_n) - f_{L,\varepsilon}(X)]| + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]|
\end{aligned}
\tag{133}
$$

Convergence of distributions

From (a), $\mathbb{E}[f_{L,\varepsilon}(X_n)]$ and $\mathbb{E}[\chi_L(X_n)]$ converge, so we have

$$\begin{aligned} \lim_n |\mathbb{E}[f(X_n) - f(X)]| \leq \\ \leq \|f\|(1 - \mathbb{E}[\chi_L(X)]) + \varepsilon + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]| \end{aligned} \tag{134}$$

which vanishes a we take $\varepsilon \to 0$ and then $L \to \infty$.

Convergence of distributions

From (a), $\mathbb{E}[f_{L,\varepsilon}(X_n)]$ and $\mathbb{E}[\chi_L(X_n)]$ converge, so we have

$$\begin{aligned} &\lim_n |\mathbb{E}[f(X_n) - f(X)]| \leq \\ &\quad \leq \|f\|(1 - \mathbb{E}[\chi_L(X)]) + \varepsilon + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]| \end{aligned} \tag{134}$$

which vanishes a we take $\varepsilon \to 0$ and then $L \to \infty$.
(b) $\Rightarrow$ (a): Trivial

Convergence of distributions

From (a), $\mathbb{E}[f_{L,\varepsilon}(X_n)]$ and $\mathbb{E}[\chi_L(X_n)]$ converge, so we have

$$\begin{aligned} &\lim_n |\mathbb{E}[f(X_n) - f(X)]| \leq \\ &\quad \leq \|f\|(1 - \mathbb{E}[\chi_L(X)]) + \varepsilon + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]| \end{aligned} \tag{134}$$

which vanishes a we take $\varepsilon \to 0$ and then $L \to \infty$.
(b) $\Rightarrow$ (a): Trivial
(b) $\Rightarrow$ (c): cos and sin are continuous bounded functions.

Convergence of distributions

From (a), $\mathbb{E}[f_{L,\varepsilon}(X_n)]$ and $\mathbb{E}[\chi_L(X_n)]$ converge, so we have

$$\begin{aligned}\lim_n |\mathbb{E}[f(X_n) - f(X)]| \leq \\ \leq \|f\|(1 - \mathbb{E}[\chi_L(X)]) + \varepsilon + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]|\end{aligned} \tag{134}$$

which vanishes a we take $\varepsilon \to 0$ and then $L \to \infty$.
(b) $\Rightarrow$ (a): Trivial
(b) $\Rightarrow$ (c): cos and sin are continuous bounded functions.
(c) $\Rightarrow$ (a): We have already seen that (by Fouries theorem) we can uniformly approximate a continuous compactly supoorted functions with a linear combination of complex exponentials. So for each $\varepsilon > 0$, there exist n and $t_1, \ldots, t_n$ and $a_1, \ldots, a_n \in \mathbb{C}$ such that $|f(x) - \sum_{k=1}^n a_k e^{it_k \cdot x}| \leq \varepsilon$. So $|E[f(X)] - \sum_k a_k \varphi_X(t_k)| \leq \varepsilon$. And the same happens for X_n. So the convergence $\mathbb{E}[f(X_n)]$ follows from a standard 2ε argument.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \liminf$.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \liminf$.
(d) $\Leftrightarrow$ (e): One statement for f, is the other one for $-f$.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \liminf$.
(d) $\Leftrightarrow$ (e): One statement for f, is the other one for $-f$.
(d) $\Rightarrow$ (f): $\mathbb{P}(O) = \mathbb{E}[1_O]$, and 1_O is lower semicontinuous if O is open.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \liminf$.
(d) $\Leftrightarrow$ (e): One statement for f, is the other one for $-f$.
(d) $\Rightarrow$ (f): $\mathbb{P}(O) = \mathbb{E}[1_O]$, and 1_O is lower semicontinuous if O is open.
(f) $\Leftrightarrow$ (g): Pass to the complement $C = O^c$.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \lim$.
(d) $\Leftrightarrow$ (e): One statement for f, is the other one for $-f$.
(d) $\Rightarrow$ (f): $\mathbb{P}(O) = \mathbb{E}[1_O]$, and 1_O is lower semicontinuous if O is open.
(f) $\Leftrightarrow$ (g): Pass to the complement $C = O^c$.
(f),(g) $\Rightarrow$ (h): Since $(-\infty, x]$ is closed, (g) implies $\overline{\lim} F_n(x) \leq F(x)$. However if x is a point of continuity of F, then $F(x) = \mathbb{P}(X \in (-\infty, x))$, so we get the other inequality from (f), since $(-\infty, x)$ is open.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \lim$.
(d) $\Leftrightarrow$ (e): One statement for f, is the other one for $-f$.
(d) $\Rightarrow$ (f): $\mathbb{P}(O) = \mathbb{E}[1_O]$, and 1_O is lower semicontinuous if O is open.
(f) $\Leftrightarrow$ (g): Pass to the complement $C = O^c$.
(f),(g) $\Rightarrow$ (h): Since $(-\infty, x]$ is closed, (g) implies $\overline{\lim} F_n(x) \leq F(x)$. However if x is a point of continuity of F, then $F(x) = \mathbb{P}(X \in (-\infty, x))$, so we get the other inequality from (f), since $(-\infty, x)$ is open.
(h) $\Rightarrow$ (a): Take f smooth and compactly supported. We have $\mathbb{E}[f(X_n)] = \int f'(x) F_n(x) dx$. Since f' is also compactly supported and bounded, and $F(x)$ is increasing and thus continuous almost everywhere, we can pass to the limit inside the integral by Lebesgue dominated convergence.

Limits in distribution

Notice that convergence in distribution cannot imply any of the previous convergences. Take for instance of $X \sim \mathcal{N}(0,1)$, and $X_n = (-1)^n X$. Then $X_n \to X$ in distribution, but clearly X_n does not convergence as a function.

However convergence in distribution is weaker than any other, and many times indeed is the only type of convergence that takes place:

Limits in distribution

Notice that convergence in distribution cannot imply any of the previous convergences. Take for instance of $X \sim \mathcal{N}(0,1)$, and $X_n = (-1)^n X$. Then $X_n \to X$ in distribution, but clearly X_n does not convergence as a function.

However convergence in distribution is weaker than any other, and many times indeed is the only type of convergence that takes place:

Proposition

If $X_n \to X$ in probability, then $X_n \to X$ in distribution.

If $X_n \to X$ in distribution and X is constant a.s. $\mathbb{P}(X = m) = 1$ for some $m \in \mathbb{R}^d$, then $X_n \to X$ in probability.

Limits in distribution

Notice that convergence in distribution cannot imply any of the previous convergences. Take for instance of $X \sim \mathcal{N}(0,1)$, and $X_n = (-1)^n X$. Then $X_n \to X$ in distribution, but clearly X_n does not convergence as a function.

However convergence in distribution is weaker than any other, and many times indeed is the only type of convergence that takes place:

Proposition

If $X_n \to X$ in probability, then $X_n \to X$ in distribution.
If $X_n \to X$ in distribution and X is constant a.s. $\mathbb{P}(X = m) = 1$ for some $m \in \mathbb{R}^d$, then $X_n \to X$ in probability.

Notice that, even if the X_n are all defined on different probability spaces, the last statement makes sense since the event $\{d(X_n, m) > \varepsilon\}$.

Limits in distribution

Notice that convergence in distribution cannot imply any of the previous convergences. Take for instance of $X \sim \mathcal{N}(0,1)$, and $X_n = (-1)^n X$. Then $X_n \to X$ in distribution, but clearly X_n does not convergence as a function.

However convergence in distribution is weaker than any other, and many times indeed is the only type of convergence that takes place:

Proposition

If $X_n \to X$ in probability, then $X_n \to X$ in distribution.
If $X_n \to X$ in distribution and X is constant a.s. $\mathbb{P}(X = m) = 1$ for some $m \in \mathbb{R}^d$, then $X_n \to X$ in probability.

Notice that, even if the X_n are all defined on different probability spaces, the last statement makes sense since the event $\{d(X_n, m) > \varepsilon\}$.

Proof: Smooth compactly suppoorted functions are Lipschitz. So for each $\varepsilon > 0$

$$|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \leq \mathbb{E}[|f(X_n) - f(X)|] \leq C\mathbb{E}[\min(1, d(X_n, X))]$$
$$\leq C(\varepsilon \mathbb{P}(d(X_n, X) \leq \varepsilon) + \mathbb{P}(d(X_n, X) \geq \varepsilon))$$

Limits in distribution

For the other implication, let $f(x)$ be a continuous function $0 \leq f \leq 1$ that equals 1 in a ball of radius $\varepsilon/2$ centered in x, and vanishes for $d(x, m) \geq \varepsilon$. Then
$\mathbb{P}(d(X_n, X) \leq \varepsilon) \geq \mathbb{E}[f(X_n)] \to \mathbb{E}[f(m)] = 1.$

Central limit theorem

Here a celebrated result.

Let X_i be a sequence of i.i.d. random variables with mean m and variance $\sigma^2 < \infty$. Let

$$Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - m) \tag{136}$$

Then, as $n \to \infty$, Z_n converges in distribution to a $\mathcal{N}(0, \sigma^2)$.

Central limit theorem

Here a celebrated result.

Let X_i be a sequence of i.i.d. random variables with mean m and variance $\sigma^2 < \infty$. Let

$$Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - m) \tag{136}$$

Then, as $n \to \infty$, Z_n converges in distribution to a $\mathcal{N}(0, \sigma^2)$.

We need to prove that for every $t \in \mathbb{R}$, $\mathbb{E}[e^{itZ_n}] \to \exp(-\sigma^2 t^2/2)$.

Indeed by independence and identical distribution of the X_i

$$\mathbb{E}[e^{itZ_n}] = \prod_{j=1}^{n} \mathbb{E}[e^{i\frac{t}{\sqrt{n}}(X_j - m)}] = \mathbb{E}[e^{i\frac{t}{\sqrt{n}}(X_1 - m)}]^n \tag{137}$$

Since $|e^{iz} - (1 + iz - z^2/2)| \leq \min(|z|^3, |z|^2)$, we have

$$\mathbb{E}[e^{itZ_n}] = \mathbb{E}[1 + i\frac{it}{\sqrt{n}}(X_1 - m) - \frac{t^2}{2n}(X_1 - m)^2 + R]^n \tag{138}$$

Central limit theorem

where
$\mathbb{E}[R] \leq \mathbb{E}[\min(t^2(X_1 - m)^2/n, |t|^3|X_1 - m|^2 n^{-3/2})] = o(1/n)$.
Therefore

$$\mathbb{E}[e^{itZ_n}] = (1 - \frac{t^2}{n2\sigma^2} + o(1/n))^n \to \exp(-t^2\sigma^2/2) \tag{139}$$

which concludes the proof.

Central limit theorem

Let $Z \in \mathcal{N}(0, \sigma^2)$. Since $F(x) := \mathbb{P}(Z \leq x)$ is continuous, we just proved that $\mathbb{P}(Z_n \in [a, b]) \to \mathbb{P}(Z \in [a, b])$ for all a, b. Acutally, it is an immediate consequence of this that the limit is uniform

$$\lim_n \sup_{a,b} |\mathbb{P}(Z_n \in [a, b]) - \mathbb{P}(Z \in [a, b])| = 0 \tag{140}$$

Central limit theorem

Let $Z \in \mathcal{N}(0, \sigma^2)$. Since $F(x) := \mathbb{P}(Z \leq x)$ is continuous, we just proved that $\mathbb{P}(Z_n \in [a, b]) \to \mathbb{P}(Z \in [a, b])$ for all a, b. Acutally, it is an immediate consequence of this that the limit is uniform

$$\lim_n \sup_{a,b} |\mathbb{P}(Z_n \in [a, b]) - \mathbb{P}(Z \in [a, b])| = 0 \tag{140}$$

This explains why we model phenomena that are influenced by many independent random factors as Gaussians.

Central limit theorem

Let $Z \in \mathcal{N}(0, \sigma^2)$. Since $F(x) := \mathbb{P}(Z \leq x)$ is continuous, we just proved that $\mathbb{P}(Z_n \in [a, b]) \to \mathbb{P}(Z \in [a, b])$ for all a, b. Acutally, it is an immediate consequence of this that the limit is uniform

$$\lim_n \sup_{a,b} |\mathbb{P}(Z_n \in [a, b]) - \mathbb{P}(Z \in [a, b])| = 0 \tag{140}$$

This explains why we model phenomena that are influenced by many independent random factors as Gaussians.
However, we may want to be more quantitative on the convergence, in particular for averages of functions.

Central limit theorem

We prove a last theorem.

Proposition

Let Y_i be a sequence of i.i.d. random variables such that $\mathbb{E}[Y_i] = 0$, $\mathbb{E}[Y_i^2] = 1$ and $\mathbb{E}[|Y_i|^3] < \infty$. Let W_i be another sequence with the same properties, but possibly a different distribution. Let g be a C^3 function on $\mathbb{R}$ with all bounded derivatives and set $C = \sup_x |g'''(x)|$. Then

$$\left|\mathbb{E}[g(\frac{Y_1 + \ldots + Y_n}{\sqrt{n}})] - \mathbb{E}[g(\frac{W_1 + \ldots + W_n}{\sqrt{n}})]\right| \leq \frac{C(\mathbb{E}[|Y_1|^3] + \mathbb{E}[|W_1|^3]}{6\sqrt{n}} \tag{141}$$

Central limit theorem

We prove a last theorem.

Proposition

Let Y_i be a sequence of i.i.d. random variables such that $\mathbb{E}[Y_i] = 0$, $\mathbb{E}[Y_i^2] = 1$ and $\mathbb{E}[|Y_i|^3] < \infty$. Let W_i be another sequence with the same properties, but possibly a different distribution. Let g be a C^3 function on $\mathbb{R}$ with all bounded derivatives and set $C = \sup_x |g'''(x)|$. Then

$$\left|\mathbb{E}[g(\frac{Y_1 + \ldots + Y_n}{\sqrt{n}})] - \mathbb{E}[g(\frac{W_1 + \ldots + W_n}{\sqrt{n}})]\right| \leq \frac{C(\mathbb{E}[|Y_1|^3] + \mathbb{E}[|W_1|^3]}{6\sqrt{n}} \tag{141}$$

Let us prove it. Set $V_k = (Y_1 + \ldots + Y_k + W_{k+2} + \ldots W_n)/\sqrt{n}$. Then by telescopic sum, the right hand side in the above formula is

$$\sum_{k=0}^{n-1} \mathbb{E}[g(V_k + Y_{k+1}/\sqrt{n})] - \mathbb{E}[g(V_k + W_{k+1}/\sqrt{n})] \tag{142}$$

Central limit theorem

We claim that each term in the sum is bounded by $\frac{C(\mathbb{E}[|Y_1|^3]+\mathbb{E}[|W_1|^3])}{6n\sqrt{n}}$, which concludes the proof. This follows from the next general lemma.

Lemma

Let V, Y, W be three independent random variables such that $E(Y) = E(Z)$, $E(Y^2) = E(Z^2)$ Then with $C := sup_x|g'''(x)|$

$$|E[g(V+Y)] - E[g(V+W)]| \leq \frac{C(\mathbb{E}[|Y|^3] + \mathbb{E}[|W|^3])}{6} \quad (143)$$

Central limit theorem

We claim that each term in the sum is bounded by $\frac{C(\mathbb{E}[|Y_1|^3]+\mathbb{E}[|W_1|^3])}{6n\sqrt{n}}$, which concludes the proof. This follows from the next general lemma.

Lemma

Let V, Y, W be three independent random variables such that $E(Y) = E(Z)$, $E(Y^2) = E(Z^2)$ Then with $C := sup_x|g'''(x)|$

$$|E[g(V+Y)] - E[g(V+W)]| \leq \frac{C(\mathbb{E}[|Y|^3] + \mathbb{E}[|W|^3])}{6} \quad (143)$$

Indeed by Taylor formula,
$g(v+y) - g(v+z) = g'(x)(y-z) + \frac{1}{2}g''(x)(y^2-z^2) + R$, where the remainder is bounded by $C(y^3+z^3)/6$. So we just need to compute it in $x = X$, $y = Y$, $z = Z$ and take the expected value.

Convergence of random variables

Let X be random variable and (X_n) be a sequence of random variables, all defined on the same probability space $X, X_n \colon \Omega \to E$. We think that E is a nice metric space (complete, separable), but you may well think $E = \mathbb{R}^n$ or even $E = \mathbb{R}$ if the general framework is too abstract.

Convergence of random variables

Let X be random variable and (X_n) be a sequence of random variables, all defined on the same probability space $X, X_n \colon \Omega \to E$. We think that E is a nice metric space (complete, separable), but you may well think $E = \mathbb{R}^n$ or even $E = \mathbb{R}$ if the general framework is too abstract.
We say that

- X_n converges to X *almost surely* or *with probability* 1 if $\mathbb{P}(\lim_n X_n = X)$.

Convergence of random variables

Let X be random variable and (X_n) be a sequence of random variables, all defined on the same probability space $X, X_n \colon \Omega \to E$. We think that E is a nice metric space (complete, separable), but you may well think $E = \mathbb{R}^n$ or even $E = \mathbb{R}$ if the general framework is too abstract.
We say that

- X_n converges to X *almost surely* or *with probability* 1 if $\mathbb{P}(\lim_n X_n = X)$.
- X_n converges to X *in probability* if $\lim_n \mathbb{P}(\mathrm{distance}(X_n, X) > \varepsilon) = 0$, for every $\varepsilon > 0$.

Convergence of random variables

Let X be random variable and (X_n) be a sequence of random variables, all defined on the same probability space $X, X_n \colon \Omega \to E$. We think that E is a nice metric space (complete, separable), but you may well think $E = \mathbb{R}^n$ or even $E = \mathbb{R}$ if the general framework is too abstract.
We say that

- X_n converges to X *almost surely* or *with probability* 1 if $\mathbb{P}(\lim_n X_n = X)$.
- X_n converges to X *in probability* if $\lim_n \mathbb{P}(\mathrm{distance}(X_n, X) > \varepsilon) = 0$, for every $\varepsilon > 0$.
- In L^p, with $p \geq 1$, if $\lim_n \mathbb{E}(\mathrm{distance}(X_n, X)^p) = 0$.

Convergence of random variables

Let X be random variable and (X_n) be a sequence of random variables, all defined on the same probability space $X, X_n : \Omega \to E$. We think that E is a nice metric space (complete, separable), but you may well think $E = \mathbb{R}^n$ or even $E = \mathbb{R}$ if the general framework is too abstract.
We say that

- X_n converges to X *almost surely* or *with probability* 1 if $\mathbb{P}(\lim_n X_n = X)$.
- X_n converges to X *in probability* if $\lim_n \mathbb{P}(\mathrm{distance}(X_n, X) > \varepsilon) = 0$, for every $\varepsilon > 0$.
- In L^p, with $p \geq 1$, if $\lim_n \mathbb{E}(\mathrm{distance}(X_n, X)^p) = 0$.

This are just probabilistic notations for the 'almost everywhere', 'in measure' and 'in L^p' convergence, that you have already encountered in measure theory.

Convergence of random variables

The following relations hold between these definitions.

Convergence of random variables

The following relations hold between these definitions.

Proposition

Let X_n, X as above. Then

(a) *If $X_n \to X$ a.s., then $X_n \to X$ in probability.*

(b) *If $\sum_n \mathbb{P}(\mathrm{distance}(X_n, X) > \varepsilon) < \infty$ for every $\varepsilon > 0$, then $X_n \to X$ a.s.. In particular, if $X_n \to X$ in probability, then there exists a subsequence such that $X_{n_k} \to X$ a.s..*

(c) *If $X_n \to X$ in L^p, then $X_n \to X$ in probability.*

(d) *If $X_n \to X$ a.s. and X_n is p-uniformly integrable, then $X_n \to X$ in L^p. Here p-uniformly integrable means that for some $x \in E$*

$$\lim_{M \to \infty} \sup_n \mathbb{E}[\mathrm{distance}(X_n, x)^p 1_{\mathrm{distance}(X_n,x)^p > M}] = 0 \qquad (144)$$

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.
We also proved $\mathbb{P}(\underline{\lim}_n A_n) \leq \underline{\lim}_n \mathbb{P}(A_n)$ and the Borel-Cantelli lemma: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\overline{\lim}_n A_n) = 0$.

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.

We also proved $\mathbb{P}(\underline{\lim}_n A_n) \leq \underline{\lim}_n \mathbb{P}(A_n)$ and the Borel-Cantelli lemma: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\overline{\lim}_n A_n) = 0$.

Proof: Write $d(x, y)$ for the distance of E, and let $A_{n,k} := \{d(X_n, X) > 1/k\}$. Notice $\{\lim_n X_n = X\} = \cap_k \underline{\lim}_n A_{n,k}^c$.

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.

We also proved $\mathbb{P}(\underline{\lim}_n A_n) \leq \underline{\lim}_n \mathbb{P}(A_n)$ and the Borel-Cantelli lemma: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\overline{\lim}_n A_n) = 0$.

Proof: Write $d(x, y)$ for the distance of E, and let $A_{n,k} := \{d(X_n, X) > 1/k\}$. Notice $\{\lim_n X_n = X\} = \cap_k \underline{\lim}_n A_{n,k}^c$.

(a) If $\mathbb{P}(\lim_n X_n = X) = 1$, then for every k we have $1 = \mathbb{P}(\underline{\lim}_n A_{n,k}^c) \leq \underline{\lim}_n \mathbb{P}(A_{n,k}^c)$. So $\overline{\lim}_n \mathbb{P}(A_{n,k}) = 0$ which means convergence in probability.

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.

We also proved $\mathbb{P}(\underline{\lim}_n A_n) \leq \underline{\lim}_n \mathbb{P}(A_n)$ and the Borel-Cantelli lemma: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\overline{\lim}_n A_n) = 0$.

Proof: Write $d(x, y)$ for the distance of E, and let $A_{n,k} := \{d(X_n, X) > 1/k\}$. Notice $\{\lim_n X_n = X\} = \cap_k \underline{\lim}_n A_{n,k}^c$.

(a) If $\mathbb{P}(\lim_n X_n = X) = 1$, then for every k we have $1 = \mathbb{P}(\underline{\lim}_n A_{n,k}^c) \leq \underline{\lim}_n \mathbb{P}(A_{n,k}^c)$. So $\overline{\lim}_n \mathbb{P}(A_{n,k}) = 0$ which means convergence in probability.

(b) By Borel-Cantelli lemma, we have that $\mathbb{P}(\overline{\lim}_n A_{n,k}) = 0$ for every k. So $\mathbb{P}(\cup_k \overline{\lim}_n A_{n,k}^c) = 0$, which means convergence a.s..

Convergence of random variables

Before going to the proof, recall that in the very first lecture we defined limsup and liminf of events: $\overline{\lim}_n A_n = \cap_n \cup_{j \geq n} A_j$ and $\underline{\lim}_n A_n = \cup_n \cap_{j \geq n} A_j$.
We also proved $\mathbb{P}(\underline{\lim}_n A_n) \leq \underline{\lim}_n \mathbb{P}(A_n)$ and the Borel-Cantelli lemma: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\overline{\lim}_n A_n) = 0$.
Proof: Write $d(x, y)$ for the distance of E, and let $A_{n,k} := \{d(X_n, X) > 1/k\}$. Notice $\{\lim_n X_n = X\} = \cap_k \underline{\lim}_n A_{n,k}^c$.
(a) If $\mathbb{P}(\lim_n X_n = X) = 1$, then for every k we have $1 = \mathbb{P}(\underline{\lim}_n A_{n,k}^c) \leq \underline{\lim}_n \mathbb{P}(A_{n,k}^c)$. So $\overline{\lim}_n \mathbb{P}(A_{n,k}) = 0$ which means convergence in probability.
(b) By Borel-Cantelli lemma, we have that $\mathbb{P}(\overline{\lim}_n A_{n,k}) = 0$ for every k. So $\mathbb{P}(\cup_k \overline{\lim}_n A_{n,k}^c) = 0$, which means convergence a.s.. To deduce that convergence in probability implies convergence a.s. along subsequences, notice that we can take

$$n_j := \inf\{m : \mathbb{P}(A_{r,j}) \leq 2^{-j} \, \forall r \geq m\} \tag{145}$$

This n_j exists since $\mathbb{P}(A_{n,k}) \to 0$ for every k. Moreover it is immediate to see the series $\sum_j \mathbb{P}(A_{n_j,k})$ converges.

Convergence of random variables

(c) This is just Markov inequality
$\mathbb{P}(d(X_n, X) \geq \varepsilon) \leq \mathbb{E}[d(X_n, X)^p]\varepsilon^{-p}$.

Convergence of random variables

(c) This is just Markov inequality
$\mathbb{P}(d(X_n, X) \geq \varepsilon) \leq \mathbb{E}[d(X_n, X)^p]\varepsilon^{-p}$.
(d) By triangular inequality,
$d(X_n, X)^p \leq c_p(d(X_n, x)^p + d(X, x)^p)$. So it is also uniformly integrable and we conclude by Lebesgue convergence.

Convergence of random variables

Proposition (Law of large numbers)

Let X_i be an i.i.d. sequence of real random variables with finite expectation $m = \mathbb{E}[X_i]$. Define

$$S_n := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{146}$$

Suppose that $\mathbb{E}[X_i^2] < \infty$. Then $S_n \to m$ in probability.
Suppose that $\mathbb{E}[X_i^4] < \infty$. Then $S_n \to m$ a.s.

Convergence of random variables

Proposition (Law of large numbers)

Let X_i be an i.i.d. sequence of real random variables with finite expectation $m = \mathbb{E}[X_i]$. Define

$$S_n := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{146}$$

Suppose that $\mathbb{E}[X_i^2] < \infty$. Then $S_n \to m$ in probability.
Suppose that $\mathbb{E}[X_i^4] < \infty$. Then $S_n \to m$ a.s.

Actually a much stronger result holds, due to Kolmogorov: *Suppose that $\mathbb{E}[|X_i|] < \infty$. Then $S_n \to m$ a.s.*.

Convergence of random variables

Proposition (Law of large numbers)

Let X_i be an i.i.d. sequence of real random variables with finite expectation $m = \mathbb{E}[X_i]$. Define

$$S_n := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{146}$$

Suppose that $\mathbb{E}[X_i^2] < \infty$. Then $S_n \to m$ in probability.
Suppose that $\mathbb{E}[X_i^4] < \infty$. Then $S_n \to m$ a.s.

Actually a much stronger result holds, due to Kolmogorov: *Suppose that $\mathbb{E}[|X_i|] < \infty$. Then $S_n \to m$ a.s.*.
We however prove the weaker version. For the first statement, note that $\mathbb{E}[S_n] = m$ and $\mathbb{D}(S_n) = \mathbb{D}(X_i)/n$. So

$$\mathbb{P}(|S_n - m| \geq \varepsilon) \leq \frac{\mathbb{D}(S_n)}{\varepsilon^2} = \frac{\mathbb{D}(X_i)}{n\varepsilon^2} \to 0 \tag{147}$$

namely convergence in probability.

Convergence of random variables

If now X_i has four moments, using the point (b) of the previous Proposition, we need to show that

$$\sum_n \mathbb{P}(|S_n - m| \geq \varepsilon) < \infty \tag{148}$$

Indeed $\mathbb{P}(|S_n - m| \geq \varepsilon) \leq \mathbb{E}[|S_n - m|^4]\varepsilon^{-4}$. But

$$\begin{aligned}\mathbb{E}[|S_n - m|^4] &= \tfrac{1}{n^4}\sum_i \mathbb{E}[(X_i - m)^4] + \tfrac{1}{n^4}\sum_{i\neq j} \mathbb{E}[(X_i - m)^2]\mathbb{E}[(X_j - m)^2] \\ &= \tfrac{1}{n^3}\mathbb{E}[(X_1 - m)^4] + \frac{n(n-1)}{n^4}\mathbb{E}[(X_1 - m)^2]^2\end{aligned} \tag{149}$$

which is summable.

Convergence of random variables

If now X_i has four moments, using the point (b) of the previous Proposition, we need to show that

$$\sum_n \mathbb{P}(|S_n - m| \geq \varepsilon) < \infty \tag{148}$$

Indeed $\mathbb{P}(|S_n - m| \geq \varepsilon) \leq \mathbb{E}[|S_n - m|^4]\varepsilon^{-4}$. But

$$\begin{aligned}\mathbb{E}[|S_n - m|^4] &= \tfrac{1}{n^4}\sum_i \mathbb{E}[(X_i - m)^4] + \tfrac{1}{n^4}\sum_{i\neq j} \mathbb{E}[(X_i - m)^2]\mathbb{E}[(X_j - m)^2] \\ &= \tfrac{1}{n^3}\mathbb{E}[(X_1 - m)^4] + \frac{n(n-1)}{n^4}\mathbb{E}[(X_1 - m)^2]^2\end{aligned} \tag{149}$$

which is summable. This Proposition tells us that as we keep playing head and tail, the average number of heads will converge to $1/2$. It is very intuitive.

Convergence of distributions

We have stressed many times that often one is not interested in a random variable, but just in its distribution. So we are going to define a most useful notion of convergence. We take $E = \mathbb{R}^d$ here, since dealing with non-locally compact spaces requires a slightly more delicate definition.

Convergence of distributions

We have stressed many times that often one is not interested in a random variable, but just in its distribution. So we are going to define a most useful notion of convergence. We take $E = \mathbb{R}^d$ here, since dealing with non-locally compact spaces requires a slightly more delicate definition.

E comes equippes with the Borel σ-algebra. We then consider the following topology on $\mathcal{P}(E)$ (the space of probability measures on E): the weakest topology such that the map $\mu \mapsto \int f \, d\mu$ is continuous for every continuous, compactly supported function $f : E \to \mathbb{R}$.

Convergence of distributions

We have stressed many times that often one is not interested in a random variable, but just in its distribution. So we are going to define a most useful notion of convergence. We take $E = \mathbb{R}^d$ here, since dealing with non-locally compact spaces requires a slightly more delicate definition.

E comes equippes with the Borel σ-algebra. We then consider the following topology on $\mathcal{P}(E)$ (the space of probability measures on E): the weakest topology such that the map $\mu \mapsto \int f \, d\mu$ is continuous for every continuous, compactly supported function $f : E \to \mathbb{R}$.

Then we will say that $X_n \to X$ *in law* or in distribution if the distribution of X_n converges to the distribution of X.

Convergence of distributions

We have stressed many times that often one is not interested in a random variable, but just in its distribution. So we are going to define a most useful notion of convergence. We take $E = \mathbb{R}^d$ here, since dealing with non-locally compact spaces requires a slightly more delicate definition.

E comes equippes with the Borel σ-algebra. We then consider the following topology on $\mathcal{P}(E)$ (the space of probability measures on E): the weakest topology such that the map $\mu \mapsto \int f \, d\mu$ is continuous for every continuous, compactly supported function $f : E \to \mathbb{R}$.

Then we will say that $X_n \to X$ *in law* or in distribution if the distribution of X_n converges to the distribution of X.

To avoid a too abstract formulation, let us give a less intrinsic but more operative definition.

Convergence of distributions

Proposition

Let X_n, X be $\mathbb{R}^d$-random variables (possibly each defined on a different probability space). The following are equivalent

(a) $\lim_n \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ *for every C^∞, compactly supported function $f : \mathbb{R}^d \to \mathbb{R}$.*

(b) $\lim_n \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ *for every continuous, bounded function $f : \mathbb{R}^d \to \mathbb{R}$*

(c) $\lim_n \varphi_{X_n}(t) = \varphi_X(t)$ *for every $t \in \mathbb{R}^d$ (pointwise convergence of the characteristic function).*

(d) $\underline{\lim}_n \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$ *for every lower semicontinuous function $f : \mathbb{R}^d \to \mathbb{R}$.*

(e) $\underline{\lim}_n \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$ *for every upper semicontinuous function $f : \mathbb{R}^d \to \mathbb{R}$.*

Convergence of distributions

(f) $\underline{\lim}_n \mathbb{E}[1_O] \geq \mathbb{E}[1_O]$ for every open set O.

(g) $\overline{\lim}_n \mathbb{E}[1_C] \leq \mathbb{E}[1_C]$ for every open set C.

(h) In $d = 1$, the cumulative distribution function $F_n(x) := \mathbb{P}(X_n \leq x)$ converges (pointwise) to $F(x) := \mathbb{P}(X \leq x)$ at every point x where F is continuous.

Convergence of distributions

(f) $\underline{\lim}_n \mathbb{E}[1_O] \geq \mathbb{E}[1_O]$ for every open set O.

(g) $\overline{\lim}_n \mathbb{E}[1_C] \leq \mathbb{E}[1_C]$ for every open set C.

(h) In $d = 1$, the cumulative distribution function $F_n(x) := \mathbb{P}(X_n \leq x)$ converges (pointwise) to $F(x) := \mathbb{P}(X \leq x)$ at every point x where F is continuous.

Proof: (a) $\Rightarrow$ (b) Assume (a). For $L > 0$, let χ_L be a C^∞ function $0 \leq \chi_L \leq 1$ such that that $\chi_L(x) = 1$ for $|x| \leq L$ and $\chi_L(x) = 0$ for $|x| \geq L$. If f is continuous and bounded, $f\chi_L$ is also compactly supported, and for every $L > 0$ and $\varepsilon > 0$ we can find a C^∞, compactly supported $f_{L,\varepsilon}$ such that $|f\chi_L - f_{L,\varepsilon}| \leq \varepsilon$. Then

$$
\begin{aligned}
|\mathbb{E}[f(X_n) - f(X)]| &\leq |\mathbb{E}[f(X_n)(1-\chi_L)(X_n)]| + |\mathbb{E}[(f\chi_L)(X_n) - f_{L,\varepsilon}(X_n)]| \\
&\quad + |\mathbb{E}[f_{L,\varepsilon}(X_n) - f_{L,\varepsilon}(X)]| + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]| \\
&\leq \|f\|(1 - \mathbb{E}[\chi_L(X_n)]) + \varepsilon \\
&\quad + |\mathbb{E}[f_{L,\varepsilon}(X_n) - f_{L,\varepsilon}(X)]| + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]|
\end{aligned}
\tag{150}
$$

Convergence of distributions

From (a), $\mathbb{E}[f_{L,\varepsilon}(X_n)]$ and $\mathbb{E}[\chi_L(X_n)]$ converge, so we have

$$\begin{aligned} &\lim_n |\mathbb{E}[f(X_n) - f(X)]| \leq \\ &\quad \leq \|f\|(1 - \mathbb{E}[\chi_L(X)]) + \varepsilon + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]| \end{aligned} \tag{151}$$

which vanishes a we take $\varepsilon \to 0$ and then $L \to \infty$.

Convergence of distributions

From (a), $\mathbb{E}[f_{L,\varepsilon}(X_n)]$ and $\mathbb{E}[\chi_L(X_n)]$ converge, so we have

$$\begin{aligned} &\lim_n |\mathbb{E}[f(X_n) - f(X)]| \leq \\ &\quad \leq \|f\|(1 - \mathbb{E}[\chi_L(X)]) + \varepsilon + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]| \end{aligned} \tag{151}$$

which vanishes a we take $\varepsilon \to 0$ and then $L \to \infty$.
(b) $\Rightarrow$ (a): Trivial

Convergence of distributions

From (a), $\mathbb{E}[f_{L,\varepsilon}(X_n)]$ and $\mathbb{E}[\chi_L(X_n)]$ converge, so we have

$$\begin{aligned} &\lim_n |\mathbb{E}[f(X_n) - f(X)]| \leq \\ &\quad \leq \|f\|(1 - \mathbb{E}[\chi_L(X)]) + \varepsilon + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]| \end{aligned} \tag{151}$$

which vanishes a we take $\varepsilon \to 0$ and then $L \to \infty$.
(b) $\Rightarrow$ (a): Trivial
(b) $\Rightarrow$ (c): cos and sin are continuous bounded functions.

Convergence of distributions

From (a), $\mathbb{E}[f_{L,\varepsilon}(X_n)]$ and $\mathbb{E}[\chi_L(X_n)]$ converge, so we have

$$\begin{aligned}\lim_n |\mathbb{E}[f(X_n) - f(X)]| \leq \\ \leq \|f\|(1 - \mathbb{E}[\chi_L(X)]) + \varepsilon + |\mathbb{E}[f_{L,\varepsilon}(X) - f(X)]|\end{aligned} \tag{151}$$

which vanishes a we take $\varepsilon \to 0$ and then $L \to \infty$.
(b) $\Rightarrow$ (a): Trivial
(b) $\Rightarrow$ (c): cos and sin are continuous bounded functions.
(c) $\Rightarrow$ (a): We have already seen that (by Fouries theorem) we can uniformly approximate a continuous compactly supoorted functions with a linear combination of complex exponentials. So for each $\varepsilon > 0$, there exist n and $t_1, \ldots, t_n$ and $a_1, \ldots, a_n \in \mathbb{C}$ such that $|f(x) - \sum_{k=1}^n a_k e^{it_k \cdot x}| \leq \varepsilon$. So $|E[f(X)] - \sum_k a_k \varphi_X(t_k)| \leq \varepsilon$. And the same happens for X_n. So the convergence $\mathbb{E}[f(X_n)]$ follows from a standard 2ε argument.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \liminf$.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \liminf$.
(d) $\Leftrightarrow$ (e): One statement for f, is the other one for $-f$.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \liminf$.
(d) $\Leftrightarrow$ (e): One statement for f, is the other one for $-f$.
(d) $\Rightarrow$ (f): $\mathbb{P}(O) = \mathbb{E}[1_O]$, and 1_O is lower semicontinuous if O is open.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \liminf$.
(d) $\Leftrightarrow$ (e): One statement for f, is the other one for $-f$.
(d) $\Rightarrow$ (f): $\mathbb{P}(O) = \mathbb{E}[1_O]$, and 1_O is lower semicontinuous if O is open.
(f) $\Leftrightarrow$ (g): Pass to the complement $C = O^c$.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \lim$.
(d) $\Leftrightarrow$ (e): One statement for f, is the other one for $-f$.
(d) $\Rightarrow$ (f): $\mathbb{P}(O) = \mathbb{E}[1_O]$, and 1_O is lower semicontinuous if O is open.
(f) $\Leftrightarrow$ (g): Pass to the complement $C = O^c$.
(f),(g) $\Rightarrow$ (h): Since $(-\infty, x]$ is closed, (g) implies $\overline{\lim} F_n(x) \leq F(x)$. However if x is a point of continuity of F, then $F(x) = \mathbb{P}(X \in (-\infty, x))$, so we get the other inequality from (f), since $(-\infty, x)$ is open.

Convergence of distributions

(a) $\Rightarrow$ (d): Lower semicontinuous functions are supremum of continuous bounded functions. $f(x) = \sup_{g \in C_b} g(x)$. So it is enough to say that $\limsup \geq \sup \lim$.
(d) $\Leftrightarrow$ (e): One statement for f, is the other one for $-f$.
(d) $\Rightarrow$ (f): $\mathbb{P}(O) = \mathbb{E}[1_O]$, and 1_O is lower semicontinuous if O is open.
(f) $\Leftrightarrow$ (g): Pass to the complement $C = O^c$.
(f),(g) $\Rightarrow$ (h): Since $(-\infty, x]$ is closed, (g) implies $\overline{\lim} F_n(x) \leq F(x)$. However if x is a point of continuity of F, then $F(x) = \mathbb{P}(X \in (-\infty, x))$, so we get the other inequality from (f), since $(-\infty, x)$ is open.
(h) $\Rightarrow$ (a): Take f smooth and compactly supported. We have $\mathbb{E}[f(X_n)] = \int f'(x) F_n(x) dx$. Since f' is also compactly supported and bounded, and $F(x)$ is increasing and thus continuous almost everywhere, we can pass to the limit inside the integral by Lebesgue dominated convergence.

Limits in distribution

Notice that convergence in distribution cannot imply any of the previous convergences. Take for instance of $X \sim \mathcal{N}(0,1)$, and $X_n = (-1)^n X$. Then $X_n \to X$ in distribution, but clearly X_n does not convergence as a function.

However convergence in distribution is weaker than any other, and many times indeed is the only type of convergence that takes place:

Limits in distribution

Notice that convergence in distribution cannot imply any of the previous convergences. Take for instance of $X \sim \mathcal{N}(0,1)$, and $X_n = (-1)^n X$. Then $X_n \to X$ in distribution, but clearly X_n does not convergence as a function.

However convergence in distribution is weaker than any other, and many times indeed is the only type of convergence that takes place:

Proposition

If $X_n \to X$ in probability, then $X_n \to X$ in distribution.

If $X_n \to X$ in distribution and X is constant a.s. $\mathbb{P}(X = m) = 1$ for some $m \in \mathbb{R}^d$, then $X_n \to X$ in probability.

Limits in distribution

Notice that convergence in distribution cannot imply any of the previous convergences. Take for instance of $X \sim \mathcal{N}(0,1)$, and $X_n = (-1)^n X$. Then $X_n \to X$ in distribution, but clearly X_n does not convergence as a function.

However convergence in distribution is weaker than any other, and many times indeed is the only type of convergence that takes place:

Proposition

If $X_n \to X$ in probability, then $X_n \to X$ in distribution.
If $X_n \to X$ in distribution and X is constant a.s. $\mathbb{P}(X = m) = 1$ for some $m \in \mathbb{R}^d$, then $X_n \to X$ in probability.

Notice that, even if the X_n are all defined on different probability spaces, the last statement makes sense since the event $\{d(X_n, m) > \varepsilon\}$.

Limits in distribution

Notice that convergence in distribution cannot imply any of the previous convergences. Take for instance of $X \sim \mathcal{N}(0,1)$, and $X_n = (-1)^n X$. Then $X_n \to X$ in distribution, but clearly X_n does not convergence as a function.

However convergence in distribution is weaker than any other, and many times indeed is the only type of convergence that takes place:

Proposition

If $X_n \to X$ in probability, then $X_n \to X$ in distribution.
If $X_n \to X$ in distribution and X is constant a.s. $\mathbb{P}(X = m) = 1$ for some $m \in \mathbb{R}^d$, then $X_n \to X$ in probability.

Notice that, even if the X_n are all defined on different probability spaces, the last statement makes sense since the event $\{d(X_n, m) > \varepsilon\}$.

Proof: Smooth compactly suppoorted functions are Lipschitz. So for each $\varepsilon > 0$

$$|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \leq \mathbb{E}[|f(X_n) - f(X)|] \leq C\mathbb{E}[\min(1, d(X_n, X))]$$
$$\leq C(\varepsilon \mathbb{P}(d(X_n, X) \leq \varepsilon) + \mathbb{P}(d(X_n, X) \geq \varepsilon))$$

Limits in distribution

For the other implication, let $f(x)$ be a continuous function $0 \leq f \leq 1$ that equals 1 in a ball of radius $\varepsilon/2$ centered in x, and vanishes for $d(x, m) \geq \varepsilon$. Then
$\mathbb{P}(d(X_n, X) \leq \varepsilon) \geq \mathbb{E}[f(X_n)] \to \mathbb{E}[f(m)] = 1.$

Central limit theorem

Here a celebrated result.

Let X_i be a sequence of i.i.d. random variables with mean m and variance $\sigma^2 < \infty$. Let

$$Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - m) \tag{153}$$

Then, as $n \to \infty$, Z_n converges in distribution to a $\mathcal{N}(0, \sigma^2)$.

Central limit theorem

Here a celebrated result.

Let X_i be a sequence of i.i.d. random variables with mean m and variance $\sigma^2 < \infty$. Let

$$Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - m) \tag{153}$$

Then, as $n \to \infty$, Z_n converges in distribution to a $\mathcal{N}(0, \sigma^2)$.

We need to prove that for every $t \in \mathbb{R}$, $\mathbb{E}[e^{itZ_n}] \to \exp(-\sigma^2 t^2/2)$.

Indeed by independence and identical distribution of the X_i

$$\mathbb{E}[e^{itZ_n}] = \prod_{j=1}^{n} \mathbb{E}[e^{i\frac{t}{\sqrt{n}}(X_j - m)}] = \mathbb{E}[e^{i\frac{t}{\sqrt{n}}(X_1 - m)}]^n \tag{154}$$

Since $|e^{iz} - (1 + iz - z^2/2)| \leq \min(|z|^3, |z|^2)$, we have

$$\mathbb{E}[e^{itZ_n}] = \mathbb{E}[1 + i\frac{it}{\sqrt{n}}(X_1 - m) - \frac{t^2}{2n}(X_1 - m)^2 + R]^n \tag{155}$$

Central limit theorem

where
$\mathbb{E}[R] \leq \mathbb{E}[\min(t^2(X_1 - m)^2/n, |t|^3|X_1 - m|^2 n^{-3/2})] = o(1/n)$.
Therefore

$$\mathbb{E}[e^{itZ_n}] = (1 - \frac{t^2}{n2\sigma^2} + o(1/n))^n \to \exp(-t^2\sigma^2/2) \tag{156}$$

which concludes the proof.

Central limit theorem

Let $Z \in \mathcal{N}(0, \sigma^2)$. Since $F(x) := \mathbb{P}(Z \leq x)$ is continuous, we just proved that $\mathbb{P}(Z_n \in [a, b]) \to \mathbb{P}(Z \in [a, b])$ for all a, b. Acutally, it is an immediate consequence of this that the limit is uniform

$$\lim_n \sup_{a,b} |\mathbb{P}(Z_n \in [a, b]) - \mathbb{P}(Z \in [a, b])| = 0 \tag{157}$$

Central limit theorem

Let $Z \in \mathcal{N}(0, \sigma^2)$. Since $F(x) := \mathbb{P}(Z \leq x)$ is continuous, we just proved that $\mathbb{P}(Z_n \in [a, b]) \to \mathbb{P}(Z \in [a, b])$ for all a, b. Acutally, it is an immediate consequence of this that the limit is uniform

$$\lim_n \sup_{a,b} |\mathbb{P}(Z_n \in [a, b]) - \mathbb{P}(Z \in [a, b])| = 0 \tag{157}$$

This explains why we model phenomena that are influenced by many independent random factors as Gaussians.

Central limit theorem

Let $Z \in \mathcal{N}(0, \sigma^2)$. Since $F(x) := \mathbb{P}(Z \leq x)$ is continuous, we just proved that $\mathbb{P}(Z_n \in [a, b]) \to \mathbb{P}(Z \in [a, b])$ for all a, b. Acutally, it is an immediate consequence of this that the limit is uniform

$$\lim_n \sup_{a,b} |\mathbb{P}(Z_n \in [a, b]) - \mathbb{P}(Z \in [a, b])| = 0 \tag{157}$$

This explains why we model phenomena that are influenced by many independent random factors as Gaussians.
However, we may want to be more quantitative on the convergence, in particular for averages of functions.

Central limit theorem

We prove a last theorem.

Proposition

Let Y_i be a sequence of i.i.d. random variables such that $\mathbb{E}[Y_i] = 0$, $\mathbb{E}[Y_i^2] = 1$ and $\mathbb{E}[|Y_i|^3] < \infty$. Let W_i be another sequence with the same properties, but possibly a different distribution. Let g be a C^3 function on $\mathbb{R}$ with all bounded derivatives and set $C = \sup_x |g'''(x)|$. Then

$$\left|\mathbb{E}[g(\frac{Y_1 + \ldots + Y_n}{\sqrt{n}})] - \mathbb{E}[g(\frac{W_1 + \ldots + W_n}{\sqrt{n}})]\right| \leq \frac{C(\mathbb{E}[|Y_1|^3] + \mathbb{E}[|W_1|^3]}{6\sqrt{n}} \tag{158}$$

Central limit theorem

We prove a last theorem.

Proposition

Let Y_i be a sequence of i.i.d. random variables such that $\mathbb{E}[Y_i] = 0$, $\mathbb{E}[Y_i^2] = 1$ and $\mathbb{E}[|Y_i|^3] < \infty$. Let W_i be another sequence with the same properties, but possibly a different distribution. Let g be a C^3 function on $\mathbb{R}$ with all bounded derivatives and set $C = \sup_x |g'''(x)|$. Then

$$\left|\mathbb{E}[g(\frac{Y_1 + \ldots + Y_n}{\sqrt{n}})] - \mathbb{E}[g(\frac{W_1 + \ldots + W_n}{\sqrt{n}})]\right| \leq \frac{C(\mathbb{E}[|Y_1|^3] + \mathbb{E}[|W_1|^3]}{6\sqrt{n}} \tag{158}$$

Let us prove it. Set $V_k = (Y_1 + \ldots + Y_k + W_{k+2} + \ldots W_n)/\sqrt{n}$. Then by telescopic sum, the right hand side in the above formula is

$$\sum_{k=0}^{n-1} \mathbb{E}[g(V_k + Y_{k+1}/\sqrt{n})] - \mathbb{E}[g(V_k + W_{k+1}/\sqrt{n})] \tag{159}$$

Central limit theorem

We claim that each term in the sum is bounded by $\frac{C(\mathbb{E}[|Y_1|^3]+\mathbb{E}[|W_1|^3])}{6n\sqrt{n}}$, which concludes the proof. This follows from the next general lemma.

Lemma

Let V, Y, W be three independent random variables such that $E(Y) = E(Z)$, $E(Y^2) = E(Z^2)$ Then with $C := sup_x|g'''(x)|$

$$|E[g(V+Y)] - E[g(V+W)]| \leq \frac{C(\mathbb{E}[|Y|^3] + \mathbb{E}[|W|^3])}{6} \quad (160)$$

Central limit theorem

We claim that each term in the sum is bounded by $\frac{C(\mathbb{E}[|Y_1|^3]+\mathbb{E}[|W_1|^3])}{6n\sqrt{n}}$, which concludes the proof. This follows from the next general lemma.

Lemma

Let V, Y, W be three independent random variables such that $E(Y) = E(Z)$, $E(Y^2) = E(Z^2)$ Then with $C := sup_x|g'''(x)|$

$$|E[g(V+Y)] - E[g(V+W)]| \leq \frac{C(\mathbb{E}[|Y|^3] + \mathbb{E}[|W|^3])}{6} \quad (160)$$

Indeed by Taylor formula,
$g(v+y) - g(v+z) = g'(x)(y-z) + \frac{1}{2}g''(x)(y^2-z^2) + R$, where the remainder is bounded by $C(y^3+z^3)/6$. So we just need to compute it in $x = X$, $y = Y$, $z = Z$ and take the expected value.