Math 6710. Probability Theory I

Taught by Lionel Levine

Notes by Linus Setiabrata

This course is an introduction to probability theory. Please let me know if you spot any mistakes! There are probably lots of typos. Things in [blue font square brackets] are personal comments. Things in [red font square brackets] are (important) announcements.

Theorem numbering unchanged since Dec 9, 2019. Last compiled August 31, 2020.

Contents

1 Measure theory preliminaries

1.1 Sep 4, 2019

Our main book, Durrett's *Probability theory and examples* (5th ed) is online on his webpage. We'll cover chapters 2, 3, and a little of 4 and 5. Another useful book is Williams' *Probability with martingales*. [This week Prof Levine's office hours will be on Thursday at 1-2 in 438 MLT, and our TA Jason's office hours will be on Wednesday at 2-4 in 218 MLT.]

Definition 1.1.1. A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a triple, where:

- Ω is a set of "outcomes",
- $\mathcal{F}$ is a set of "events", so $\mathcal{F} \subseteq 2^\Omega = \{\text{all subsets of } \Omega\}$,
- $\mathbb{P}\colon \mathcal{F} \to [0,1]$, where $\mathbb{P}(A)$ is interpreted as the probability that A occurs. △

Example 1.1.2. Consider:

1. Perhaps $\Omega = \{1,2,3,4,5,6\}$, and we are rolling a die. If it is fair, then $\mathbb{P}(\{1,2,3,4\}) = \frac{4}{6}$.
2. Perhaps $\Omega = \mathbb{N} = \{0,1,\dots\}$. Perhaps the experiment consists of flipping a coin until we get heads, and the outcome is the number of tails before the first heads. If the coin is fair, it is clear that $\mathbb{P}(\{n\}) = \frac{1}{2^{n+1}}$. Since any subset $A \subset \mathbb{N}$ is a countable union of singleton sets $A = \{n_1, n_2, \dots\}$, we have $\mathbb{P}(A) = \sum_{i\geq 1} \mathbb{P}(\{n_i\})$. The moral is that for Ω countable, we can do probability theory while avoiding measure theory.
3. Suppose $\Omega = [0,1]^2$, and we are dropping a pin on a square table. Since Ω is (necessarily!) uncountable here, there are a variety of "paradoxes" which require measure theory to resolve.
4. Suppose $\Omega = \{0,1\}^{\mathbb{N}}$ (cf. HW 0); perhaps we are flipping an infinite sequence of coins.

   If the coin flips are independent and fair, for $A = \{\omega \in \Omega\colon \omega_2 = \omega_3 = 1\}$, we have $\mathbb{P}(A) = \frac{1}{4}$. For $B = \{\omega \in \Omega\colon \omega_n = 1 \text{ for all } n \geq 100\}$ we have $\mathbb{P}(B) = 0$. Let $C = \{\omega \in \Omega\colon \omega_n = 1 \text{ eventually}\}$. Then

$$\begin{aligned} C &= \{\omega\colon \text{there is } N \text{ so that } \omega_n = 1 \text{ for all } n \geq N\} \\ &= \bigcup_{N\in\mathbb{N}} \{\omega\colon \omega_n = 1 \text{ for all } n \geq N\} \\ &= \bigcup_{N\in\mathbb{N}} \bigcap_{n\geq N} \{\omega_n = 1\}, \end{aligned}$$

   where $\{\omega_n = 1\}$ is shorthand for $\{\omega \in \Omega\colon \omega_n = 1\}$. We will later show the union bound, which implies

$$\mathbb{P}(C) \leq \sum_{N\in\mathbb{N}} \mathbb{P}\Big(\bigcap_{n\geq N} \{\omega_n = 1\}\Big),$$

   which is a countable sum of zeros. Suppose

$$D = \left\{\omega \in \Omega\colon \lim_{n\to\infty} \frac{\omega_1 + \cdots + \omega_n}{n} = \frac{1}{2}\right\}.$$

   Then $\mathbb{P}(D) = 1$; this is the strong law of large numbers. △

Let's review measure theory. It can be thought of as an abstraction of the idea of length/area/volume, and along with this comes an abstraction of the idea of integration. Let us begin with

Definition 1.1.3. Let $\mathcal{F} \subseteq 2^\Omega$. We say $\mathcal{F}$ is a σ-field (also known as a σ-algebra) if:

(i) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$,

(ii) $A_1, A_2, \cdots \in \mathcal{F} \implies \bigcup_{n\geq 1} A_n \in \mathcal{F}$,

(iii) $\emptyset \in \mathcal{F}$.

From these axioms it follows that $\Omega \in \mathcal{F}$, and that $\mathcal{F}$ is stable under countable intersections. A tuple $(\Omega, \mathcal{F})$ is called a measurable space, and the elements of $\mathcal{F}$ are called measurable sets. $\triangle$

Definition 1.1.4. A measure is a function

$$\mu \colon \mathcal{F} \to \mathbb{R} \cup \{\infty\}$$

satisfying, for $A, A_1, A_2, \cdots \in \mathcal{F}$,

(i) $\mu(A) \geq \mu(\emptyset) = 0$, and

(ii) if $A_1, A_2, \ldots$ are disjoint (meaning $A_i \cap A_j = \emptyset$ for $i \neq j$), then

$$\mu\Big(\bigcup_{n\geq 1} A_n\Big) = \sum_{n\geq 1} \mu(A_n). \qquad \triangle$$

To have the second condition above, one needs to be careful (e.g. with well-definedness). In particular, one cannot always just pick $\mathcal{F}$ to be 2^Ω.

Definition 1.1.5. A probability measure also satisfies $\mu(\Omega) = 1$. $\triangle$

We'll use the notation μ for a general measure and $\mathbb{P}$ for a probability measure.

Lemma 1.1.6. *Let μ be a measure on $(\Omega, \mathcal{F})$. We have*

(i) If $A, B \in \mathcal{F}$ satisfy $A \subseteq B$, then $\mu(A) \leq \mu(B)$.

(ii) If $A_1, A_2, \cdots \in \mathcal{F}$, not necessarily disjoint, then

$$\mu\Big(\bigcup_{n\geq 1} A_n\Big) \leq \sum_{n\geq 1} \mu(A_n).$$

(iii) If $A_i \uparrow A$, then $\mu(A_i) \uparrow \mu(A)$.

(iv) (cf. HW 1) If $A_i \downarrow A$, and $\mu(A_i) < \infty$ for some i, then $\mu(A_i) \downarrow \mu(A)$.

(For sets A_i, A, the notation $A_i \uparrow A$ means $A_1 \subseteq A_2 \subseteq \ldots$ and $\cup A_i = A$, and for $x_i, x \in \mathbb{R}$, the notation $x_i \uparrow x$ means $x_1 \leq x_2 \leq \ldots$ and $\lim_{i\to\infty} x_i = x$.)

Proof. Part (i) follows from writing $B = A \sqcup (B \setminus A)$ and applying axiom (ii) required of a measure. Note that $B \setminus A = B \cap A^c \in \mathcal{F}$.

Part (ii) follows from defining $B_i = A_i \setminus (A_1 \cup \cdots \cup A_{i-1})$; note that the $B_i \subseteq A_i$ are disjoint and satisfy

$$\bigcup_{i\geq 1} A_i = \bigcup_{i\geq 1} B_i.$$

Hence we have

$$\mu\Big(\bigcup_{n\geq 1} A_n\Big) = \sum_{n\geq 1} \mu(B_n) \leq \sum_{i\geq 1} \mu(A_n),$$

where the inequality is part (i) of the lemma.

Part (iii) follows from writing, as before, $A = \sqcup_{i\geq 1} B_i$. Moreover, we have $A_n = \sqcup_{i=1}^n B_i$. Hence

$$\mu(A) = \lim_{n\to\infty} \sum_{i=1}^n \mu(B_i) = \lim_{n\to\infty} \mu(A_n),$$

as desired.

Part (iv) will be in HW 1. $\square$

In HW 1 we will also prove the following. For any $\mathcal{A} \subseteq 2^\Omega$, we let $\sigma(\mathcal{A})$ be the intersection of all σ-fields containing $\mathcal{A}$. Then $\sigma(\mathcal{A})$ is in fact a σ-field.

Definition 1.1.7. Let T be a topological space. The Borel σ-field is $\sigma(\mathcal{A})$, where $\mathcal{A}$ consists of all open subsets of T. △

The most important examples for probability arise from $T = \mathbb{R}^d$ or $T = \{0,1\}^{\mathbb{N}}$. The latter is topologized by taking the open sets to be cylinder sets, which are sets of the form

$$\{\omega \in T \colon \omega_1 = a_1, \ldots, \omega_n = a_n\}$$

for fixed $(a_1, \ldots, a_n) \in \{0,1\}^n$.

Let's see why we can't always take every subset to be measurable.

Example 1.1.8. Let $T = \mathbb{R}/\mathbb{Z}$ be the unit circle. Let us show that there does not exist a rotationally invariant probability measure on $(T, 2^T)$. (Here, rotational invariance means $\mathbb{P}(A) = \mathbb{P}(A + c)$ for all $c \in \mathbb{R}$.)

Define an equivalence relation $\equiv$ on T by $x \equiv y$ if $x - y \in \mathbb{Q}$. By the axiom of choice, we may consider the set A consisting of one representative from each equivalence class. Note that for each $q \in \mathbb{Q}/\mathbb{Z}$, we may consider $A_q = A + q$; it follows that

$$\mathbb{R}/\mathbb{Z} = \bigcup_{q \in \mathbb{Q}/\mathbb{Z}} A_q \implies 1 = \mathbb{P}(\mathbb{R}/\mathbb{Z}) = \sum_{q \in \mathbb{Q}/\mathbb{Z}} \mathbb{P}(A_q).$$

By rotational invariance, $\mathbb{P}(A_q)$ are all equal to each other. We obtain a contradiction by considering separately the cases $\mathbb{P}(A) = 0$ and $\mathbb{P}(A) > 0$. △

1.2 Sep 9, 2019

We'll continue with our quick review of measure theory today. Problem set 1 will be posted later today, due a week from now. [We'll have weekly problem sets. If the class size remains the same, we will have a takehome final, but if it drops we may replace the final with presentations.]

Definition 1.2.1. A collection of sets $\mathcal{A} \subseteq 2^\Omega$ is called a (boolean) algebra if:

(i) $A \in \mathcal{A} \implies A^c \in \mathcal{A}$, and

(ii) $A_1, \dots, A_n \in \mathcal{A} \implies A_1 \cup \dots \cup A_n \in \mathcal{A}$. △

Boolean algebras are often easier to handle than σ-algebras, because we only require finite unions.

Example 1.2.2. Some examples of boolean algebras:

(i) The open and closed sets of a topological space form a boolean algebra.

(ii) For $\Omega = \mathbb{Z}$, the set $\mathcal{A} = \{\text{finite subsets}\} \cup \{\text{cofinite subsets}\}$ forms a boolean algebra. (A cofinite set is one in which the complement is finite.) △

Definition 1.2.3. We call $\mu \colon \mathcal{A} \to \mathbb{R} \cup \{\infty\}$ a measure if

(i) $\mu(A) \geq \mu(\emptyset) = 0$, and

(ii) $\mu(\sqcup_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mu(A_i)$, whenever both $A_i \in \mathcal{A}$ and $\sqcup_{i=1}^\infty A_i \in \mathcal{A}$.

We say μ is σ-finite if there exist $A_1, A_2, \dots$ so that

$$\bigcup_{i \geq 1} A_i = \Omega \quad \text{and} \quad \mu(A_i) < \infty \text{ for all } i. \qquad \triangle$$

The following theorem says that we don't really have to worry about σ-algebras:

Theorem 1.2.4 (Caratheodory Extension Theorem)**.** *If $\mathcal{A}$ is a boolean algebra and $\mu \colon \mathcal{A} \to \mathbb{R} \cup \{\infty\}$ is a σ-finite measure on $\mathcal{A}$, then there exists a unique extension of μ to a measure on the σ-algebra $\sigma(\mathcal{A})$ generated by $\mathcal{A}$.*

The following theorem can be deduced from the Caratheodory Extension theorem.

Theorem 1.2.5. *Suppose $F \colon \mathbb{R} \to \mathbb{R}$ is a nondecreasing and right-continuous function. Then there exists a unique measure μ on the Borel sets $(\mathbb{R}, \mathcal{B})$ such that $\mu(a, b] = F(b) - F(a)$.*

(Here, right-continuous means $\lim_{y \downarrow x} F(y) = F(x)$, and the Borel sets are the σ-algebra generated by open sets in $\mathbb{R}$.) The most important case of the above theorem is $F(x) = x$. The resulting measure is called Lebesgue measure.

Proof idea of Theorem 1.2.5. Let us consider the boolean algebra

$$\mathcal{A} = \{(a_1, b_1] \cup \dots \cup (a_n, b_n] \colon -\infty \leq a_1 \leq b_1 \leq \dots \leq b_n \leq \infty\},$$

where by convention $(a, a] \stackrel{\text{def}}{=} \{a\}$. (One should check that this actually is an algebbra.)

To get the conclusion of Theorem 1.2.5 it is clear that we should define

$$\mu\Big((a_1, b_1] \cup \dots \cup (a_n, b_n]\Big) = F(b_1) - F(a_1) + \dots + F(b_n) - F(a_n);$$

the real issue is well-definedness. For example, with $a < b < c$, we have $(a, c] = (a, b] \cup (b, c]$, and so we potentially have two formulas for $\mu(a, c]$. Fortunately, the formulas agree.

So if you believe that μ is well defined on $\mathcal{A}$, the Caratheodory extension theorem guarantees the existence (and uniqueness!) of a measure on $\sigma(\mathcal{A}) = \mathcal{B}$. □

Definition 1.2.6. A family $\mathcal{I} \subseteq 2^\Omega$ is called a π-system if $I_1, I_2 \in \mathcal{I} \implies I_1 \cap I_2 \in \mathcal{I}$. (Intuitively, think about intervals in $\mathbb{R}$.) △

The following theorem is an observation of Dynkin.

Theorem 1.2.7 ("Thm 1"). *Let $\mathcal{I}$ be a π-system and let $\mathcal{F} = \sigma(\mathcal{I})$. If $\mu_1, \mu_2 \colon \mathcal{F} \to \mathbb{R} \cup \{\infty\}$ are measures such that $\mu_1(\Omega) = \mu_2(\Omega) < \infty$, and $\mu_1 = \mu_2$ on $\mathcal{I}$, then $\mu_1 = \mu_2$ on $\mathcal{F}$.*

Example 1.2.8. Suppose $\Omega = [0,1]$ and $\mathcal{I} = \{[0,x] \colon 0 \le x \le 1\}$. One can check that the Borel subsets $\mathcal{B}([0,1])$ is equal to $\sigma(\mathcal{I})$, so by Theorem 1.2.7 defining a measure on $\mathcal{I}$ is enough to define a measure on $\mathcal{B}([0,1])$. △

Definition 1.2.9. A family $\mathcal{L} \subseteq 2^\Omega$ is called a λ-system if

(i) $\Omega \in \mathcal{L}$,

(ii) $A, B \in \mathcal{L}$ with $A \subseteq B$, then $B \setminus A \in \mathcal{L}$, and

(iii) If $A_n \in \mathcal{L}$ and $A_n \uparrow A$, then $A \in \mathcal{L}$. △

Lemma 1.2.10 ("Lemma 2"). *A family of sets $\mathcal{F} \subseteq 2^\Omega$ is a σ-algebra if and only if $\mathcal{F}$ is both a π-system and a λ-system.*

We'll be using this a lot.

Proof. The forward direction is trivial.

The other direction follows from the following observations:

(i) $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$, by λ-system's axioms (i) and (ii).

(ii) For a countable family of sets $A_n \in \mathcal{F}$, define $B_n = A_1 \cup \cdots \cup A_n$. By definition of π-system, and λ-system axiom (ii), $B_n \in \mathcal{F}$. The $B_n \uparrow \cup_m A_m$, so by λ-system axiom (iii) we get $\cup_m A_m \in \mathcal{F}$. □

Theorem 1.2.11 (Dynkin π-λ theorem). *Let $\mathcal{I}$ be a π-system and $\mathcal{L}$ be a λ-system. If $\mathcal{I} \subseteq \mathcal{L}$, then $\sigma(I) \subseteq \mathcal{L}$.*

Proof. Let $\lambda(\mathcal{I})$ be the smallest λ-system containing $\mathcal{I}$. (Formally, this means we consider all λ-systems containing $\mathcal{I}$ and intersect them. [It is immediate to check that an arbitrary intersection of λ-systems is again a λ-system.])

We must show that $\lambda(\mathcal{I}) = \sigma(\mathcal{I})$, and by Lemma 1.2.10 it is enough to show that $\lambda(\mathcal{I})$ is a π-system. Let

$$\mathcal{L}_1 = \{B \in \lambda(\mathcal{I}) \colon B \cap C \in \lambda(\mathcal{I}) \, \forall C \in \mathcal{I}\}.$$

Since $\mathcal{I}$ is a π-system, $\mathcal{I} \subseteq \mathcal{L}_1$. The claim is that $\mathcal{L}_1$ is a λ-system, because we then get $\mathcal{L}_1 = \lambda(\mathcal{I})$. We check the three axioms required of a λ-system as follows:

(i) Note that $\Omega \in \mathcal{L}_1$.

(ii) Let us now take $B_1, B_2 \in \mathcal{L}_1$ with $B_1 \subseteq B_2$. Recall that $(B_2 \setminus B_1) \cap C = (B_2 \cap C) \setminus (B_1 \cap C)$; furthermore, $B_2 \cap C \in \lambda(\mathcal{I})$ and $B_1 \cap C \in \lambda(\mathcal{I})$ by assumption. Hence their difference is also in $\lambda(\mathcal{I})$.

(iii) Finally, consider a countable family of sets $B_n \in \mathcal{L}_1$ with $B_n \uparrow B$, and let $C \in \mathcal{I}$. Since $(B_n \cap C) \uparrow (B \cap C)$, it follows that $B \cap C \in \lambda(\mathcal{I})$. This proves that $B \in \mathcal{L}_1$.

Since $\mathcal{L}_1 = \lambda(\mathcal{I})$, we need to prove that $\mathcal{L}_1$ is a π-system. We've shown that it is stable under intersections with $\mathcal{I}$. To show it is stable under intersections in $\lambda(\mathcal{I})$, we iterate this process, defining $\mathcal{L}_2$ and arguing analogously to above. (This detail will be omitted.) □

Proof of Theorem 1.2.7. Let $\mathcal{L} = \{A \in \mathcal{F} \colon \mu_1(A) = \mu_2(A)\}$. We claim that $\mathcal{L}$ is a λ-system. We check the axioms required of a λ-system as follows:

(i) By hypothesis, $\mathcal{L}$ contains Ω.

(ii) If $A \subseteq B \in \mathcal{L}$, then

$$\mu_1(B \setminus A) = \mu_1(B) - \mu_1(A) = \mu_2(B) - \mu_2(A) = \mu_2(B \setminus A)$$

and $B \setminus A \in \mathcal{L}$. Note that the above computation only holds when $\mu_1(B) < \infty$, which is true by hypothesis.

(iii) If $A_n \in \mathcal{L}$ is such that $A_n \uparrow A$, then

$$\mu_1(A) = \lim_{n\to\infty} \mu_1(A_n) = \lim_{n\to\infty} \mu_2(A_n) = \mu_2(A)$$

and $A \in \mathcal{L}$.

By the π-λ Theorem (Theorem 1.2.11), and the assumption that $\mathcal{I} \subseteq \mathcal{L}$, we get $\mathcal{F} = \sigma(\mathcal{I}) \subseteq \mathcal{L}$. □

In Durrett's book, Appendix A1.5 in the 4th edition contains a σ-finite version of Theorem 1.2.7. To wrap up our measure theoretic excursion, let's talk about why Caratheodory Extension works.

Proof idea for Theorem 1.2.4. There are four steps to this proof:

1. For *any* $E \subseteq \Omega$, define the outer measure

$$\mu^*(E) = \inf_{\substack{\{A_i\}:\ A_i \in \mathcal{A} \\ E \subseteq \cup A_i}} \sum_i \mu(A_i).$$

2. Let us define E to be measurable if for all $F \subseteq \Omega$,

$$\mu^*(F) = \mu^*(E \cap F) + \mu^*(E^c \cap F).$$

3. Now we can check that if $A \in \mathcal{A}$, then $\mu^*(A) = \mu(A)$. Moreover, A is measurable as in the definition above.

4. Finally, let

$$\mathcal{A}^* = \{E \subseteq \Omega \colon E \text{ is measurable}\}$$

and check that $\mathcal{A}^*$ is a σ-field and that the restriction of μ^* to $\mathcal{A}^*$ is a measure.

The key properties of μ^* that allow us to do steps 3 and 4 above are:

(i) $\mu^*(\emptyset) = 0$,

(ii) $E \subseteq F \implies \mu^*(E) \leq \mu^*(F)$, and

(iii) $F \subseteq \cup_{n\geq 1} F_n \implies \mu^*(F) \leq \sum_{n\geq 1} \mu^*(F_n)$. □

1.3 Sep 11, 2019

We have just a little bit of abstract measure theory nonsense to cover, before getting into the concrete probability theory.

We'd like to understand which sets are measurable. We have, for $\mathcal{A}$ an algebra and $\mathcal{A}^*$ the measurable sets, the inclusions

$$\sigma(\mathcal{A}) \subseteq \mathcal{A}^* \subseteq 2^\Omega,$$

and for most interesting applications these inclusions are strict. [For example, $\Omega = \mathbb{R}, \mathcal{A} = \{\text{Borel sets} \subset 2^{\mathbb{R}}\}, \mathcal{A}^* = \{\text{Lebesgue measurable sets} \subset 2^{\mathbb{R}}\}$.]

Definition 1.3.1. A set $N \subseteq \Omega$ is a null set if $\mu^*(N) = 0$. Note that any null set is measurable, since for all $F \in 2^\Omega$, we have $F \subseteq (F \cap N^c) \sqcup N$ and

$$\mu^*(F) \leq \mu^*(F \cap N^c) + \underbrace{\mu^*(N)}_{=0} \leq \mu^*(F),$$

so these inequalities are equalities and N is measurable. △

The philosophy is that we don't need to worry about null sets. In particular:

Definition 1.3.2. A measure space $(\Omega, \mathcal{F}, \mu)$ is called complete if whenever $A \subseteq B$ and $B \in \mathcal{F}$ with $\mu(B) = 0$, then $A \in \mathcal{F}$. △

Theorem 1.3.3. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. There exists a complete measure space $(\Omega, \overline{\mathcal{F}}, \overline{\mu})$ that extends the original space, that is,*

1. $E \in \overline{\mathcal{F}} \iff E = A \cup N$ *for* $A \in \mathcal{F}$ *and* $N \subseteq B \in \mathcal{F}$ *with* $\mu(B) = 0$,
2. $\overline{\mu} = \mu$ *on* $\mathcal{F}$.

(The triple $(\Omega, \overline{\mathcal{F}}, \overline{\mu})$ is called the completion of $(\Omega, \mathcal{F}, \mu)$.)

Example 1.3.4. Let $\mathcal{B}$ be the Borel sets, and λ denote the Lebesgue measure on $\mathbb{R}$. The completion is $(\mathbb{R}, \overline{\mathcal{B}}, \overline{\lambda})$, where $\overline{\mathcal{B}}$ are Lebesgue sets in the usual sense and $\overline{\lambda}$ is Lebesgue measure in the usual sense. (There exist null sets $N \notin \mathcal{B}$.) △

There is this notion of a Borel Hierarchy. Let us take $\Omega = \mathbb{R}$ and $\mathcal{A}_0 = \{\text{open intervals} \subset \mathbb{R}\}$. Let $\mathcal{A}_n = \mathcal{A}_{n-1}^*$, where

$$\mathcal{A}^* = \left\{ A^c, \bigcup_{k \geq 1} A_k : A, A_k \in \mathcal{A} \right\}.$$

Note that $\mathcal{B} \supseteq \mathcal{A}_n$ for every n. One might guess

$$\mathcal{B} = \bigcup_{n \geq 0} \mathcal{A}_n = \mathcal{A}_\infty,$$

but no!

We'd need a transfinite hierarchy to exhaust $\mathcal{B}$.

Theorem 1.3.5 (Durrett, 4th ed, Appendix A2.2)**.** *For all Lebesgue measurable $E \subseteq \mathbb{R}$, we can write*

$$E = A \setminus N,$$

for $A \in \mathcal{A}_2$ an element of the second level of the Borel Hierarchy, and N a null set.

Let's start interpreting things in probabilistic language.

2 Probability preliminaries

2.3 Sep 11, 2019

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space (that is, a measure space with $\mathbb{P}(\Omega) = 1$).

Definition 2.3.1. We say $X : \Omega \to \mathbb{R}$ is a random variable (R.V.) if $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}$. △

Definition 2.3.2. The distribution of X is the measure on $(\mathbb{R}, \mathcal{B})$ defined by

$$\mu_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}).$$

We write

$$\mathbb{P}(X \in B) \stackrel{\text{def}}{=} \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$$

to suppress the ω from the notation. △

Example 2.3.3. Consider the indicator random variable of $\mathcal{F}$ is defined as follows: for $A \in \mathcal{F}$, we set

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \text{for } \omega \in A \\ 0 & \text{for } \omega \notin A. \end{cases}$$

The map $\mathbb{1}_A : \Omega \to \mathbb{R}$ is a random variable, because

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

(More to come later.) △

Definition 2.3.4. In general, given measurable spaces $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$, a function $X : \Omega \to S$ is called measurable if $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{S}$. △

Hence, a (real-valued) random variable is the case when Ω has a probability measure and $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B})$. But one might be interested in, e.g., random graphs, random trees, random complex numbers, or random extended reals (so S is a set of graphs, a set of trees, $\mathbb{C}$, or $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$). In general, an S-valued random variable is a measurable function $(\Omega, \mathbb{P}) \to (S, \mathcal{S})$.

Definition 2.3.5. If $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (S, \mathcal{S})$ is a measurable function, the pushforward μ of P is defined as

$$\mu(B) \stackrel{\text{def}}{=} \mathbb{P}(X^{-1}(B)),$$

for all $B \in \mathcal{S}$. △

Hence, the case $S = \mathbb{R}, \mu = \mu_X$ is a special case of pushforwards of measures.

Lemma 2.3.6. *With notation as in Definition 2.3.5, μ is a measure.*

Proof. We check $\mu(\emptyset) = \mathbb{P}(X^{-1}(\emptyset)) = \mathbb{P}(\emptyset) = 0$, and

$$\mu\left(\bigsqcup_{n \geq 1} B_n\right) = \mathbb{P}\left(X^{-1}\left(\bigsqcup_{n \geq 1} B_n\right)\right) = \mathbb{P}\left(\bigsqcup_{n \geq 1} X^{-1}(B_n)\right) = \sum_{n \geq 1} \mathbb{P}(X^{-1}(B_n)) = \sum_{n \geq 1} \mu(B_n).$$

□

Definition 2.3.7. The distribution function (sometimes denoted C.D.F.) of a random variable $X : \Omega \to \mathbb{R}$ is the function $F = F_X : \mathbb{R} \to \mathbb{R}$ defined by

$$F(x) = \mathbb{P}(X \leq x) \stackrel{\text{def}}{=} \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}).$$ △

We note that since $\{\omega \in \Omega \colon X(\omega) \le x\} = X^{-1}((-\infty, x])$, we really have

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

Although F_X only contains the information of μ_X on sets of the form $(-\infty, x]$, observe that F_X actually determines μ_X. This is because:

1. $\mathcal{B} = \sigma(\{(-\infty, x] \colon x \in \mathbb{R}\})$, where $\{(-\infty, x] \colon x \in \mathbb{R}\}$ is a π-system.
2. By the π-λ theorem (Theorem 1.2.11), the values $\mu_X((-\infty, x])$ for $x \in \mathbb{R}$ determine the values of $\mu_X(B)$ for all $B \in \mathcal{B}$.

Note also that F_X has the following properties:

1. It is nondecreasing, that is, $x \le y$ means $F(x) \le F(y)$,
2. $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$, and
3. It is right continuous, so $\lim_{y\downarrow x} F(y) = F(x)$.

Example 2.3.8 (More random variables)**.** Consider:

1. Let $X = \mathbb{1}_A$. Then $\mathbb{P}(X \le x) = F_X(x)$ is given piecewise by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \mathbb{P}(A^c) & \text{if } 0 \le x < 1 \\ 1 & \text{if } x \ge 1. \end{cases}$$

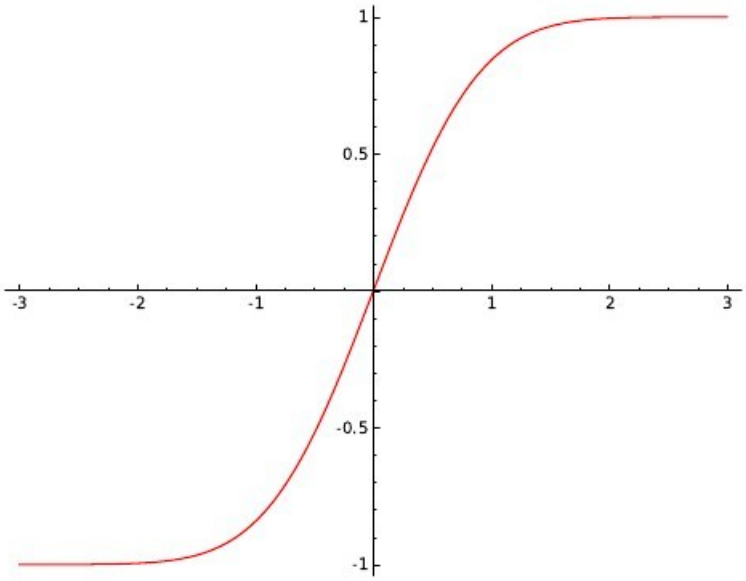
2. For a $\lambda > 0$ we have the exponential random variable $\text{Exp}(\lambda)$, whose distribution function is

$$F(x) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \ge 0. \end{cases}$$

3. We have the normal distribution $N(0, 1)$ with

$$F(x) \stackrel{\text{def}}{=} \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

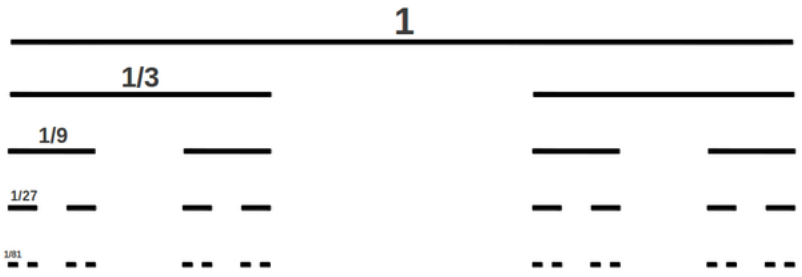
The graph of F looks like



4. Consider the (usual) Cantor set

$$\mathcal{C} = \bigcap_{n \geq 1} C_n$$

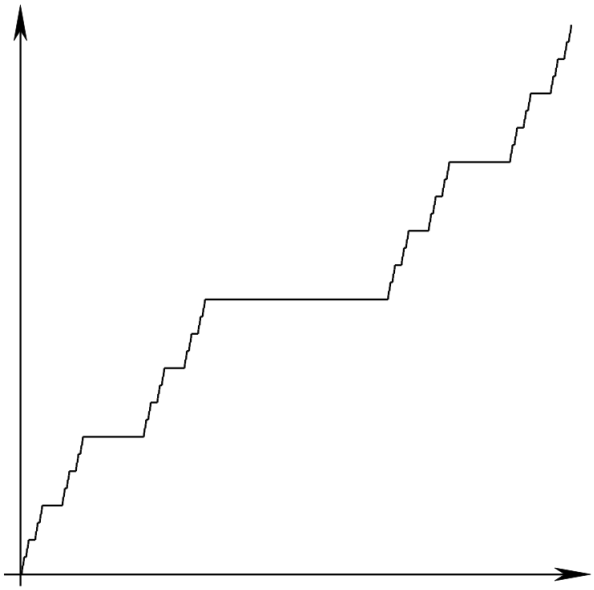
for $C_0 = [0, 1]$, $C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$, $C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{3}{9}] \cup [\frac{6}{9}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$, and so on, as in the following picture.



Note that a choice of an element in the Cantor set corresponds exactly to a choice of countable sequence of L's and R's. Now let $X =$ "a uniform random element of $\mathcal{C}$". Let $F = F_X$; we have

$$F(x) = \begin{cases} \frac{1}{2} & \forall x \in [\frac{1}{3}, \frac{2}{3}] \\ \frac{1}{4} & \forall x \in [\frac{1}{9}, \frac{2}{9}] \\ \frac{3}{4} & \forall x \in [\frac{7}{9}, \frac{8}{9}] \\ \vdots & \vdots \end{cases}$$

(An approximation of) the graph of F is:



Note that F is continuous (since if F had a jump, say at $x \in [0, 1]$, then X would assign a positive probability of choosing x.) $\triangle$

2.4 Sep 16, 2019

Last time we saw some random variables, which were measurable maps $X\colon (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$.

We had a notion of distribution, denoted $\mu_X(B) = \mathbb{P}(\{\omega \in \Omega\colon X(\omega) \in B\}) = \mathbb{P}(\{X \in B\})$, and a notion of the distribution function $F_X(x) = \mathbb{P}(X \le x)$. The π-λ theorem implies that F_X determines μ_X.

We talked about the uniform distribution on the Cantor set, namely example #4 in 2.3.8. Even though the distribution function is continuous, this is not always the case. However, it is always right-continuous.

Definition 2.4.1. Let us denote by

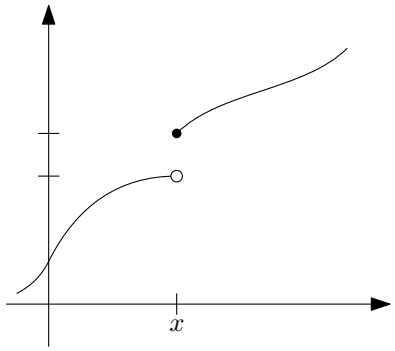
$$F(x_-) \stackrel{\text{def}}{=} \lim_{y \uparrow x} F(y) = \lim_{y \uparrow x} \mathbb{P}(X \le y) = \mathbb{P}(X < x).$$

(Note that the inequality is strict.) The last equality follows from

$$\bigcup_{y<x} \{X \le y\} = \{X < x\}. \qquad \triangle$$

Thus $F(x) - F(x_-) = \mathbb{P}(X \le x) - \mathbb{P}(X < x) = \mathbb{P}(X = x)$, and discontinuities occur at points which are assigned a positive probability by P.

Definition 2.4.2. We say x is an <u>atom</u> of μ_X if $\mathbb{P}(X = x) > 0$. (See the picture below for an example of a distribution function which has an atom.)



$\triangle$

Definition 2.4.3. We say X is <u>discrete</u> if $\mu_X(S) = 1$ for some countable set S. $\triangle$

Example 2.4.4. The standard example is that of an indicator function, see Example 2.3.3. (We denote this distribution by Bernoulli(p).) $\triangle$

Example 2.4.5. Let $q_1, q_2, \ldots$ be an enumeratorion of $\mathbb{Q}$. Note that

$$F(x) = \sum_{n \ge 1} 2^{-n} \mathbb{1}_{[q_n, \infty)}$$

is the distribution of a random variable X with $\mathbb{P}(X = q_n) = 2^{-n}$. Note that F is discontinuous at every rational. (!) $\triangle$

In HW 2, we'll see that F is the distribution of some $\mathbb{R} \cup \{\pm\infty\}$-valued random variable if and only if F is nondecreasing and right-continuous. (If we wanted $\mathbb{R}$-valued random variables, we'd need some condition on the limit.)

Recall our abstract framework about S-valued random variables. That is, a function $X\colon (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ is called <u>measurable</u> if $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{S}$.

Definition 2.4.6. The <u>σ-field generated by X</u> is denoted $\sigma(X)$ and is defined as

$$\sigma(X) \stackrel{\text{def}}{=} \sigma(\{X \in B\}\colon B \in \mathcal{S}).$$

$\triangle$

We should think of this σ-field as all the information we know if we are given the value of X. Put another way, it should be thought of as all the yes-no questions we can answer (namely, "is X in B?"), if we know the value of X.

In HW 2, we'll see that if $\mathcal{C} \subset \mathcal{S}$ generates S (i.e., $\sigma(\mathcal{C}) = \mathcal{S}$), then the sets $\{X \in C : C \in \mathcal{C}\}$ generate $\sigma(X)$.

Example 2.4.7. We've seen an example of this, namely with $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B})$ and $\mathcal{C} = \{(-\infty, x] : x \in \mathbb{R}\}$. This is why F_X determines μ_X. △

Lemma 2.4.8. *Let X, f be maps*

$$(\Omega, \mathcal{F}) \xrightarrow{X} (S, \mathcal{S}) \xrightarrow{f} (T, \mathcal{T})$$

such that X and f are measurable. Then $f \circ X$ is also measurable.

This is trivial to check from the definitions, but it gives rise to very rich functions:

Example 2.4.9. Let $S, T = \mathbb{R}$. Lemma 2.4.8 in this case implies that if X is a random variable, then so is X^2, $|X|$, e^X, $\sin(X)$, and so on. △

Lemma 2.4.10. *If $f : (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}, \mathcal{B})$ is measurable, and $X_1, \dots, X_n$ are random variables, then so is $Y = f(X_1, \dots, X_n)$.*

(Here $\mathcal{B}^n$ denotes the Borel sets of $\mathbb{R}^n$.)

Proof. We saw in HW 1 that

$$\begin{aligned} \mathcal{B}^n &= \sigma(\text{half-open rectangles}) \\ &= \sigma(A_1 \times \cdots \times A_n : A_i \in \mathcal{B}). \end{aligned}$$

The Lemma now follows from the observation that

$$\{Y \in A_1 \times \cdots \times A_n\} = \bigcap_{i=1}^{n} \{X_i \in A_i\}. \qquad \square$$

Lemma 2.4.10 also gives you many new examples from old ones, namely

Example 2.4.11. For random variables $X_1, \dots, X_n$, Lemma 2.4.10 implies that $X_1 + \cdots + X_n$, $X_1 \dots X_n$, $\max(X_1, \dots, X_n)$, and $\min(X_1, \dots, X_n)$ are random variables, too. △

Lemma 2.4.12. *If $X_1, X_2, \dots$ are $\mathbb{R}$-valued random variables, then $\inf X_n$, $\sup X_n$, $\liminf X_n$, and $\limsup X_n$ are $(\mathbb{R} \cup \{\pm\infty\})$-valued random variables.*

Proof. We write

$$\{\inf X_n < a\} = \bigcup_{n \geq 1} \{X_n < a\} \in \mathcal{F}, \text{ for all } a \in \mathbb{R}.$$

Checking that $\sup X_n$ is also a $(\mathbb{R} \cup \{\pm\infty\})$-valued random variable is completely analogous. We also have

$$\liminf_{n\to\infty} X_n = \sup_{N \geq 1} \inf_{n \geq N} X_n, \quad \text{and} \quad \limsup_{n\to\infty} X_n = \inf_{N \geq 1} \sup_{n \geq N} X_n. \qquad \square$$

Recall HW 0, where we were wondering whether

$$A = \{\omega \in \Omega : \lim_{n\to\infty} X_n \text{ exists}\} = \bigcup_{\ell \in \mathbb{R}} \{\omega \in \Omega : \lim_{n\to\infty} X_n = \ell\}$$

was measurable. Unfortunately the union is uncountable, so the problem is not immediate from the measurability of all $\{\omega \in \Omega : \lim_{n\to\infty} X_n = \ell\}$. Here's a quick solution to this problem, now that we have some theory:

Observe that $A = f^{-1}(\{0\})$, where $f : \Omega \to \mathbb{R}$ is given by $f(\omega) = \limsup_{n\to\infty} X_n(\omega) - \liminf_{n\to\infty} X_n(\omega)$. (Note that f is measurable.)

This set A is important.

Definition 2.4.13. We say the random variables X_n converge almost surely if $\mathbb{P}(A) = 1$, where

$$A = \{\omega \in \Omega \colon \lim_{n\to\infty} X_n \text{ exists}\}.$$

The notation

$$X_n \to X \quad \text{a.s.}$$

means $\mathbb{P}(\omega \in \Omega \colon \lim_{n\to\infty} X_n(\omega) = X(\omega)) = 1$. △

Example 2.4.14. The strong law of large numbers says that for $\Omega = \{0,1\}^{\mathbb{N}}$, and $X_n = \frac{\omega_1+\cdots+\omega_n}{n}$, and $\mathbb{P} = \prod \text{Bernoulli}(\frac{1}{2})$, then $X_n \to \frac{1}{2}$ a.s.. △

(As a fun exercise, think of an example of $X_n \to X$ a.s. where X is nonconstant.)

We take a break from probability to develop some more measure theory (in particular, the theory of integration, which will be needed to take expected values.)

3 Integration

3.4 Sep 16, 2019

We let $(\Omega, \mathcal{F}, \mu)$ be a σ-finite measure space. Our goal is to define the integral

$$\int f \, d\mu$$

for as many functions $f \colon \Omega \to \mathbb{R}$ as possible. The integral has alternate notations

$$\int f \, d\mu = \int_\Omega f(\omega) \, d\mu(\omega) = \int_\Omega f(\omega) \, \mu(d\omega).$$

Let us build up our integral:

Definition 3.4.1 (Integral)**.**

0. If we are reasonable, we should have
$$\int \mathbb{1}_A \, d\mu \stackrel{\text{def}}{=} \mu(A).$$

1. We want the integral to be linear. Thus, let $\varphi(\omega)$ denote a simple function, which, by definition, is a finite sum
$$\varphi = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$$
for $a_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$. We can assume A_i are disjoint (so $A_i \cap A_j = \emptyset$ for $i \neq j$), and in this case we define
$$\int \varphi \, d\mu \stackrel{\text{def}}{=} \sum_{i=1}^n a_i \mu(A_i).$$
Intuitively, we can think of φ as a density and μ as a volume, and then the integral of φ with respect to μ can be thought of as mass. Alternatively, we can think of the integral as the "area under the graph". △

Definition 3.4.2. We say $f \leq g$ almost everywhere if $\mu(\{\omega \colon g(\omega < f(\omega)\}) = 0$. △

(The philosophy of integration is that measure zero sets shouldn't matter, so the integral shouldn't know whether $f \leq g$ everywhere or just almost everywhere.)

Here are some properties of the integral that we would like:

(i) If $f \geq 0$ almost everywhere, then we should have
$$\int f \, d\mu \geq 0.$$

(ii) We should have
$$\int (af) \, d\mu = a \int f \, d\mu.$$

(iii) We should also have
$$\int (f + g) \, d\mu = \int f \, d\mu + \int g \, d\mu.$$

Let's prove these properties about integrals of simple functions:

Proof. If f is a simple function with $f \geq 0$ almost everywhere, then $a_i \geq 0$ for all i such that $\mu(A_i) > 0$. Then

$$\int f \, d\mu = \sum a_i \mu(A_i) \geq 0.$$

The second property is obvious.

The third property is a bit more nuanced. First observe that if f, g are simple, then $f + g$ is simple. For

$$f = \sum a_i \mathbb{1}_{A_i} \quad \text{and} \quad g = \sum b_j \mathbb{1}_{B_j},$$

then

$$f + g = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) \mathbb{1}_{A_i \cap B_j}$$

and

$$\int (f + g) \, d\mu = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) \mu(A_i \cap B_j).$$

To write $\int (f + g) \, d\mu$ as $\int f \, d\mu + \int g \, d\mu$, first observe that we can assume that $\cup A_i = \cup B_j = \Omega$ (otherwise, we could add the complement and give it a coefficient of 0.) Then

$$\sum_{i=1}^{n} a_i \mu(A_i) = \sum_{i=1}^{n} a_i \sum_{j=1}^{m} \mu(A_i \cap B_j).$$

Then

$$\int f \, d\mu + \int g \, d\mu = \sum_{i=1}^{n} a_i \sum_{j=1}^{m} \mu(A_i \cap B_j) + \sum_{j=1}^{m} b_j \sum_{i=1}^{n} \mu(A_i \cap B_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) \mu(A_i \cap B_j) = \int (f + g) \, d\mu.$$

□

3.5 Sep 18, 2019

[Office Hours: Jason will have his on Wednesdays from 11-1 in 218, and Prof Levine's will be on Thursdays from 1-2 in 438.]

Let's recall the properties of integration (which we will prove in general):

(i) If $f \geq 0$ almost everywhere, then $\int f \, d\mu \geq 0$.

(ii) We have $\int (af) \, d\mu = a \int f \, d\mu$ for $a \in \mathbb{R}$,

(iii) We have $\int (f+g) \, d\mu = \int f \, d\mu + \int g \, d\mu$,

(iv) If $f \leq g$ almost everywhere, then $\int f \, d\mu \leq \int g \, d\mu$,

(v) If $f = g$ almost everywhere, then $\int f \, d\mu = \int g \, d\mu$, and

(vi) We have $|\int f \, d\mu| \leq \int |f| \, d\mu$.

Lasttime we defined $\int f \, d\mu$ for simple functions $f = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$, for $A_1, \ldots, A_n \in \mathcal{F}$ and $a_1, \ldots, a_n \in \mathbb{R}$, and checked (i)-(iii).

Lemma 3.5.1. *For any theory of integration, properties (iv)-(vi) follow from (i)-(iii).*

Proof. For property (iv), write $g = f + (g - f)$ and observe that by property (iii) we have

$$\int g \, d\mu = \int f \, d\mu + \int (g - f) \, d\mu,$$

and since $g - f \geq 0$ almost everywhere, the last term in the above equation is nonnegative by (i).

For property (v), we just apply property (iv) twice: if $f = g$ almost everywhere then $f \geq g$ almost everywhere and $f \leq g$ almost everywhere. Hence, by property (iv) we have

$$\int f \, d\mu \leq \int g \, d\mu \quad \text{and} \quad \int f \, d\mu \geq \int g \, d\mu,$$

so they're equal.

For property (vi), note that

$$\int f \, d\mu \leq \int |f| \, d\mu \quad \text{and} \quad \int (-f) \, d\mu \leq \int |f| \, d\mu,$$

since $f \leq |f|$ and $-f \leq |f|$. By property (ii) we get

$$\int (-f) \, d\mu = -\int f \, d\mu,$$

so combining the resulting inequalities gives

$$\left| \int f \, d\mu \right| \leq \int |f| \, d\mu.$$

□

We want to extend our theory of integration to more than just the simple functions. Recall that throughout we are working with a σ-finite measure space $(\Omega, \mathcal{F}, \mu)$: this means that there exist subsets $A_n \uparrow \Omega$ with $\mu(A_n) < \infty$. Let's continue to build up our integral:

Definition 3.5.2 (Integral; cf. Definition 3.4.1)**.**

2. Let us consider (horizontally and vertically) bounded functions: fix an $E \in \mathcal{F}$ with $\mu(E) < \infty$, and fix $M \in \mathbb{R}$, and let us consider $f\colon (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ satisfying $|f| \le M$ and $f(\omega) = 0$ for all $\omega \in \Omega \setminus E$.

   We approximate f from below by simple functions φ.

   We define

$$\int f\,d\mu \stackrel{\text{def}}{=} \sup_{\substack{\varphi \le f \\ \varphi \text{ simple} \\ \operatorname{supp}(\varphi) \subseteq E}} \left\{ \int \varphi\,d\mu \right\}. \qquad \triangle$$

Lemma 3.5.3. *With the assumptions on f as above, we have*

$$\int f\,d\mu = \inf_{\substack{\psi \ge f \\ \psi \text{ simple} \\ \operatorname{supp}(\psi) \subseteq E}} \left\{ \int \psi\,d\mu \right\}.$$

Proof. By property (iv), we have

$$\int \varphi\,d\mu \le \int \psi\,d\mu$$

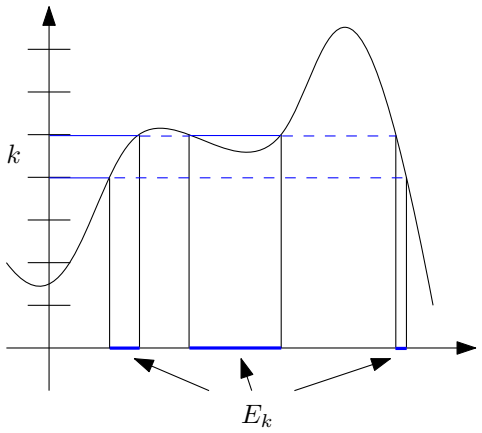
for all $\varphi \le f \le \psi$. Hence we have

$$\sup_{\varphi} \left\{ \int \varphi\,d\mu \right\} \le \inf_{\psi} \left\{ \int \psi\,d\mu \right\}$$

To prove the other inequality, we fix n and let

$$E_k = \left\{ x \in E\colon \frac{kM}{n} \ge f(x) > \frac{(k-1)M}{n} \right\}$$

so that we're partitioning the y-axis into equal intervals, and taking the preimage, as below:



For $k \in \{-n, \dots, n\}$, we let

$$\psi_n = \sum_{k=-n}^{n} \frac{kM}{n} \mathbb{1}_{E_k} \quad \text{and} \quad \varphi_n = \sum_{k=-n}^{n} \frac{(k-1)M}{n} \mathbb{1}_{E_k}.$$

By construction we have $\varphi_n \le f \le \psi_n$. Note that we have $\psi_n - \varphi_n = \frac{M}{n} \mathbb{1}_E$, and hence

$$\int \psi_n\,d\mu - \int \varphi_n\,d\mu = \int (\psi_n - \varphi_n)\,d\mu = \int \frac{M}{n} \mathbb{1}_E\,d\mu = \frac{M}{n}\mu(E) \downarrow 0 \text{ as } n \to \infty. \qquad \square$$

Note that the Riemann integral insists on breaking up the x-axis and then approximating the area under the curve with "vertical" rectangles. The Lebesgue integral breaks up the x-axis into the E_k, and this base is adapted to the function f. This is one reason why the Lebesgue integral is more flexible.

In the proof of Lemma 3.5.3, we used that f is measurable when we asserted φ_n and ψ_n were simple: we need $E_k = f^{-1}((\frac{(k-1)M}{n}, \frac{kM}{n}]) \in \mathcal{F}$.

Let us check properties (i) to (vi) for bounded functions. For property (i), we suppose $f \geq 0$ almost everywhere, and let $N = \{\omega \in \Omega \colon f(\omega) < 0\}$. Since $\mu(N) = 0$, letting $\varphi = -M\mathbb{1}_N$, we have $f \geq \varphi$ everywhere. Thus we get

$$\int f\,d\mu \geq \int \varphi\,d\mu = -M\mu(N) = 0.$$

Let us check property (ii). The case where $a > 0$ is boring, so let us consider the case $a < 0$. We have

$$\int (af)\,d\mu = \sup_{\substack{a\varphi \text{ simple} \\ a\varphi \leq af}} \int (a\varphi)\,d\mu.$$

Note that (since $a < 0$) we have $a\varphi \leq af$ if and only if $\varphi \geq f$. Thus

$$\sup_{\substack{a\varphi \text{ simple} \\ a\varphi \leq af}} \int (a\varphi)\,d\mu = \inf_{\substack{\varphi \text{ simple} \\ \varphi \geq f}} \int (a\varphi)\,d\mu = a \inf_{\substack{\varphi \text{ simple} \\ \varphi \geq f}} \int \varphi\,d\mu$$

where the second equality is property (ii) for integrals of simple functions. Lemma 3.5.3 guarantees that

$$a \inf_{\substack{\varphi \text{ simple} \\ \varphi \geq f}} \int \varphi\,d\mu = a \int f\,d\mu.$$

For property (iii), note that if $\psi_1 \geq f$ and $\psi_2 \geq g$ then $\psi_1 + \psi_2 \geq f + g$. Hence

$$\int (f+g)\,d\mu = \inf_{\substack{\psi \geq f+g \\ \psi \text{ simple}}} \int \psi\,d\mu \leq \inf_{\substack{\psi_1 \geq f \text{ simple} \\ \psi_2 \geq g \text{ simple}}} \int (\psi_1 + \psi_2)\,d\mu = \inf_{\substack{\psi_1 \geq f \text{ simple} \\ \psi_2 \geq g \text{ simple}}} \int \psi_1\,d\mu + \int \psi_2\,d\mu.$$

For the other inequality, do exactly the same proof with $\varphi_1 \leq f$ and $\varphi_2 \leq g$, and use sup instead of inf. (Alternatively, we can use property (ii) and plug in $-f, -g$ into the inequality above.)

As we observed earlier (Lemma 3.5.1), properties (iv)-(vi) follow from (i)-(iii).

Definition 3.5.4 (Integration; cf. Definitions 3.4.1 and 3.5.2)**.**

3. Let us consider nonnegative (but possibly unbounded) functions. For $f \geq 0$ we define

$$\int f\,d\mu \stackrel{\text{def}}{=} \sup\left\{ \int h\,d\mu \colon 0 \leq h \leq f \text{ and } h \text{ bounded} \right\},$$

where h bounded means that there exist $M \in \mathbb{R}, E \in \mathcal{F}$ so that $\mu(E) < \infty$, $h(\omega) = 0$ for all $\omega \in E^c$, and $h \leq M$ everywhere.

△

Lemma 3.5.5. *If $E_n \uparrow \Omega$ with $\mu(E_n) < \infty$, then*

$$\int_{E_n} (f \wedge n) \uparrow \int f\,d\mu.$$

To describe the notation in Lemma 3.5.5, we need to introduce: for real numbers a, b we write $a \wedge b = \min(a, b)$, and for functions f, g we write $(f \wedge g)(\omega) = f(\omega) \wedge g(\omega)$. Now for $A \in \mathcal{F}$ we write

$$\int_A f\,d\mu = \int (f\mathbb{1}_A)\,d\mu.$$

Note that

$$(f\mathbb{1}_A)(\omega) = \begin{cases} f(\omega) & \text{if } \omega \in A \\ 0 & \text{otherwise.} \end{cases}$$

Proof of Lemma 3.5.5. Let $h_n = (f \wedge n)\mathbb{1}_{E_n} \leq f$. Note that

$$\int h_n \, d\mu \leq \int f \, d\mu.$$

For notational conciseness we let $\mathcal{H}$ be the family of functions

$$\mathcal{H} \stackrel{\text{def}}{=} \{h \colon 0 \leq h \leq f \text{ and } h \text{ bounded}\},$$

where "bounded" is in the sense of Definition 3.5.4. Since it is bounded, there is M so that $h \leq M$. Then take $n \geq M$ so that

$$\int (f \wedge n) \, d\mu \geq \int_{E_n} h \, d\mu = \int \Omega h \, d\mu - \int_{E_n^c} h \, d\mu,$$

and

$$\int_{E_n} h \, d\mu \leq M\mu(E_n^c \cap F) \downarrow 0 \text{ as } n \to \infty.$$

For every $h \in \mathcal{H}$, we now have

$$\liminf_{n\to\infty} \int_{E_n} (f \wedge n) \, d\mu \geq \int_{\Omega} h \, d\mu.$$

Passing to the supremum,

$$\liminf_{n\to\infty} \int_{E_n} (f \wedge n) \, d\mu \geq \sup_{h\in\mathcal{H}} \int h, d\mu = \int f \, d\mu.$$

□

With this lemma we'll be able to check properties (i)-(vi), and then check it for general f.

3.6 Sep 23, 2019

We're finishing up the definition of the integral today. We've seen how to integrate simple functions, bounded functions, and nonnegative functions. We note that the integral of a nonnegative function is $\mathbb{R} \cup \{\pm\infty\}$-valued, since it was defined to be the limit of integrals of bounded functions.

Definition 3.6.1. We say a measurable function f is integrable if

$$\int |f| \, d\mu < \infty. \qquad \triangle$$

In this case, we can write $f = f^+ - f^-$, where

$$f^+ \stackrel{\text{def}}{=} f \vee 0; \quad f^- \stackrel{\text{def}}{=} (-f) \vee 0,$$

where the notation $x \vee y$ means $\max(x, y)$, and $(f \vee g)(\omega)$ means the function $f(\omega) \vee g(\omega)$. Then $|f| = f^+ + f^-$.

Definition 3.6.2 (Integration; see Definitions 3.4.1, 3.5.2, and 3.5.4)**.** If f is integrable, we set

$$\int f \, d\mu \stackrel{\text{def}}{=} \int f^+ \, d\mu - \int f^- \, d\mu. \tag{1}$$

$\triangle$

In case $\int f^+ \, d\mu = \infty$ and $\int f^- \, d\mu < \infty$, we can still make sense of equation (1) and set $\int f \, d\mu = \infty$. Likewise, if $\int f^+ \, d\mu < \infty$ and $\int f^- \, d\mu = \infty$, we can still make sense of equation (1) and set $\int f \, d\mu = -\infty$. But if both $\int f^+ \, d\mu = \infty$ and $\int f^- \, d\mu = \infty$, then $\int f \, d\mu$ will be undefined.

We should check again properties (i)-(iii), but this is routine.

[As SL reminded me, the four step process in Definitions 3.4.1, 3.5.2, 3.5.4, and 3.6.2 is sometimes called the standard machine.]

3.7 Sep 23, 2019

[This subsection used to be part of Ch 4, but I feel it is better classified as part of Ch 3. Sorry!] Let's consider special cases of the theory of integration we've developed in Section 3.

1. Let's take $(\Omega, \mathcal{F}, \mu) = (\mathbb{R}, \mathcal{B}, \lambda)$: then
$$\int_{-\infty}^{\infty} f(x)\,dx \stackrel{\text{def}}{=} \int f\,d\lambda.$$
2. For Ω countable and $\mathcal{F} = 2^{\Omega}$, with the counting measure $\mu(A) = \#A$, we have
$$\int f\,d\mu = \sum_{\omega\in\Omega} f(\omega).$$

So one of the things that is nice about the measure theoretic integration in Section 3 is that it unifies the two settings above.

Let's talk about Jensen's inequality.

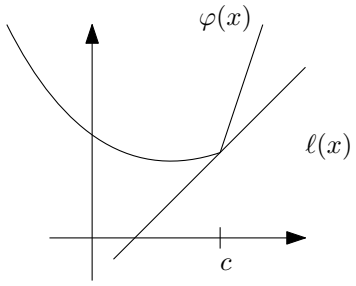
Definition 3.7.1. Recall that $\varphi\colon \mathbb{R} \to \mathbb{R}$ is called convex if
$$\lambda\varphi(x) + (1-\lambda)\varphi(y) \geq \varphi(\lambda x + (1-\lambda)y)$$
for all $x, y \in \mathbb{R}$ and all $\lambda \in [0,1]$. △

Examples of convex functions include x^2, $|x|$, $e^{\alpha x}$, and so on.

Proposition 3.7.2 (Jensen's Inequality)**.** *If μ is a probability measure and φ is convex, with f and $\varphi(f)$ both integrable, then*
$$\varphi\left(\int f\,d\mu\right) \leq \int \varphi(f)\,d\mu.$$

(The intuition of the proof is that $\int f\,d\mu$ is a weighted average of the values of f.)

Proof of Proposition 3.7.2. We claim that for all $c \in \mathbb{R}$, there is a supporting line (at c), call it $\ell(x) = ax + b$, so that $\ell(x) \leq \varphi(x)$ with $\ell(c) = \varphi(c)$, as in the picture below:



Rigorously, the supporting line $\ell(c)$ exists because
$$\lim_{h\uparrow 0} \frac{\varphi(h+c)-\varphi(c)}{h} \quad \text{and} \quad \lim_{h\downarrow 0} \frac{\varphi(h+c)-\varphi(c)}{h}$$
both exist, by convexity of φ. Furthermore,
$$\lim_{h\uparrow 0} \frac{\varphi(h+c)-\varphi(c)}{h} \leq \lim_{h\downarrow 0} \frac{\varphi(h+c)-\varphi(c)}{h}$$
by the convexity of φ.

Let us take $c = \int f\,d\mu$ and a supporting line ℓ at c. Now observe that
$$\int \varphi(f)\,d\mu \geq \int \ell(f)\,d\mu = \int (af+b)\,d\mu = a\int f\,d\mu + \int b\,d\mu = \ell\left(\int f\,d\mu\right) = \varphi\left(\int f\,d\mu\right). \qquad \square$$

Let's also talk about Hölder's inequality.

Definition 3.7.3. Let $1 \leq p < \infty$. We define the L_p norm of a measurable function f as follows:

$$||f||_p \stackrel{\text{def}}{=} \left(\int |f|^p \, d\mu \right)^{\frac{1}{p}}.$$

△

Proposition 3.7.4 (Hölder's Inequality)**.** *For any integrable f, g, with $1 < p < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$, we have*

$$\int |fg| \, d\mu \leq ||f||_p ||g||_q.$$

The case $p = q = 2$ is often called the Cauchy-Schwarz inequality.

Proof. First note that for a constant $a \in \mathbb{R}$, we have

$$||af||_p = |a| \cdot ||f||_p.$$

Furthermore, if $||f||_p = 0$ then $|f|^p = 0$ almost everywhere (cf. HW 2 (!)). This implies in particular that $|fg| = 0$ almost everywhere, so Proposition 3.7.4 holds in this case, too.

In light of the discussion above, we can assume $||f||_p, ||g||_q \neq 0$, and since we can scale by constants, it suffices to consider the case $||f||_p = ||g||_q = 1$.

Now we can use the inequality

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}, \qquad \text{for all } x, y \in \mathbb{R}.$$

This follows from standard calculus (take some partials, and check when they're equal). Then, for $x = |f(\omega)|$ and $y = |g(\omega)|$, we get

$$\int |fg| \, d\mu \leq \int \frac{|f|^p}{p} \, d\mu + \int \frac{|g|^q}{q} \, d\mu = \frac{1}{p} + \frac{1}{q} = ||f||_p ||g||_q,$$

as desired. □

Although the two inequalities we've just proven (Jensen and Hölder) are very nice, in Probability we often have sequences of measurable functions and would like to understand their limit. With this in mind:

Let $(\Omega, \mathcal{F}, \mu)$ be a σ-finite measure space and let $f_1, f_2, \ldots : \Omega \to \mathbb{R}$ be a sequence of functions. When does

$$\lim_{n \to \infty} \int f_n \, d\mu = \int (\lim_{n \to \infty} f_n) \, d\mu? \tag{2}$$

To make this well-posed, we need to talk about what we mean by $\lim_{n \to \infty} f_n$. There is:

- Pointwise convergence everywhere, where

$$\lim_{n \to \infty} f_n \stackrel{\text{def}}{=} f, \qquad \text{where} \qquad f_n(\omega) \to f(\omega) \text{ for all } \omega \in \Omega.$$

- Pointwise convergence almost everywhere, where

$$\lim_{n \to \infty} f_n \stackrel{\text{def}}{=} f, \qquad \text{where} \qquad f_n(\omega) \to f(\omega) \text{ for all } \omega \in E, \text{ with } \mu(\Omega \setminus E) = 0.$$

- Convergence in measure, where

$$\lim_{n \to \infty} f_n \stackrel{\text{def}}{=} f, \qquad \text{where} \qquad \mu(\{\omega \colon |f_n(\omega) - f(\omega)| > \varepsilon\}) \downarrow 0 \text{ as } n \to \infty \text{ for every } \varepsilon > 0.$$

When $\mu(\Omega) = 1$, convergence in measure is sometimes called convergence in probability.

There are other types of convergence (e.g. convergence in L^p), which we'll talk about eventually, but this is enough for now. In HW 3 we'll show that convergence in measure is slightly weaker than pointwise convergence almost everywhere, i.e. we'll show that if $f_n \to f$ almost everywhere then $f_n \to f$ in measure.

Let's talk about potential counterexamples to our question (2).

1. Consider $f_n = \frac{1}{n}\mathbb{1}_{[0,n]}$. Then
$$\int f_n \, d\lambda = \frac{1}{n} \cdot n = 1,$$
even though $f_n \to 0$ pointwise. This is a counterexample because
$$0 = \int (\lim f_n) \, d\lambda \neq \lim_{n\to\infty} \int f_n \, d\lambda = 1.$$

2. Consider $f_n = n\mathbb{1}_{[0,n]}$, where as before
$$\int f_n \, d\lambda = n \cdot \frac{1}{n} = 1,$$
even though $f_n \to 0$ pointwise (a.e.). This is a counterexample for the same reason as before.

Basically, these two examples are representative counterexamples.

Lemma 3.7.5 (Fatou's Lemma)**.** *If $f_n \geq 0$, then*
$$\liminf_{n\to\infty} \int f_n \, d\mu \geq \int (\liminf_{n\to\infty} f_n) \, d\mu.$$

We'll prove this in a bit.

Lemma 3.7.6 (Bounded Convergence Theorem)**.** *Suppose $E \in \mathcal{F}$ with $\mu(E) < \infty$, and suppose $f_n(\omega) = 0$ for all $\omega \notin E$. Suppose also that $|f_n| \leq M$ for some $M \in \mathbb{R}$ independent of n. Then, if $f_n \to f$ in measure, then*
$$\int f_n \, d\mu \to \int f \, d\mu.$$

Proof of Lemma 3.7.6. We have
$$\left| \int f \, d\mu - \int f_n \, d\mu \right| = \left| \int (f - f_n) \, d\mu \right| \leq \int_\Omega |f - f_n| \, d\mu.$$
Fix a $\varepsilon > 0$ independent of n. For each n, we'll split Ω into two sets $\Omega = G_n \sqcup B_n$, where $G_n = \{\omega \colon |f(\omega) - f_n(\omega)| < \varepsilon\}$ and $B_n = \Omega \setminus G_n$. Since $f_n \to f$ in measure, $\mu(B_n) \downarrow 0$ as $n \to \infty$.

Continuing the computation above, we have
$$\int_\Omega |f - f_n| \, d\mu = \int_{G_n} |f - f_n| \, d\mu + \int_{B_n} |f - f_n| \, d\mu \leq \varepsilon\mu(E) + (2M + \varepsilon) \cdot \mu(B_n).$$
As $n \to \infty$, the term $(2M + \varepsilon) \cdot \mu(B_n) \to 0$, and since ε was fixed independent of n, we can send it to 0 and $\varepsilon\mu(E) \to 0$. We arrive at
$$\lim_{n\to\infty} \left| \int f \, d\mu - \int f_n \, d\mu \right| = 0. \qquad \square$$

Proof of Lemma 3.7.5. Note that
$$\liminf_{n\to\infty} f_n = \lim_{n\to\infty} \inf_{n\geq m} f_n.$$
Let us define
$$g_m \stackrel{\text{def}}{=} \inf_{n\geq m} f_n.$$

We have $f_m \geq g_m \geq 0$, and $g_m \uparrow g$, where $g \stackrel{\text{def}}{=} \liminf f_n$. Thus it's enough to show

$$\liminf_{n\to\infty} \int g_n \, d\mu \geq \int g \, d\mu.$$

Let $E_m \uparrow \Omega$ with $\mu(E_m) < \infty$. For a fixed m, note that

$$(g_n \wedge m)\mathbb{1}_{E_m} \to (g \wedge m)\mathbb{1}_{E_m}.$$

By Bounded Convergence (Lemma 3.7.6), we get

$$\int_{E_m} (g_n \wedge m) \to \int_{E_m} (g \wedge m)$$

for fixed m, as $n \to \infty$. We obtain in particular

$$\liminf_{n\to\infty} \int_{E_m} (g_n \wedge m) = \int_{E_m} (g \wedge m)$$

and taking $m \to \infty$ we get

$$\liminf_{n\to\infty} \int_{\Omega} g_n = \int_{\Omega} g.$$

[... is this right?] □

3.8 Sep 25, 2019

[This subsection used to be part of Ch 4, but I feel it is better classified as part of Ch 3. Sorry!] [The office hours are fixed for the rest of the semester. TA Jason's office hours will be on Wednesdays, from 11-1, at 218 MLT. Professor Levine's office hours will be on Thursdays, from 1-2, at 438 MLT.]

We begin today with the monotone convergence theorem, which states the following:

Theorem 3.8.1 (Monotone Convergence)**.** *If $f_n \geq 0$, and $f_n \uparrow f$, then*

$$\int f_n \, d\mu \to \int f \, d\mu.$$

(Recall that $f_n \uparrow f$ means that $f_n(\omega) \uparrow f(\omega)$ for all $\omega \in \Omega$.)

Proof. The hard part is Fatou's Lemma (Lemma 3.7.5). If $f_n \uparrow f$, then

$$\limsup_{n\to\infty} \int f_n \, d\mu \leq \int f \, d\mu \leq \liminf_{n\to\infty} \int f_n \, d\mu,$$

with the second inequality being Lemma 3.7.5. Note that we use $f_n \geq 0$ in this step. We'll see in the HW 3 some ways to weaken the assumption $f_n \geq 0$. □

Theorem 3.8.2 (Dominated Convergence)**.** *If $f_n \to f$ almost everywhere, and $|f_n| \leq g$, where*

$$\int g \, d\mu < \infty,$$

then

$$\int f_n \, d\mu \to \int f \, d\mu.$$

Proof. Observe that $|f_n| \leq g$ implies that $f_n + g \geq 0$. We can use Fatou's Lemma (Lemma 3.7.5) on these functions, that is,

$$\liminf \int (f_n + g) \, d\mu \geq \int \liminf (f_n + g) \, d\mu = \int (f + g) \, d\mu.$$

By linearity, we write

$$\int g \, d\mu + \liminf \int f_n \, d\mu \geq \int g \, d\mu + \int f \, d\mu$$

and the integrals of g cancel. The same argument for $-f_n + g$ gives

$$\limsup \int f_n \, d\mu \leq \int f \, d\mu,$$

and these two inequalities together give the conclusion. □

With these theorems in place, we talk about expected values of random variables. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}^* = \mathbb{R} \cup \{\pm\infty\}$ be a random variable. We define

Definition 3.8.3. The expectation of X is

$$\mathbb{E}X \stackrel{\text{def}}{=} \int_\Omega X \, d\mathbb{P}.$$ △

Note that $\mathbb{E}X$ is defined if $X \geq 0$. In this case, $\mathbb{E}X \in \mathbb{R} \cup \{\infty\}$. It's also defined when X is integrable, in which case $\mathbb{E}X = \mathbb{E}(X^+) - \mathbb{E}(X^-)$. It's undefined when $\mathbb{E}(X^+) = \mathbb{E}(X^-) = \infty$.

Lemma 3.8.4 (Markov's Inequality)**.** *Let $B \in \mathcal{B}$ be a Borel set, and let $\varphi : \mathbb{R} \to [0, \infty)$. Then*

$$\mathbb{P}(X \in B) \leq \frac{\mathbb{E}[\varphi(X)]}{i_B}, \qquad i_B \stackrel{\text{def}}{=} \inf_{x \in B} \{\varphi(x)\}.$$

Proof. Consider the random variables $i_B \mathbb{1}_{\{X \in B\}} \le \varphi(X) \mathbb{1}_{\{X \in B\}} \le \varphi(X)$. Note that these inequalities are pointwise (we use $\varphi \ge 0$ for the second inequality). Hence

$$i_B \mathbb{P}(X \in B) = \int i_B \mathbb{1}_{\{X \in B\}} \, d\mathbb{P} \le \int \varphi(X) \mathbb{1}_{\{X \in B\}} \, d\mathbb{P} \le \int \varphi(X) \, d\mathbb{P} = \mathbb{E}[\varphi(X)],$$

and dividing both sides by i_B gives the desired inequality. □

A very important special case of Markov's inequality is the Chebyshev inequality, obtained by taking $\varphi(x) = x^2$. For fixed $a \in \mathbb{R}$, we have

$$\mathbb{P}(|X| \ge a) \le \frac{\mathbb{E}(X^2)}{a^2}.$$

While Definition 3.8.3 is a good abstract definition, in practice we compute expectations by change of variables. Suppose we have measurable functions

$$(\Omega, \mathcal{F}) \xrightarrow{X} (S, \mathcal{S}) \xrightarrow{f} (\mathbb{R}, \mathcal{B}).$$

Let $\mu(A) = \mathbb{P}(X \in A)$ be the distribution of X [It's a measure on S]. Note that X may not have an expectation (since it may not be real valued). However:

Lemma 3.8.5 (Change of variables)**.** *With notation as above, we have*

$$\mathbb{E}[f(X)] = \int_S f(y) \, \mu(dy).$$

Note that by definition, we have

$$\mathbb{E}[f(x)] \stackrel{\text{def}}{=} \int_\Omega f(X(\omega)) \, \mathbb{P}(d\omega),$$

in particular an integral over Ω rather than over S.

Proof. We use the "four step machine", building up the truth of the lemma for increasingly complicated classes of functions, like how we defined our integral.

1. If $f = \mathbb{1}_B$, for $B \in \mathcal{S}$, chasing definitions gives

$$\mathbb{E}[\mathbb{1}_B(X)] = \mathbb{P}(X \in B) = \mu(B) = \int_S \mathbb{1}_B \, d\mu.$$

2. For simple functions, it suffices to observe that both expectation and integration are linear.

3. For nonnegative functions $f \ge 0$, we truncate

$$f_n(x) = \frac{\lfloor 2^n f(x) \rfloor}{2^n} \wedge n.$$

Then f_n is simple and $f_n \uparrow f$. We have

$$\mathbb{E}[f(X)] = \lim_{n \to \infty} \mathbb{E}[f_n(X)] = \lim_{n \to \infty} \int_S f_n \, d\mu = \int_S f \, d\mu,$$

with the first and last equality being monotone convergence (Theorem 3.8.1).

4. For integrable functions $f = f^+ - f^-$, we use linearity to write

$$\mathbb{E}[f(X)] = \mathbb{E}[f^+(X)] - \mathbb{E}[f^-(X)] = \int_S f^+ \, d\mu - \int_S f^- \, d\mu = \int f \, d\mu.$$ □

Probability gets very interesting when we start talking about independence of random variables. On the measure theory side, this will correspond to a product measure:

Definition 3.8.6. Let $(X, \mathcal{A}, \mu_1)$ and $(Y, \mathcal{B}, \mu_2)$ be σ-finite measure spaces. Let:

$$\begin{aligned}\Omega &= X \times Y = \{(x, y) \colon x \in X, y \in Y\},\\ \mathcal{I} &= \{A \times B \colon A \in \mathcal{A}, B \in \mathcal{B}\},\\ \mathcal{F} &= \sigma(\mathcal{I}).\end{aligned}$$ △

Intuitively, we think of elements of $\mathcal{I}$ as rectangles (think $X, Y = \mathbb{R}$). Durrett uses the notation $\mathcal{A} \times \mathcal{B}$ for $\mathcal{F}$, although we should note that it's not actually a Cartesian product.

Theorem 3.8.7. *There exists a unique measure μ on $(\Omega, \mathcal{A} \times \mathcal{B})$ satisfying*

$$\mu(A \times B) = \mu_1(A)\mu_2(B)$$

for all $A \in \mathcal{A}, B \in \mathcal{B}$.

Example 3.8.8. Let us consider $D = \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 < 1\}$. Note that it is not a rectangle, yet $D \in \mathcal{A} \times \mathcal{B}$, where $\mathcal{A}, \mathcal{B} \subseteq 2^{\mathbb{R}}$, since D can be written as a countable union of almost disjoint rectangles. Note that for any decomposition $D = \sqcup_{n \geq 1} R_n$, we always have

$$\sum_{n \geq 1} \mu(R_n) = \pi.$$ △

Proof of Theorem 3.8.7. Let's prove uniqueness first. Note that $\mathcal{I}$ is a π-system, since

$$(A \times B) \cap (C \times D) = (A \cap C) \times (B \times D).$$

The π-λ theorem (Theorem 1.2.11), or Theorem 1.2.7, imply that if μ, ν are measures agreeing on $\mathcal{I}$ then they agree on $\mathcal{A} \times \mathcal{B}$.

To prove existence, note that

$$(A \times B)^c = A^c \times B \cup A \times B^c \cup A^c \times B^c.$$

Hence the algebra generated by $\mathcal{I}$ is just the sets of the form

$$\bigsqcup_{k=1}^{n} R_k, \qquad R_k \in \mathcal{I}.$$

By the Caratheodory extension theorem (Theorem 1.2.4), it is enough to check that

$$\mu\left(\bigsqcup_{k=1}^{n} R_k\right) = \sum_{k=1}^{n} \mu(R_k)$$

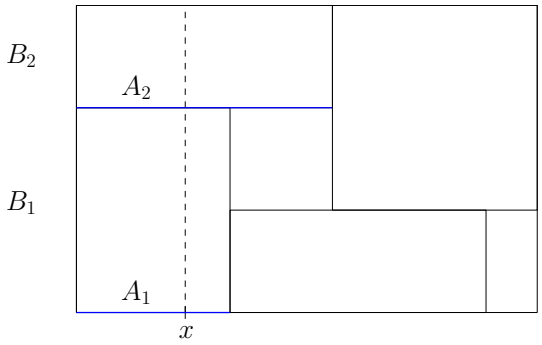
is countably additive on the algebra. Let us suppose that

$$A \times B = \bigsqcup_{i \geq 1} A_i \times B_i;$$

we need to check that

$$\mu_1(A)\mu_2(B) = \sum_{i \geq 1} \mu_1(A_i)\mu_2(B_i).$$

For $x \in A$, let $I(x) = \{i \colon x \in A_i\}$. See the picture below, where we have labelled the first two rectangles $A_1 \times B_1$ and $A_2 \times B_2$ in a decomposition of $A \times B$ into 6 such rectangles:



Observe that

$$\{x\} \in B = \bigsqcup_{i \in I(x)} \{x\} \times B_i.$$

Note that $B = \sqcup_{i \in I(x)} B_i$ implies that

$$\mu_2(B) = \sum_{i \in I(x)} \mu_2(B_i),$$

hence

$$\mathbb{1}_A(x)\mu_2(B) = \sum_{i \geq 1} \mathbb{1}_{A_i}(x)\mu_2(B_i).$$

These are functions of $x \in X$, so we can integrate both sides over X (with respect to $d\mu_1$) to get

$$\int_X \mathbb{1}_A(x)\mu_2(B)\, d\mu_1 = \int_X \sum_{i \geq 1} \mathbb{1}_{A_i}(x)\mu_2(B_i)\, d\mu_1.$$

The left side is just $\mu_1(A)\mu_2(B)$, whereas the right side is

$$\int_X \sum_{i \geq 1} \mathbb{1}_{A_i}(x)\mu_2(B_i)\, d\mu_1 = \sum_{i \geq 1} \left(\int_X \mathbb{1}_{A_i}\, d\mu_1 \right) \mu_2(B_i) = \sum_{i \geq 1} \mu_1(A_i)\mu_2(B_i),$$

where the first equality follows from monotone convergence (Theorem 3.8.1) applied to

$$\sum_{i=1}^{n} \mathbb{1}_{A_i} \uparrow \sum_{i \geq 1} \mathbb{1}_{A_i}.$$

See also HW 3. □

4 Independence

4.9 Sep 30, 2019

We'll talk about Fubini's Theorem today.

Let $(X, \mathcal{A}, \mu_1)$and $(Y, \mathcal{B}, \mu_2)$ be σ-finite measure spaces, and let us define the measure space $(\Omega, \mathcal{F}, \mu)$ with

$$\begin{aligned}\Omega &= X \times Y, \\ \mathcal{F} &= \mathcal{A} \times \mathcal{B} = \sigma(\{A \times B \colon A \in \mathcal{A}, B \in \mathcal{B}\}), \\ \mu &= \mu_1 \times \mu_2 \colon \mathcal{F} \to \mathbb{R} \cup \{\infty\}.\end{aligned}$$

Here μ is the product measure; see Theorem 3.8.7.

Theorem 4.9.1 (Fubini Theorem)**.** *Suppose $f \colon X \times Y \to \mathbb{R}$ is measurable and either $f \geq 0$ or $\int_{X\times Y} |f| \, d\mu < \infty$, so that $\int_{X\times Y} f \, d\mu$ is defined. Then*

$$\int_{X\times Y} f \, d\mu = \int_X \underbrace{\int_Y f(x,y) \, \mu_2(dy)}_{\text{function of } x} \mu_1(dx).$$

The nonnegative f case is due to Tonelli, whereas the integrable f case is due to Fubini. (The bounded f case is due to Lebesgue, and the continuous f case was known to Euler.)

Let's have some sanity checks before proving this theorem. We should definitely have:

(a) For fixed $x \in X$ we'd need the function $h(y) \stackrel{\text{def}}{=} f(x,y)$ is $\mathcal{B}$-measurable, and

(b) The function $g(x) \stackrel{\text{def}}{=} \int f(x,y) \mu_2(dy)$ is $\mathcal{A}$-measurable.

To see (a), consider the commutative diagram

$$y \xmapsto{\iota_x} (x,y) \xmapsto{f} f(x,y), \qquad y \xmapsto{h} f(x,y)$$

so we need to check $\iota_x \colon Y \to X \times Y$ is measurable, in light of Lemma 2.4.8.

Note that

$$\iota_x^{-1}(A \times B) = \{y \colon (x,y) \in A \times B\} = \begin{cases} \emptyset, & x \notin A \\ B, & x \in A \end{cases}$$

This implies ι_x is measurable for every $x \in X$; here we use Exercise 2 on HW 2, which asserted that it suffices to check preimages of a generating set are measurable.

Part (b) is a bit trickier, and it will lead us to the proof of Fubini. Let's see the simplest case first, where $f = \mathbb{1}_E$ for $E \in \mathcal{F}$. Let $E_x = \iota_x^{-1}(E) = \{y \in Y \colon (x,y) \in E\}$. We have $\mathbb{1}_E(x,y) = \mathbb{1}_{E_x}(y)$. Then

$$g_E(x) = \int \mathbb{1}_{E_x}(y) \, \mu_2(dy) = \mu_2(E_x).$$

The following will be useful:

Lemma 4.9.2. *If $E \in \mathcal{F}$ then*

(i) g_E is $\mathcal{A}$-measurable, and

(ii) $\int_X g_E \, \mu_1(dx) = \mu(E)$.

In other words, we can get the product measure of E by integrating the measure of the slices.

Proof Strategy. As usual, the rectangles form a π-system. Let us define

$$\mathcal{L} = \{E \in \mathcal{F} \colon \text{(i) and (ii) hold}\}.$$

We want to check $\mathcal{L}$ is a λ-system and then use π-λ (Theorem 1.2.11).

We claim that if $E = A \times B$ is a rectangle, then $E \in \mathcal{L}$. This is because, as we noted before already,

$$(A \times B)_x = \begin{cases} \emptyset, & x \notin A \\ B, & x \in A \end{cases}$$

Then $g_E(x) = \mu_2(B)\mathbb{1}_A(x)$ is clearly measurable and its integral is

$$\int_X g_E \, d\mu_1 = \mu_2(B) \int_X \mathbb{1}_A(x) \, d\mu_1 = \mu_2(B)\mu_1(A) = \mu(A \times B) = \mu(E).$$

Our next claim is that if $E_n \in \mathcal{L}$ and $E_n \uparrow E$ then $E \in \mathcal{L}$. To see that E satisfies (i), we observe that $(E_n)_x \uparrow E_x$ and hence $g_{E_n}(x) = \mu_2((E_n)_x) \uparrow \mu_2(E) = g_E(x)$. It follows that

$$g_E(x) = \sup_n g_{E_n}(x)$$

is a pointwise supremum of measurable functions, hence is measurable (Lemma 2.4.12). To see that E satisfies (ii), we observe that

$$\int g_E \, d\mu_1 = \lim_{n\to\infty} \int g_{E_n} \, d\mu_1 = \lim_{n\to\infty} \mu(E_n) = \mu(E),$$

where the first equality is monotone convergence theorem (Theorem 3.8.1) and the second equality is because $E_n \in \mathcal{L}$.

Since $(X, \mathcal{A}, \mu_1)$ and $(Y, \mathcal{B}, \mu_2)$ are σ-finite, the previous claim allows us to reduce to the case where $\mu_1(X), \mu_2(Y) < \infty$, since otherwise we can take sequences $X_n \uparrow X$ and $Y_n \uparrow Y$, with $\mu_1(X_n), \mu_2(Y_n) < \infty$ and observe that $X_n \times Y_n \uparrow X \times Y$.

Finally we claim that $\mathcal{L}$ is a λ-system. We need to check that if $E \subseteq F$ with $E, F \in \mathcal{L}$, we have $F \setminus E \in \mathcal{L}$ too. Indeed,

$$g_{F\setminus E}(x) = \mu_2((F \setminus E)_x) = \mu_2(F_x \setminus E_x) = \mu_2(F_x) - \mu_2(E_x) = g_F(x) - g_E(x),$$

where we used $\mu_2(Y) < \infty$ to split $\mu_2(F_x \setminus E_x) = \mu_2(F_x) - \mu_2(E_x)$.

Part (ii) follows from the computation

$$\int g_{F\setminus E} \, d\mu_1 = \int g_F \, d\mu_1 - \int g_E \, d\mu_1 = \mu(F) - \mu(E) = \mu(F \setminus E).$$

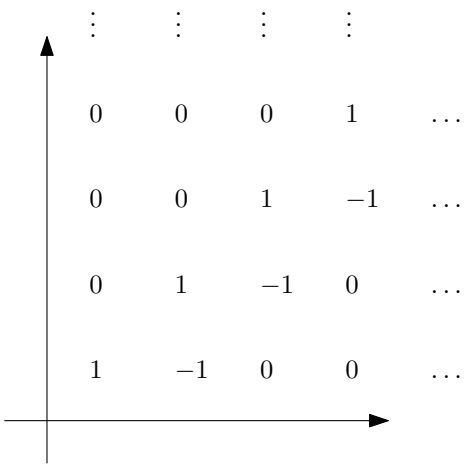
Since $\mathcal{L}$ is a λ-system we can apply Dynkin π-λ (Theorem 1.2.11) to get $\mathcal{F} = \sigma(\{\text{rectangles}\}) \subseteq \mathcal{L}$. □

Proof of Fubini-Tonelli, Theorem 4.9.1. We four-step-machine it:

1. Lemma 4.9.2 gives the theorem for indicators,
2. Linearity gives the theorem for simples,
3. Monotone convergence gives the theorem for nonnegatives,
4. $f = f^+ - f^-$ gives the theorem for integrables. □

Here are two counterexamples to Fubini's theorem when we mess with the assumptions. This is perhaps more important than the proof of Fubini.

1. We may have $\int_X \int_Y f$ and $\int_Y \int_X f$ both finite and unequal! For example, let $X = Y = \mathbb{N}$ and take the counting measure. Consider



Formally, we have

$$f(x,y) = \begin{cases} 1, & x = y \\ -1, & x = y+1 \\ 0, & \text{otherwise} \end{cases}$$

and observe that the sum of the rowsums is 0, whereas the sum of the columnsums is 1. Note that Fubini doesn't apply because f is nonnegative and

$$\int_{X\times Y} |f|\, d\mu = \sum_{(x,y)\in\mathbb{N}^2} |f(x,y)| = \infty.$$

2. For the second counterexample, let $X = Y = (0,1)$, and let μ_1 be the Lebesgue measure and μ_2 be the counting measure. Let us consider

$$f(x,y) = \begin{cases} 1, & x = y \\ 0, & \text{otherwise} \end{cases}$$

Then

$$\int_X f(x,y)\, d\mu_1 = 0 \text{ for all } y \in Y; \quad \text{hence} \quad \int_Y \int_X f(x,y)\, d\mu_1\, d\mu_2 = 0.$$

On the other hand, $\int_Y f(x,y)\, d\mu_2 = 1$.

In this case, it turns out that the product measure μ on $X \times Y$ isn't even defined, because we assumed σ-finiteness in the definition of product of measures (see Theorem 3.8.7). So $\int_{X\times Y} f\, d\mu$ isn't defined, either.

Geometrically, we won't be able to approximate $\{x = y\}$ by finite-measure rectangles in $\mathcal{A} \times \mathcal{B}$.

Let's collect some words together.

Dictionary 4.9.3. *We have the following dictionary between measure theory and probability.*

$$\begin{aligned}
(\Omega, \mathcal{F}, \mu) \text{ measure space} &\longleftrightarrow (\Omega, \mathcal{F}, \mathbb{P}), \mathbb{P}(\Omega) = 1 \text{ probability space} \\
\text{Sets } A \in \mathcal{F} &\longleftrightarrow \text{Events} \\
\text{Length, Area, Volume } \mu(A) &\longleftrightarrow \text{Probability} \\
\text{Measurable functions } f &\longleftrightarrow \text{Random variables } X\colon \Omega \to \mathbb{R} \\
\text{Integrals } \int f\, d\mu &\longleftrightarrow \text{Expectation } \mathbb{E}X = \int X\, d\mathbb{P} \\
\text{Product measures } \mu_1 \times \mu_2 &\longleftrightarrow \text{Independence}
\end{aligned}$$

The new word on this dictionary is independence.

Definition 4.9.4. Events $A, B \in \mathcal{F}$ are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Random variables $X, Y : \Omega \to \mathbb{R}$ are independent if

$$\mathbb{P}(X \in C, Y \in D) = \mathbb{P}(X \in C)\mathbb{P}(Y \in D) \text{ for all } C, D \text{ Borel.}$$

Collections of events $\mathcal{A}, \mathcal{B} \subseteq \mathcal{F}$ are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \text{ for all } A \in \mathcal{A} \text{ and } B \in \mathcal{B}. \qquad \triangle$$

Remark 4.9.5. Note that A and B are independent events if and only if the random variables $\mathbb{1}_A$ and $\mathbb{1}_B$ are independent random variables. Also, X and Y are independent random variables if and only if $\sigma(X), \sigma(Y)$ are independent collections of events. $\triangle$

Definition 4.9.6. The collections of events $\mathcal{A}_1, \ldots \mathcal{A}_n$ are independent if

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \ldots \mathbb{P}(A_{i_k}) \tag{3}$$

for any $1 \leq i_1 \leq \cdots \leq i_k \leq n$ and any $A_{i_j} \in \mathcal{A}_{i_j}$. $\triangle$

Note that $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are independent if and only if $\mathcal{A}_1 \cup \{\Omega\}, \ldots, \mathcal{A}_n \cup \{\Omega\}$ are independent, since we are in a probability space where $\mathbb{P}(\Omega) = 1$. Thus we can assume that $\Omega \in \mathcal{A}_i$ for all i, and Condition (3) is equivalent to

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \prod_{i=1}^{n} \mathbb{P}(A_i), \text{ for all } A_1 \in \mathcal{A}_1, \ldots, A_n \in \mathcal{A}_n.$$

Example 4.9.7. Let us define independent random variables X_1, X_2, X_3 with $\mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = \frac{1}{2}$. Define events

$$A_{ij} = \{X_i = X_j\}, \text{ for } i, j \in [3],$$

and observe that

$$\mathbb{P}(A_{12} \cap A_{23}) = \mathbb{P}(\{X_1 = X_2 = X_3\}) = \frac{1}{4} = \mathbb{P}(A_{12})\mathbb{P}(A_{23}).$$

However, as a triple, they're not independent, since

$$\mathbb{P}(A_{12} \cap A_{13} \cap A_{23}) = \mathbb{P}(\{X_1 = X_2 = X_3\}) = \frac{1}{4} \neq \mathbb{P}(A_{12})\mathbb{P}(A_{13})\mathbb{P}(A_{23}),$$

so the events $\{A_{12}, A_{13}, A_{23}\}$ are pairwise independent but not independent. $\triangle$

4.10 Oct 2, 2019

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Theorem 4.10.1. *If $\mathcal{A}_1, \dots, \mathcal{A}_n \subseteq \mathcal{F}$ are π-systems which are also independent collections of events, then $\sigma(\mathcal{A}_1), \dots, \sigma(\mathcal{A}_n)$ are independent collections of events.*

Proof. The proof boils down to the π-λ theorem (Theorem 1.2.11). Fix $A_2 \in \mathcal{A}_2, \dots, A_n \in \mathcal{A}_n$. Let $F = A_2 \cap \cdots \cap A_n$. Let

$$\mathcal{L} \stackrel{\text{def}}{=} \{A \in \mathcal{F} \colon \mathbb{P}(A \cap F) = \mathbb{P}(A)\mathbb{P}(F)\}.$$

We want to show that $\sigma(\mathcal{A}_1) \subseteq \mathcal{L}$. By assumption, we are given $\mathcal{A}_1 \subseteq \mathcal{L}$, since

$$\mathbb{P}(A \cap F) = \mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2)\dots\mathbb{P}(A_n) = \mathbb{P}(A_1)\mathbb{P}(F), \qquad \text{for all } A_1 \in \mathcal{A}_1.$$

Let us check that $\mathcal{L}$ is a λ-system. For $A \subseteq B$ with $A, B \in \mathcal{L}$, we should check that $B \setminus A \in \mathcal{L}$. (This is property (i) of a λ-system.)

Indeed, because $(B \setminus A) \cap F = (B \cap F) \setminus (A \cap F)$, we have

$$\mathbb{P}((B \setminus A) \cap F) = \mathbb{P}(B \cap F) - \mathbb{P}(A \cap F) = \mathbb{P}(B)\mathbb{P}(F) - \mathbb{P}(A)\mathbb{P}(F) = \mathbb{P}(B \setminus A)\mathbb{P}(F).$$

Now suppose $B_k \in \mathcal{L}$ with $B_k \uparrow B$. We should check $B \in \mathcal{L}$. (This is property (ii) of a λ-system.)

Indeed, observe that $B_k \cap F \uparrow B \cap F$, hence $\mathbb{P}(B_k \cap F) \uparrow \mathbb{P}(B \cap F)$. Since $\mathbb{P}(B_k) \uparrow \mathbb{P}(B)$, we also get $\mathbb{P}(B_k)\mathbb{P}(F) \uparrow \mathbb{P}(B)\mathbb{P}(F)$. (Since limits are unique,) we get $\mathbb{P}(B \cap F) = \mathbb{P}(B)\mathbb{P}(F)$, and $B \in \mathcal{L}$ as desired.

By the π-λ theorem, $\sigma(\mathcal{A}_1) \subseteq \mathcal{L}$. This shows that $\sigma(\mathcal{A}_1), \mathcal{A}_2, \dots, \mathcal{A}_n$ are independent. Repeating this $n-1$ times for $\mathcal{A}_2$, $\mathcal{A}_3$, and so on, shows that $\sigma(\mathcal{A}_1), \sigma(\mathcal{A}_2), \dots$ are independent. □

Corollary 4.10.2. *The real-valued random variables $X_1, \dots, X_n$ are independent (equivalently, the collections of events $\sigma(X_1), \dots, \sigma(X_n)$ are independent [see Remark 4.9.5]) if and only if*

$$\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \le x_i) \text{ for all } x_1, \dots, x_n \in \mathbb{R}.$$

If the X_i were extended-real valued independent random variables, then we should check the condition for $x_i \in \mathbb{R}^* = \mathbb{R} \cup \{\pm\infty\}$.

Proof. Because the $\{X_i \le x\}$ are π-systems generating $\sigma(X_i)$, Theorem 4.10.1 guarantees that all the $\sigma(X_i)$ are independent. □

Theorem 4.10.3. *If $\{X_{ij} \colon 1 \le i \le n, 1 \le j \le m_i\}$ are independent, then so are $\{Y_i \colon 1 \le i \le n\}$, where $Y_i = f_i(X_{i1}, \dots, X_{im_i})$ for measurable functions $f_i \colon \mathbb{R}^{m_i} \to \mathbb{R}$.*

For example, if $X_1, \dots, X_6$ are independent, then so are $X_1 + X_2$, e^{X_3}, and $X_4 X_5 X_6^2$.

Proof. Let $\mathcal{F}_{ij} = \sigma(X_{ij})$, and say $\mathcal{G}_i = \sigma(\{X_{ij} \colon 1 \le j \le m_i\}) = \sigma(\cup_{j=1}^{m_i} \mathcal{F}_{ij})$. We are given that the $\{\mathcal{F}_{ij}\}$ are independent, and we want to show that the $\{\mathcal{G}_i\}$ are independent. Let

$$\mathcal{A}_i = \left\{ \bigcap_{j=1}^{m_i} A_j \colon A_j \in \mathcal{F}_{ij} \right\}.$$

By definition, $\mathcal{A}_i$ is a π-system for each i. Furthermore, $\mathcal{A}_i$ contains $\mathcal{F}_{ij}$ for all j, and hence contains $\cup_{j=1}^{m_i} \mathcal{F}_{ij}$. Since the $\{\mathcal{F}_{ij}\}$ are independent, so is $\{\mathcal{A}_i\}$. By Theorem 4.10.1, the $\{\sigma(\mathcal{A}_i)\}$ are independent too. We are done since $\mathcal{G}_i = \sigma(\mathcal{A}_i)$. □

Theorem 4.10.4. *If $X_1, \dots, X_n$ are independent and X_i has distribution μ_i, then $(X_1, \dots, X_n)$ has distribution $\mu_1 \times \cdots \times \mu_n$.*

Proof. We have

$$\mathbb{P}((X_1,\dots,X_n)\in A_1\times\cdots\times A_n)=\mathbb{P}(X_1\in A_1)\dots\mathbb{P}(X_n\in A_n)=\mu_1(A_1)\dots\mu_n(A_n)=(\mu_1\times\cdots\times\mu_n)(A_1\times\cdots\times A_n).$$

Since the rectangles $A_1\times\cdots\times A_n$ are a π-system generating $\mathcal{B}(\mathbb{R}^n)$, we are done by Theorem 4.10.1. □

Definition 4.10.5. If X is a random variable, we write $X\sim\mu$ to say that X has distribution μ. △

Theorem 4.10.6. *If X and Y are independent random variables, with $X\sim\mu$ and $Y\sim\nu$, and $h\colon\mathbb{R}^2\to\mathbb{R}$ is measurable with either $h\geq 0$ or $\mathbb{E}|h(X,Y)|<\infty$, then*

$$\mathbb{E}h(X,Y)=\iint h(x,y)\mu(dx)\nu(dy).$$

Proof. We have

$$\mathbb{E}h(X,Y)=\int_\Omega h(X,Y)\,d\mathbb{P}=\int_{\mathbb{R}^2}h\,d(\mu\times\nu),$$

with the second equality following from change of variables (Lemma 3.8.5); note that $\mu\times\nu$ is the distribution of (X,Y) by Theorem 4.10.4. By Fubini (Theorem 4.9.1), this is equal to

$$\int_{\mathbb{R}}\int_{\mathbb{R}}h(x,y)\,\mu(dx)\nu(dy).$$

□

Remark 4.10.7. Consider the special case $h(x,y)=f(x)g(y)$ for $f,g\colon\mathbb{R}\to\mathbb{R}$ measurable [and independent, right...?], and either $f,g\geq 0$ or $\mathbb{E}|f(X)|,\mathbb{E}|g(Y)|<\infty$. Then

$$\mathbb{E}(f(X)g(Y))=\mathbb{E}f(X)\mathbb{E}g(Y).$$ △

Proof. We have

$$\mathbb{E}(f(X)g(Y))=\iint f(x)g(y)\,\mu(dx)\nu(dy)=\int g(y)\int f(x)\,\mu(dx)\nu(dy)=\int g(y)\mathbb{E}f(X)\,\nu(dy)=\mathbb{E}f(X)\mathbb{E}g(Y).$$

□

Theorem 4.10.8. *If $X_1,\dots,X_n$ are independent and either $X_i\geq 0$ or $\mathbb{E}|X_i|<\infty$, then*

$$\mathbb{E}(X_1\dots X_n)=\prod_{i=1}^n\mathbb{E}X_i.$$

Proof. Let $X=X_1$ and $Y=X_2\dots X_n$. Let $f=g=(x\mapsto|x|)$. If each $\mathbb{E}|X_i|<\infty$ then

$$\mathbb{E}|X_1\dots X_n|=(\mathbb{E}|X_1|)\cdot(\mathbb{E}|X_2\dots X_n|)=\mathbb{E}|X_1|\dots\mathbb{E}|X_n|<\infty,$$

with the second equality by induction, and hence

$$\mathbb{E}(X_1\dots X_n)=\mathbb{E}(X_1)\cdot\mathbb{E}(X_2\dots X_n)=\mathbb{E}(X_1)\dots\mathbb{E}(X_n),$$

with the second equality also by induction. □

Theorem 4.10.8 can fail for infinite products. Consider independent random variables $X_1,X_2,\dots$ given by

$$\mathbb{P}(X_i=0)=\mathbb{P}(X_i=2)=\frac{1}{2}.$$

Then $\mathbb{E}X_i=\frac{1}{2}(0)+\frac{1}{2}(2)=1$ but $\prod X_i=0$ almost surely (!). Thus

$$1=\prod\mathbb{E}X_i\neq\mathbb{E}\prod X_i=0.$$

Note also that independence is a strong condition. We'll see in the homework that there are four independent random variables that are 3-wise independent but not independent; there are examples of $(n-1)$-wise independent random variables which are not n-wise independent.

Definition 4.10.9. Random variables X and Y are uncorrelated if $\mathbb{E}X^2, \mathbb{E}Y^2 < \infty$ and $\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$. △

Pairwise independence implies uncorrelatedness, but not the other way around. For example, consider the random variable

$$(X,Y) = \begin{cases} (1,0) \\ (-1,0) \\ (0,1) \\ (0,-1) \end{cases}$$

each with probability $\frac{1}{4}$. Then $\mathbb{E}X = \mathbb{E}Y = 0$ and $XY = 0$ (so $\mathbb{E}(XY) = 0$ too), and yet

$$0 = \mathbb{P}(X = 0, Y = 0) \neq \mathbb{P}(X = 0)\mathbb{P}(Y = 0) = \frac{1}{4}.$$

Earlier in our discussion of what could go wrong with Theorem 4.10.8 for infinite products, we assumed the existence of countably many independent random variables $X_1, X_2, \dots$. In light of the fact that the property of being independent gets stronger and stronger as we add more random variables, it is a subtle (but nontrivial) point to construct countably many independent random variables. Our next goal is to make this precise.

Let's talk about infinite products of measures. The goal is that given distribution functions $F_1, F_2, \dots$, we want to construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting independent random variables $X_1, X_2, \dots : \Omega \to \mathbb{R}$ such that each X_i has distribution F_i.

In the finite case, $n = 1$ follows from Caratheodory Extension (Theorem 1.2.4), and for other finite n we can take $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$, and $\mathbb{P} = \mu_{X_1} \times \cdots \times \mu_{X_n}$. We use this probability measure because we'd want

$$\mathbb{P}((a_1, b_1] \times \cdots \times (a_n, b_n]) = \mathbb{P}(X_1 \in (a_1, b_1], \dots, X_n \in (a_n, b_n]) = \prod_{i=1}^{n} (F_i(b_i) - F_i(a_i)).$$

In the infinite case, we would need to take $\Omega = \mathbb{R}^{\mathbb{N}} = \{(\omega_1, \omega_2, \dots) : \omega_i \in \mathbb{R}\}$. (This is also called sequence space; cf. $\{0,1\}^{\mathbb{N}}$.)

Definition 4.10.10. A cylinder set in Ω is a set of the form

$$\begin{aligned} A &= (a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R} \times \mathbb{R} \times \dots \\ &= \{\omega \in \Omega : a_1 < \omega_1 \leq b_i, \dots, a_n < \omega_n \leq b_n, \text{ no other restrictions on other } \omega_i\}. \end{aligned}$$

[As before, cf. the case with $\{0,1\}^{\mathbb{N}}$, as we saw in HW 0.] △

Let $\mathcal{F} = \sigma(\mathcal{S})$, where $\mathcal{S}$ denotes the cylinder sets.

Theorem 4.10.11 (Kolmogorov Extension Theorem). *For $n \geq 1$, let $\mathbb{P}_n$ be the probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ such that*

$$\mathbb{P}_{n+1}((a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R}) = \mathbb{P}_n((a_1, b_n] \times \cdots \times (a_n, b_n]) \quad \text{for all } a_i \leq b_i, i \in [n].$$

Then there exists a unique probability measure $\mathbb{P}$ on $(\mathbb{R}^{\mathbb{N}}, \mathcal{F}) = (\mathbb{R}^{\mathbb{N}}, \sigma(\mathcal{S}))$ such that

$$\mathbb{P}(\underbrace{(a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R} \times \mathbb{R} \times \dots}_{\in \mathcal{S}}) = \mathbb{P}_n((a_1, b_1] \times \cdots \times (a_n, b_n]).$$

Although this result seems obvious, it's important that $\mathbb{P}_n$ are *probability* measures, and that they're *real*-valued.

5 Laws of large numbers

5.11 Oct 7, 2019

Let's talk about the different notions of convergence. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $Y_n, Y : \Omega \to \mathbb{R}$ be random variables.

Definition 5.11.1. We say $Y_n \to Y$ in probability if for all $\varepsilon > 0$, we have $\mathbb{P}(|Y_n - Y| > \varepsilon) \to 0$ as $n \to \infty$. △

Definition 5.11.2. We say $Y_n \to Y$ in L^p if $\mathbb{E}|Y_n - Y|^p \to 0$ as $n \to \infty$. △

Lemma 5.11.3. *If $Y_n \to Y$ in L^p for any p, then $Y_n \to Y$ in probability.*

Proof. We can reduce to the case $Y = 0$. Then $Y_n \to Y$ if and only if $Y_n - Y \to 0$. Markov's inequality (Lemma 3.8.4) says that

$$\mathbb{E}|Y_n|^p \geq \varepsilon^p \mathbb{P}(|Y_n| > \varepsilon).$$

Thus if $Y_n \to 0$ in L^p, we get $\mathbb{E}|Y_n|^p \to 0$, hence $\mathbb{P}(|Y_n| > \varepsilon) \to 0$ as well. This means that $Y_n \to 0$ in probability. □

Theorem 5.11.4 (L^2 weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be uncorrelated (this means $\mathbb{E}X_i < \infty$, and $\mathbb{E}(X_i X_j) = \mathbb{E}X_i \cdot \mathbb{E}X_j$), with $\mathbb{E}X_i = \mu$ for all i. Assume further that $\mathrm{Var}(X_i) \leq C < \infty$ (see Definition 5.11.5). Let $S_n = X_1 + \cdots + X_n$. Then,*

$$\frac{S_n}{n} \to \mu \quad \text{in probability.}$$

Proof. We have

$$\mathbb{E}\left(\frac{S_n}{n}\right) = \frac{1}{n}\mathbb{E}S_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \frac{1}{n}\cdot n \cdot \mu = \mu.$$

Furthermore,

$$\mathbb{E}\left(\frac{S_n}{n} - \mu\right)^2 = \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(S_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i),$$

where the last equality above uses the fact that the X_i are uncorrelated. Recall that we have a uniform bound $\mathrm{Var}(X_i) \leq C$ for all i. Hence, we get

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \leq \frac{1}{n^2}(Cn) = \frac{C}{n} \to 0$$

Hence $\frac{S_n}{n} \to \mu$ in L^2, and Lemma 5.11.3 implies the claim. □

Let's talk more about variance.

Let X be a random variable with $\mathbb{E}X^2 < \infty$. Then, $(\mathbb{E}X)^2 \leq \mathbb{E}X^2$ by Jensen's inequality (Proposition 3.7.2). In particular we have $\mathbb{E}|X| < \infty$.

Definition 5.11.5. The variance of X, denoted $\mathrm{Var}(X)$, is by definition

$$\mathrm{Var}(X) \stackrel{\text{def}}{=} \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X^2) - 2(\mathbb{E}X)(\mathbb{E}X) + (\mathbb{E}X)^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2.$$ △

Thus $\mathrm{Var}(X)$ is measures how "spread out" the random variable is. Notice:

1. The variance $\mathrm{Var}(X)$ is always nonnegative (by Jensen; Proposition 3.7.2).
2. We have $\mathrm{Var}(X) = 0$ if and only if $\mathbb{P}(X = \mu) = 1$. In words we say that X is almost surely constant, or X is an a.s. constant.
3. Similarly, $\mathrm{Var}(X) = \mathbb{E}X^2$ if and only if $\mu = 0$.

4. If $Y = aX + b$,then $\mathbb{E}Y = a\mathbb{E}X + b$, and hence

$$(Y - \mathbb{E}Y)^2 = (aX - a\mathbb{E}X)^2 = a^2(X - \mathbb{E}X)^2.$$

In particular, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

Note that item 4 above was what really made the proof of Theorem 5.11.4 work.

Lemma 5.11.6. *If $X_1, \ldots, X_n$ are uncorrelated, then*

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i).$$

Proof. Let $Y_i = X_i - \mathbb{E}X_i$. Note that $\mathbb{E}Y_i = 0$. Since we are shifting by a constant, we have

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \mathrm{Var}\left(\sum_{i=1}^n Y_i\right).$$

Since $\sum Y_i$ has mean zero, the variance is the expectation of the square, that is,

$$\mathrm{Var}\left(\sum_{i=1}^n Y_i\right) = \mathbb{E}\left(\sum_{i=1}^n Y_i\right)^2 = \mathbb{E}\left(\sum_{i=1}^n \sum_{j=1}^n Y_i Y_j\right) = \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}(Y_i Y_j) = \sum_{i=1}^n \mathbb{E}Y_i^2 + \underbrace{\sum_{i \neq j} (\mathbb{E}Y_i) \cdot (\mathbb{E}Y_j)}_{=0}.$$

Note that $\mathbb{E}Y_i^2 = \mathrm{Var}(Y_i) = \mathrm{Var}(X_i)$, and the claim follows. The last equality uses the fact that the Y_i are uncorrelated; this follows from the more general claim that if X and Y are uncorrelated, then the translations $X + a$ and $Y + b$ (with $a, b \in \mathbb{R}$) are uncorrelated too. □

Remark 5.11.7. Later we will prove the strong law of large numbers (cf. the weak law, Theorem 5.11.4), which says that $\frac{S_n}{n} \to \mu$ almost surely. In other words, we have $\mathbb{P}(\lim_{n\to\infty} \frac{S_n}{n} = \mu) = 1$.

Let's see an example of a sequence of random variables for which $X_n \to X$ in probability but $X_n \not\to X$ almost surely. Let $A_1, A_2, \ldots$ be independent random variables with $\mathbb{P}(A_n) = \frac{1}{n}$. Let $X_n = \mathbb{1}_{A_n}$. Then

$$\mathbb{P}(|X_n| > \varepsilon) = \frac{1}{n} \to 0.$$

We will be able to prove soon (via the Borel-Cantelli lemma) that

$$\mathbb{P}(\lim_{n\to\infty} X_n = 0) = 0.$$ △

Let's talk about densities.

Definition 5.11.8. A random variable X has a density function f if

$$\mathbb{P}(X \in B) = \int_B f(t)\, d\lambda(t)$$

for all Borel sets $B \in \mathcal{B}$. As usual, λ here denotes the Lebesgue measure. Note that it's enough to check this for a generating set of $\mathcal{B}$, i.e. it's sufficient to check that

$$\mathbb{P}(X \leq x) = \int_{-\infty}^x f(t)\, d\lambda(t).$$ △

Note that any nonnegative function $f\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ with $\int_{\mathbb{R}} f = 1$ is the density function of some random variable. However, some random variables have no density function. (Examples include discrete random variables, the Cantor example from Example 2.3.8, and anything with an atom.)

Lemma 5.11.9. *Let $(S, \mathcal{S}, \mu)$ be a σ-finite measure space. Let $f\colon S \to \mathbb{R}_{\geq 0}$ with $\int f\, d\mu < \infty$. Define*

$$\nu(A) \stackrel{\text{def}}{=} \int_A f\, d\mu.$$

Then we have

$$\int g(x)\, d\nu(x) = \int f(x)g(x)\, d\mu(x).$$

Example 5.11.10. Consider

$$(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} \mathbb{R} \xrightarrow{\varphi} \mathbb{R}.$$

Suppose X has density f; recall that

$$\mu_X(B) = \mathbb{P}(X \in B) = \int_B f\, d\lambda.$$

Note also that

$$\mathbb{E}\varphi(X) = \int \varphi\, d\mu_X = \int \varphi(x) f(x)\, d\lambda,$$

with the second equality from Lemma 5.11.9. To make things explicit, suppose $X \sim \text{Unif}[-1, 1]$ which has density $\frac{1}{2}\mathbb{1}_{[-1,1]}$. Suppose also that $\varphi(x) = x^n$. Then,

$$\mathbb{E}\varphi(X) = \int_{-\infty}^{\infty} x^n \frac{1}{2}\mathbb{1}_{[-1,1]}(x)\, d\lambda(x) = \int_{-1}^{1} \frac{1}{2}x^n\, d\lambda(x) = \begin{cases} 0 & n \text{ odd} \\ \frac{1}{n+1} & n \text{ even} \end{cases}$$

Thus, densities make it easy to compute expected values. △

Example 5.11.11. Let $X_1, X_2, \ldots$ be independent uniform distributions on $(-1, 1)$. Let $Y_i = X_i^2$. Then the Y_i are independent. By the previous computation (in Example 5.11.10) we have $\mathbb{E}Y_i = \frac{1}{3}$ and $\mathbb{E}Y_i^2 = \frac{1}{5}$, the L^2 weak law of large numbers (Theorem 5.11.4) says

$$\frac{Y_1 + \cdots + Y_n}{n} \to \frac{1}{3} \quad \text{in probability.}$$

In other words,

$$\mathbb{P}\left(\left|\frac{X_1^2 + \cdots + X_n^2}{n} - \frac{1}{3}\right| > \varepsilon\right) \to 0.$$

Now $X_1^2 + \cdots + X_n^2$ is the squared length of the random vector $(X_1, \ldots, X_n)$; in other words, we pick a random point in the cube $[-1, 1]^n$, and observed that the point is lying very close to the sphere of radius $\sqrt{n/3}$: more precisely,

$$\mathbb{P}\left(\left|||(X_1, \ldots, X_n)||_2^2 - \frac{n}{3}\right| > n\varepsilon\right) \to 0.$$

Stated another way, we can consider the "thin shell" around the sphere of radius $\sqrt{n/3}$ given by

$$A_{n,\varepsilon} = \left\{x \in \mathbb{R}^n\colon (1-\varepsilon)\frac{n}{3} < ||x||_2^2 < (1+\varepsilon)\frac{n}{3}\right\},$$

and we get

$$\frac{\lambda(A_{n,\varepsilon} \cap [-1, 1]^n)}{\lambda([-1, 1]^n)} \to 1.$$

△

5.12 Oct 9, 2019

[Office hours will be on Wednesday 1-2PM in 438 MLT and Friday from 11-1 in 438 MLT.]
We'll prove variations of the Weak Law of Large Numbers (Theorem 5.11.4) today.

Theorem 5.12.1 (Weak Law of Large Numbers with different means). *Let $S_1, S_2, \dots$ be random variables with $\mathbb{E}S_n^2 < \infty$. Denote by $\mu_n = \mathbb{E}S_n$ and $\sigma_n = \sqrt{\mathrm{Var}(S_n)}$. Let (b_n) be any sequence such that $\frac{\sigma_n}{b_n} \to 0$. Then*

$$\frac{S_n - \mu_n}{b_n} \to 0 \quad \text{in probability.}$$

(As you may remember, $\sqrt{\mathrm{Var}(S_n)}$ is called the standard deviation.)

Proof. We compute

$$\mathbb{E}\left(\frac{S_n - \mu_n}{b_n}\right)^2 = \frac{1}{b_n^2}\mathbb{E}(S_n - \mu_n)^2 = \frac{\mathrm{Var}(S_n)}{b_n^2} = \left(\frac{\sigma_n}{b_n}\right)^2 \to 0.$$

Hence $\frac{S_n - \mu_n}{b_n} \to 0$ in L^2, and hence in probability too. □

This theorem allows you to treat combinatorial situations where we almost have, but don't quite have, independence.

Example 5.12.2. Consider:

1. Suppose we are putting n numbered balls in n numbered boxes, with ball k going in box $B_k \sim \mathrm{Unif}\{1, \dots, n\}$, with $B_1, \dots, B_n$ independent. Let

$$N_n = \#(\underbrace{\{1, \dots, n\} \setminus \{B_1, \dots, B_n\}}_{\text{empty boxes}}),$$

so that N_n is the number of empty boxes; note that $0 \le N_n \le n - 1$.

Let's find $\mathbb{E}N_n$ and $\mathrm{Var}(N_n)$. Let

$$A_i = \{B_k \neq i \text{ for all } k\},$$

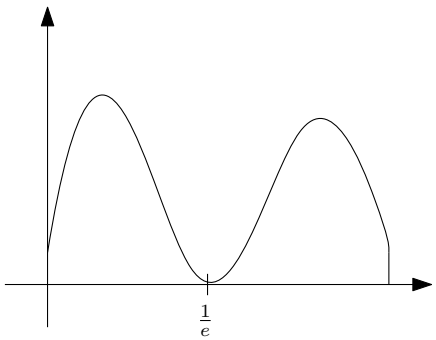
i.e., A_i is the event that box i is empty. Then $N_n = \mathbb{1}_{A_1} + \cdots + \mathbb{1}_{A_n}$, and

$$\mathbb{E}N_n = \sum_{i=1}^n \mathbb{E}\mathbb{1}_{A_i} = \sum_{i=1}^n \mathbb{P}(A_i) = n\left(1 - \frac{1}{n}\right)^n,$$

where $\mathbb{P}(A_i) = (1 - \frac{1}{n})^n$ follows from independence of B_k. In particular, we see that

$$\frac{\mathbb{E}N_n}{n} \to \frac{1}{e}. \tag{4}$$

Our goal is to show that the sequence of random variables $\frac{N_n}{n} \to \frac{1}{e}$ in probability. Note that Equation (4) is a statement about a sequence of real numbers, while our claim is a statement about limiting distributions. In particular, a limiting picture like below would be consistent with $\frac{\mathbb{E}N_n}{n} \to \frac{1}{e}$, but not with $\frac{N_n}{n} \to \frac{1}{e}$.



To rule out limiting picture that look like the above, one needs to control the variance. The idea is to use the fact that $\mathbb{1}_{A_i}$ are *nearly* uncorrelated. In particular, we have

$$\mathbb{P}(A_k \cap A_\ell) = \left(1 - \frac{2}{n}\right)^n \to \frac{1}{e^2} \quad \text{and} \quad \mathbb{P}(A_k)\mathbb{P}(A_\ell) = \left(1 - \frac{1}{n}\right)^{2n} \to \frac{1}{e^2},$$

so even though they're not the same, they have the same limit. Thus,

$$\mathbb{E}(N_n^2) = \mathbb{E} \sum_{k,\ell=1}^{n} \mathbb{1}_{A_k} \mathbb{1}_{A_\ell} = \sum_{k,\ell} \mathbb{P}(A_k \cap A_\ell),$$

whereas

$$(\mathbb{E}N_n)^2 = \left(\sum_k \mathbb{P}(A_k)\right)^2 = \sum_{k,\ell} \mathbb{P}(A_k)\mathbb{P}(A_\ell).$$

In particular, the variance is equal to

$$\operatorname{Var}(N_n) = \sum_{k \neq \ell} \left(\left(1 - \frac{2}{n}\right)^n - \left(1 - \frac{1}{n}\right)^{2n} \right) + \sum_k \left(\left(1 - \frac{1}{n}\right)^2 - \left(1 - \frac{1}{n}\right)^{2n} \right).$$

Note that both sums above have at most n^2 terms and are independent of k and ℓ. With notation as in Theorem 5.12.1 we set $b_n = n$, and observe that

$$\frac{\operatorname{Var}(N_n)}{n^2} \to 0,$$

since $(1 - \frac{2}{n})^n$ and $(1 - \frac{1}{n})^{2n}$ both converge to $\frac{1}{e^2}$.

2. Suppose we are trying to collect coupons (or toys from cereal boxes). Formally, for $n \geq 1$ define independent uniform distributions $X_1, X_2, \cdots \sim \{1, \ldots, n\}$. Let

$$T_n = \min\{m \colon \{X_1, \ldots, X_m\} = \{1, \ldots, n\}\}$$

denote the time required to collect all n coupons. For example, with $n = 4$ we might have

m	1	2	3	4	5	6	7	8	9	...
X_m	1	2	2	1	4	1	3	3	4	...

and in this example $T_4 = 7$. Observe that

$$T_n = \sum_{k=1}^{n} G_k,$$

where G_k is additional time to collect the kth coupon, i.e.

$$G_k = T_k - T_{k-1}, \quad \text{where} \quad T_k = \min\{m \colon \#\{X_1, \ldots X_m\} = k\},$$

with $T_0 = 0$ and $T_1 = 1$.

Lemma 5.12.3. *We have* $(X_{T_k+1}, X_{T_k+2}, \ldots) \stackrel{d}{=} (X_1, X_2, \ldots)$ *and is independent of* $(X_1, \ldots, X_{T_{k-1}})$.

(Here, the notation $\stackrel{d}{=}$ means "equal in distribution".)

We have

$$\begin{aligned}\mathbb{P}(G_k > j) &= \mathbb{P}(X_{T_{k-1}+1}, \ldots, X_{T_{k-1}+j} \in \{X_1, \ldots, X_{T_{k-1}}\}) \\ &= \prod_{i=1}^{j} \mathbb{P}(X_{T_{k-1}+i} \in \{X_1, \ldots, X_{T_{k-1}}\}) \\ &= \prod_{i=1}^{j} \left(\frac{k-1}{n}\right) \\ &= \left(\frac{k-1}{n}\right)^j.\end{aligned}$$

Thus, G_k obeys the geometric distribution $\text{Geom}(1 - \frac{k-1}{n})$, which is defined in the following way: if $G \sim \text{Geom}(p)$, then $\mathbb{P}(G = j) = p(1-p)^{j-1}$. Thus

$$\mathbb{P}(G > j) = p(1-p)^{j-1} \sum_{i \geq 1} (1-p)^i = p(1-p)^{j-1} \cdot \frac{1-p}{p} = (1-p)^j.$$

The story to think about is that G is the random variable corresponding to the first heads for a sequence of flips of a coin that comes up heads with probability p. The mean and variance of geometric distributions is not hard to compute:

$$\mathbb{E}G = \sum_{j \geq 1} \mathbb{P}(G \geq j) = \sum_{j \geq 1} (1-p)^{j-1} = \frac{1}{p},$$

and

$$\text{Var}(G) = \frac{1-p}{p^2} \leq \frac{1}{p^2}..$$

Going back to our problem, observe that that Lemma 5.12.3 implies that $G_1, \ldots, G_n$ are independent. Since $G_k \sim \text{Geom}(\frac{n-k+1}{n})$. Thus,

$$\mathbb{E}T_n = \sum_{k=1}^{n} \mathbb{E}G_k = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} = n\left(\sum_{j=1}^{n} \frac{1}{j}\right) \sim n \log n.$$

We also have

$$\text{Var}(T_n) = \sum_{k=1}^{n} \text{Var}(G_k) \leq \sum_{k=1}^{n} \left(\frac{n}{n-k+1}\right)^2 = n^2 \sum_{j=1}^{n} \frac{1}{j^2} \leq n^2 \frac{\pi^2}{6}.$$

In the notation of Theorem 5.12.1, let $b_n = n \log n$ and observe that

$$\frac{\sigma_n}{b_n} = \frac{n\pi/\sqrt{6}}{n \log n} \to 0$$

so Theorem 5.12.1 applies. In particular,

$$\frac{T_n - \mathbb{E}T_n}{n \log n} \to 0 \quad \text{in probability.}$$

In other words,

$$\frac{T_n}{n \log n} \to 1 \quad \text{in probability.}$$

This kind of phenomenon is called "concentration".

△

Definition 5.12.4. A sequence of random variables X_n is concentrated if

$$\frac{X_n}{\mathbb{E}X_n} \to 1 \quad \text{in probability.} \qquad \triangle$$

After the break we will talk about Borel Cantelli lemmas. Here is the setup. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A_1, A_2, \cdots \in \mathcal{F}$. We denote by

$$\{A_n \text{ i.o.}\} \stackrel{\text{def}}{=} \{\omega \in \Omega \colon \omega \in A_n \text{ for infinitely many } n\};$$

here "i.o." stands for infinitely often, and denote by

$$\{A_n \text{ eventually}\} \stackrel{\text{def}}{=} \{\omega \in \Omega \colon \omega \in A_n \text{ for all but finitely many } n\}.$$

The Borel-Cantelli lemma says:

Lemma 5.12.5 (Borel-Cantelli Lemma 1)**.** *If* $\sum_{n\geq 1} \mathbb{P}(A_n) < \infty$*, then* $\mathbb{P}(\{A_n \text{ i.o.}\}) = 0$*.*

5.13 Oct 16, 2019

We'll be talking about Borel-Cantelli lemmas today.

As always, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $A_1, A_2, \cdots \in \mathcal{F}$ be events. We defined the notation

$$\{A_n \text{ i.o.}\} \stackrel{\text{def}}{=} \{\omega \in \Omega \colon \omega \in A_n \text{ for infinitely many } n\} = \bigcap_{N \geq 1} \bigcup_{n \geq N} A_n;$$

here "i.o." stands for infinitely often, and also

$$\{A_n \text{ eventually}\} \stackrel{\text{def}}{=} \{\omega \in \Omega \colon \omega \in A_n \text{ for all but finitely many } n\} = \bigcup_{N \geq 1} \bigcap_{n \geq N} A_n.$$

Observe that

$$\mathbb{1}_{\{A_n \text{ i.o.}\}} = \limsup \mathbb{1}_{A_n} \quad \text{and} \quad \mathbb{1}_{\{A_n \text{ eventually}\}} = \liminf \mathbb{1}_{A_n}.$$

Note that

$$\mathbb{P}(A_n \text{ i.o.}) \geq \limsup \mathbb{P}(A_n) \quad \text{and} \quad \mathbb{P}(A_n \text{ eventually}) \leq \liminf \mathbb{P}(A_n);$$

cf. HW 5, Problem 2. Note also that

$$X_n \to X \text{ a.s.} \qquad \text{if and only if} \qquad \text{for all } \varepsilon > 0,\ \mathbb{P}(|X_n - X| > \varepsilon \text{ i.o.}) = 0;$$

cf. HW 5, Problem 3.

Lemma 5.13.1 (Borel-Cantelli Lemma 1; cf. Lemma 5.12.5)**.** *If* $\sum_{n \geq 1} \mathbb{P}(A_n) < \infty$, *then* $\mathbb{P}(\{A_n \text{ i.o.}\}) = 0$.

Proof. Let

$$N = \sum_{n \geq 1} \mathbb{1}_{A_n},$$

which counts the number of events that occur. Note that

$$\mathbb{E}N = \sum_{n \geq 1} \mathbb{E}\mathbb{1}_{A_n} = \sum_{n \geq 1} \mathbb{P}(A_n) < \infty,$$

where the first equality is monotone convergence (Theorem 3.8.1). It follows that $\mathbb{P}(N = \infty) = 0$, as desired. □

The converse of Lemma 5.13.1 fails. Consider the probability space $([0,1], \mathcal{B}, \mathbb{P})$ with Lebesgue measure, and let $A_n = (0, \frac{1}{n})$. Then

$$\sum \mathbb{P}(A_n) = \sum_{n \geq 1} \frac{1}{n} = \infty$$

but

$$\mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}\left(\bigcap A_n\right) = 0.$$

However, there is a partial converse to lemma 5.13.1:

Lemma 5.13.2 (Borel-Cantelli Lemma 2)**.** *If* $A_1, A_2, \ldots$ *are independent, and* $\sum \mathbb{P}(A_n) = \infty$, *then* $\mathbb{P}(A_n \text{ i.o.}) = 1$.

To warm up for this proof, let us recall

Observation 5.13.3. If $p_n \in [0,1]$ then

$$\sum_{n \geq 1} p_n = \infty \quad \text{implies} \quad \prod_{n \geq 1} (1 - p_n) = 0. \qquad \triangle$$

Proof. The proof boils down to the observation that $1 - x \leq e^{-x}$. Then

$$\prod_{n \geq 1} (1 - p_n) \leq \prod_{n \geq 1} e^{-p_n} = e^{-\sum p_n} = 0. \qquad \square$$

With this observation we can now prove the Borel-Cantelli Lemma 2:

Proof of Lemma 5.13.2. We have

$$\mathbb{P}\Big(\bigcap_{n\geq N} A_n^c\Big) = \prod_{n\geq N} \mathbb{P}(A_n^c) = \prod_{n\geq N} (1-\mathbb{P}(A_n)) = 0,$$

where the first equality is independence of A_i and the last equality is Observation 5.13.3. Taking complements, we have

$$\mathbb{P}\Big(\bigcup_{n\geq N} A_n\Big) = 1 - \mathbb{P}\Big(\bigcap_{n\geq N} A_n^c\Big) = 1$$

for all N. Then

$$\bigcup_{n\geq N} A_n \downarrow \{A_n \text{ i.o.}\}, \qquad \text{so} \qquad 1 = \mathbb{P}\Big(\bigcup_{n\geq N} A_n\Big) \downarrow \mathbb{P}(A_n \text{ i.o.}). \qquad \square$$

Lemma 5.13.4. *If $X_n \to X$ a.s., then $X_n \to X$ in probability.*

Proof. Fix $\varepsilon > 0$. Let

$$A_n = \{\omega \in \Omega \colon |X_n(\omega) - X(\omega)| > \varepsilon\}.$$

Thus (by HW 5 problem 3, mentioned earlier) $\mathbb{P}(A_n \text{ i.o.}) = 0$. Note that

$$\{A_n \text{ i.o.}\} \subseteq \{X_n > X + \varepsilon \text{ i.o.}\} \cup \{X_n < X - \varepsilon \text{ i.o.}\} \subseteq \{X_n \to X\}^c.$$

Then

$$0 = \mathbb{P}(A_n \text{ i.o.}) \geq \limsup \mathbb{P}(A_n),$$

with the inequality from HW 5, Problem 2. We've thus shown $\mathbb{P}(A_n) \to 0$ as $n \to \infty$, which is what it means for $X_n \to X$ in probability. $\square$

Theorem 5.13.5. *The random variables $X_n \to X$ in probability if and only if for every (deterministic) subsequence n_k there is a further (deterministic) subsequence n_{k_j} such that*

$$X_{n_{k_j}} \to X \text{ a.s.}$$

Consider for example independent random variables $X_n = \mathbb{1}_{A_n}$ with $\mathbb{P}(A_n) = \frac{1}{n}$. We have $X_n \to 0$ in probability but not almost surely.

Proof of Theorem 5.13.5. The forward direction is more interesting. Fix $\varepsilon > 0$. For each $j \geq 1$ choose k_j so that

$$\mathbb{P}(|X_{n_{k_j}} - X| > \varepsilon) < 2^{-j}.$$

By Borel-Cantelli 1 (Lemma 5.13.1), $\mathbb{P}(|X_{n_{k_j}} - X| > \varepsilon \text{ i.o.}) = 0$, since $\sum 2^{-j} < \infty$. Hence $X_{n_{k_j}} \to X$ almost surely. (To be fully rigorous, one might have to take nesting subsequences $n_{k_j;\ell}$ for each $\varepsilon = \frac{1}{\ell}$, and then show that $X_{n_{k_\ell;\ell}} \to X$ a.s. as $\ell \to \infty$.) $\square$

Remark 5.13.6. In the setting of Theorem 5.13.5, there *can* exist a random subsequence that does not converge to X almost surely. For example, if A_n are independent events with $\mathbb{P}(A_n) = \frac{1}{n}$, then by Borel-Cantelli 2 (Lemma 5.13.2) we have $\mathbb{P}(A_n \text{ i.o.}) = 1$. Thus with probability 1 there exists a (random!) subsequence $N_1, N_2, \ldots$ so that $X_{N_k} = 1$ for all $k \geq 1$. In particular, this sequence is not converging to 0.

By "with probability 1 there exists", we mean

$$\mathbb{P}\Big(\bigcup_{(n_1,n_2,\ldots)} \{X_{n_k}(\omega) = 1 \text{ for all } k \geq 1\}\Big) = 1.$$

(For example, when flipping infinitely many coins it's possible that they all come up tails, and there is no n_k with $X_{n_k}(\omega) = 1$, but that's a probability zero event.) $\triangle$

Let's talk about convergence in the topological space $(Y, \mathcal{U})$, where $\mathcal{U}$ denotes the open sets of Y. Let us consider $y_n, y \in Y$.

Definition 5.13.7. We say $y_n \to y$ if for all open $U \ni y$, all but finitely many $y_n \in U$. △

In the special case that (Y, d) is a metric space, the subset $U \in \mathcal{U}$ if and only if for all $y \in U$ there exists $\varepsilon > 0$ so that $B(y, \varepsilon) \subseteq U$. Important examples for us include $\mathbb{R}$, $\mathbb{R}^n$, and $L^p(\Omega, \mathcal{F}, \mathbb{P})$.

Lemma 5.13.8. *We have $y_n \to y$ if and only if for all sequences n_k, there exists a subsequence n_{k_j} such that $y_{n_{k_j}} \to y$.*

Corollary 5.13.9. *There does not exist a topology on the set of random variables such that convergence in the sense of Definition 5.13.7 is a.s. convergence of random variables.*

Remark 5.13.10. On the other hand, there *is* a metric so that convergence in the topological space corresponds to convergence in probability. This metric is given by

$$d(X, Y) = \mathbb{E}\left(\frac{|X - Y|}{1 + |X - Y|}\right).$$

[If you were in 6110 with me – this remark is essentially problem 4 on our final (!)] △

Corollary 5.13.11. *If $X_n \to X$ in probability and $f : \mathbb{R} \to \mathbb{R}$ is continuous, then*

(i) $f(X_n) \to f(X)$ in probability, and

(ii) if f is also bounded, then $\mathbb{E}f(X_n) \to \mathbb{E}f(x)$.

Proof. For part (i), we apply (the forward direction of) Theorem 5.13.5: for all subsequences n_k there exist subsubsequences n_{k_j} so that $X_{n_{k_j}} \to X$ almost surely. Then $f(X_{n_{k_j}}) \to f(X)$ almost surely, and by (the backward direction of) Theorem 5.13.5 we have $f(X_n) \to f(X)$ in probability.

For part (ii), bounded convergence (Lemma 3.7.6) says $\mathbb{E}f(X_{n_{k_j}}) = \mathbb{E}f(X)$. Then $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ by Lemma 5.13.8. □

Let's talk about the strong law of large numbers.

Definition 5.13.12. We write $X \overset{\mathrm{d}}{=} Y$ if $\mu_X = \mu_Y$. △

Definition 5.13.13. For $p > 0$, the pth moment of X is $\mathbb{E}|X|^p$. △

In HW 3, we showed that $\mathbb{E}|X|^p < \infty$ then $\mathbb{E}|X|^q < \infty$ for all $q < p$.

Theorem 5.13.14 (Strong Law of Large Numbers)**.** *Let $X_1, \ldots, X_n$ be i.i.d. ("independent and identically distributed", so $X_1 \overset{\mathrm{d}}{=} X_2 \overset{\mathrm{d}}{=} X_3 \overset{\mathrm{d}}{=} \ldots$), with $\mathbb{E}X_1 = \mu$ and $\mathbb{E}X_1^4 < \infty$. Let $S = X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to \mu \quad \text{a.s.}$$

Proof. Without loss of generality, we may assume $\mu = 0$ (we may let $X_i' = X_i - \mu$). Then

$$\mathbb{E}S_n^4 = \mathbb{E}(X_1 + \cdots + X_n)^4 = \sum_{i,j,k,\ell} \mathbb{E}(X_i X_j X_k X_\ell).$$

Many of the terms appearing in the sum are zero: since the X_i are independent, expectations multiply (Theorem 4.10.8). If $X_i X_j X_k X_\ell$ contains a variable in degree 1, then the expectation of that term is zero, so not many terms remain:

$$\mathbb{E}S_n^4 = n\mathbb{E}X_1^4 + n(n-1)\mathbb{E}(X_1^2 X_2)^2 \leq Cn^2$$

for some absolute constant C. Markov's inequality (Lemma 3.8.4) says that

$$\mathbb{P}(|S_n| > n\varepsilon) \cdot (n\varepsilon)^4 \leq \mathbb{E}S_n^4 \leq Cn^2,$$

so

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| > \varepsilon\right) \leq \frac{Cn^2}{(n\varepsilon)^4} = \frac{C}{n^2\varepsilon^4}.$$

Summing the above inequality over n, we see that we may apply the Borel-Cantelli Lemma (Lemma 5.13.1); hence

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| > \varepsilon \text{ i.o.}\right) = 0,$$

that is to say, $\frac{S_n}{n} \to 0$ a.s.. □

5.14 Oct 21, 2019

We begin with a souped-up Borel-Cantelli.

Theorem 5.14.1. *Let $A_1, A_2, \dots$ be pairwise independent events, and assume $\sum_{n\geq 1}\mathbb{P}(A_n) = \infty$. Let $S_n = \mathbb{1}_{A_1} + \cdots + \mathbb{1}_{A_n}$ be the number of events that occur by "time" n. Then*

$$\frac{S_n}{\mathbb{E}S_n} \to 1 \quad \text{a.s..}$$

(cf. Borel-Cantelli 2 (Lemma 5.13.2), which only says $\mathbb{P}(A_n \text{ i.o.}) = 1$.)

When one sees almost sure convergence, especially in the presence of assumptions on independence, one might want to reach for the strong law of large numbers (Theorem 5.13.14).

Let us define $X_n \overset{\text{def}}{=} \mathbb{1}_{A_n}$; note that $\mathbb{E}X_n^4 = \mathbb{E}X_n \leq 1$ because X_n is just an indicator. However, to make the proof work, we needed 4-wise independence of the X_i, so Theorem 5.14.1 doesn't quite follow.

Proof of Theorem 5.14.1. We have

$$\operatorname{Var}(S_n) = \sum_{k=1}^{n}\operatorname{Var}(\mathbb{1}_{A_k}) = \sum_{k=1}^{n}\mathbb{P}(A_k)(1-\mathbb{P}(A_k)) \leq \sum_{k=1}^{n}\mathbb{P}(A_k) = \mathbb{E}S_n.$$

Recall Chebyshev's inequality (which follows from the $\varphi = x^2$ case of Markov's inequality, see Lemma 3.8.4 and the discussion immediately following the proof); for us, the inequality says

$$\mathbb{P}\left(\left|\frac{S_n}{\mathbb{E}S_n} - 1\right| > \delta\right) = \mathbb{P}(|S_n - \mathbb{E}S_n| > \delta\mathbb{E}S_n) \leq \frac{\operatorname{Var}(S_n)}{(\delta\mathbb{E}S_n)^2} \leq \frac{\mathbb{E}S_n}{\delta^2(\mathbb{E}S_n)^2} = \frac{1}{\delta^2\mathbb{E}S_n} \to 0, \tag{5}$$

so we get $S_n/\mathbb{E}S_n \to 1$ in probability. We want to upgrade this convergence to almost sure convergence; the tool we will use to do this is Theorem 5.13.5. Let us define $n_k = \min\{n\colon \mathbb{E}S_n \geq k^2\}$ and $T_k = S_{n_k}$. Then $\mathbb{E}T_k \geq k^2$. Equation (5) says

$$\mathbb{P}\left(\left|\frac{T_k}{\mathbb{E}T_k} - 1\right| > \delta\right) \leq \frac{1}{\delta^2\mathbb{E}T_k} \leq \frac{1}{\delta^2 k^2}.$$

By Borel-Cantelli 1 (Lemma 5.12.5),

$$\mathbb{P}\left(\left|\frac{T_k}{\mathbb{E}T_k} - 1\right| > \delta \text{ i.o.}\right) = 0.$$

Hence,

$$\frac{T_k}{\mathbb{E}T_k} \to 1 \quad \text{a.s..}$$

Note that $\mathbb{P}(A_k) \leq 1$ for all k implies $\mathbb{E}T_k \leq k^2 + 1$ (since the numbers $\{\mathbb{E}S_i\}$ jump up by $\mathbb{P}(A_{i+1}) \leq 1$ when we increment i to $i+1$). If $n_k \leq n < n_{k+1}$, then

$$\frac{S_n}{\mathbb{E}S_n} \leq \frac{S_{n_{k+1}}}{\mathbb{E}S_{n_k}} = \frac{T_{k+1}}{\mathbb{E}T_{k+1}} \cdot \frac{\mathbb{E}T_{k+1}}{\mathbb{E}T_k}.$$

The bounds on $\mathbb{E}T_k$ allow us to control the ratio $\mathbb{E}T_{k+1}/\mathbb{E}T_k$, and $T_{k+1}/\mathbb{E}T_{k+1} \to 1$ almost surely implies

$$\frac{T_{k+1}}{\mathbb{E}T_{k+1}} \cdot \frac{\mathbb{E}T_{k+1}}{\mathbb{E}T_k} \to 1 \quad \text{a.s..}$$

Thus $S_n/\mathbb{E}S_n$ is bounded from above by a sequence converging to 1 almost surely. Likewise,

$$\frac{T_k}{\mathbb{E}T_k} \cdot \frac{\mathbb{E}T_k}{\mathbb{E}T_{k+1}} = \frac{S_{n_k}}{S_{n_{k+1}}} \leq \frac{S_n}{\mathbb{E}S_n},$$

and $\frac{T_k}{\mathbb{E}T_k} \cdot \frac{\mathbb{E}T_k}{\mathbb{E}T_{k+1}} \to 1$ almost surely too. Hence $S_n/\mathbb{E}S_n \to 1$ almost surely. □

Theorem 5.14.2 (Sharp Strong Law of Large Numbers; cf. Theorem 5.13.14)**.** *Let $X_1, X_2, \dots$ be pairwise independent with the same distribution $X_1 \stackrel{\mathrm{d}}{=} X_2 \stackrel{\mathrm{d}}{=} \dots$ with $\mathbb{E}|X_n| < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to \mathbb{E}X_1 \quad \text{a.s..}$$

This version is stronger than last week's version in two ways: require only pairwise independence instead of i.i.d., and only $\mathbb{E}|X_n| < \infty$ instead of $\mathbb{E}|X_n|^4 < \infty$.

Example 5.14.3. There are lots of random variables with $\mathbb{E}|X| < \infty$ and $\mathbb{E}|X|^4 = \infty$. One such example for $\Omega = (0,1)$ is given by $X(\omega) = \omega^{-1/4}$.

Another example is given by taking a random variable so that $\mathbb{P}(X = n) = C/n^3$, for $n \geq 1$; we see $\mathbb{E}X = \sum_{n\geq 1} nC/n^3 < \infty$, and $\mathbb{E}X^4 = \sum n^4 C/n^3 = \infty$. This kind of random variable, with polynomial decay at infinity, is called "heavy-tailed". △

Remark 5.14.4. It's not so much of an exaggeration to say that most tools of probability are suited to light-tail distributions; for example, one likes to take expectations, and when they're infinite, they're not very useful. This poses a problem for practitioners, since many distributions in the real world are in fact heavy-tailed. There are examples of people fitting heavy tailed random variables to Gaussians, causing their models to be inaccurate. △

Example 5.14.5 (Record values)**.** Suppose $X_1, X_2, \dots$ are i.i.d. with *continuous* distribution function F. This implies $\mathbb{P}(X_i = X_j) = 0$ for $i \neq j$, by HW 4, Ex. 4. Let us define

- $A_k = \{X_k = \max\{X_1, \dots, X_k\}\}$ to be the event that the kth trial sets a record,
- $R_n = \sum_{k=1}^n \mathbb{1}_{A_k}$ be the number of new records in n trials, and
- $\pi_n \in \mathfrak{S}_n$ a permutation with $\pi_n(i) = \#\{j \leq n \colon X_j \geq X_i\}$. (Here, $\mathfrak{S}_n$ is the set of all permutations of $[n]$.)

Our claim is that $\pi_n \sim \mathrm{Unif}(\mathfrak{S}_n)$, that is to say, $\mathbb{P}(\pi_n = \sigma) = \frac{1}{n!}$ for all $\sigma \in \mathfrak{S}_n$.

To prove this, note that

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \prod_{i=1}^n \mathbb{P}(X_i \leq x_i) = \prod_{i=1}^n F(x_i) = \prod_{i=1}^n F(x_{\sigma(i)}) = \mathbb{P}(X_{\sigma(1)} \leq x_1, \dots, X_{\sigma(n)} \leq x_n)$$

implies

$$\mathbb{P}(\pi = \sigma) = \mathbb{P}(\pi = \mathrm{id}) \text{ for all } \sigma \in \mathfrak{S}_n.$$

Aside 5.14.6. As an aside, here's how to sample from $\mathrm{Unif}(\mathfrak{S}_n)$ (to *sample* is to generate a random variable with the uniform distribution on $\mathfrak{S}_n$). One might be able to prove various things about distributions, but sampling from that distribution (efficiently) is a hard or sometimes even an open problem.

One method is to generate independent random variables $X_1, \dots, X_n \sim \mathrm{Unif}(0,1)$ and then take π_n as defined in Example 5.14.5.

Another method is to use insertion: begin with the string "1", and in the kth step insert k into the string in one of the k possible places, all equally likely, and independent of the past. At the end we'll have a string $\pi = \pi_1 \dots \pi_n$ consisting of each element of $[n] = \{1, \dots, n\}$ exactly once, giving an element $\pi \in \mathfrak{S}_n$ (by $\pi \colon i \mapsto \pi_i$). One advantage of this method is that it's now clear that if $\pi \sim \mathrm{Unif}(\mathfrak{S}_n)$, then $f(\pi) \sim \mathrm{Unif}(\mathfrak{S}_{n-1})$, where $f(\pi)$ is the permutation of $[n-1]$ obtained by "crossing out" n.

Not all sets are easy to sample from! An example is the set of proper k-colorings of a fixed graph with n vertices. △

Continuing with the example, let us suppose that $A_1, \dots, A_n$ are pairwise independent and $\mathbb{P}(A_k) = \frac{1}{k}$. (Actually, they're independent, but that's not necessary for us.) To see this, observe that

$$A_k = \{\pi_k(k) = k\} = \{\pi_n(k) > \pi_n(i) \text{ for all } i \leq k\},$$

so

$$\mathbb{P}(A_k) = \frac{\#\{\sigma \in \mathfrak{S}_k : \sigma(k) = k\}}{\#\mathfrak{S}_k} = \frac{1}{k}.$$

Note that if $j < k$, then

$$\begin{aligned}
\mathbb{P}(A_j \cap A_k) &= \mathbb{P}(\pi_k(k) = k, \pi_k(j) > \pi_k(i) \text{ for all } i < j) \\
&= \frac{1}{k!}\#\{\sigma \in \mathfrak{S}_k : \sigma(k) = k, \sigma(j) > \sigma(i) \text{ for all } i < j\} \\
&= \frac{1}{k!}\#\{\sigma \in \mathfrak{S}_{k-1} : \sigma(j) < \sigma(i) \text{ for all } i < j\} \\
&= \frac{1}{k}\left(\frac{1}{(k-1)!}\#\{\sigma \in \mathfrak{S}_{k-1} : \sigma(j) < \sigma(i) \text{ for all } i < j\}\right) \\
&= \frac{1}{k}\mathbb{P}(A_j) \\
&= \frac{1}{kj}.
\end{aligned}$$

Because the A_i are pairwise independent, we have

$$\mathbb{E}R_n = \sum_{i=1}^{n} \mathbb{E}\mathbb{1}_{A_i} = \sum_{i=1}^{n} \mathbb{P}(A_i) = 1 + \frac{1}{2} + \cdots + \frac{1}{n} \approx \log n.$$

As a corollary of Theorem 5.14.1, we obtain

$$\frac{R_n}{\log n} \to 1 \quad \text{a.s..}$$ △

5.15 Oct 23, 2019

Today, we'll focus on applications of the (sharp) strong law of large numbers (Theorem 5.14.2).

Let's talk about empirical distribution functions. Suppose we had random variables $X_1, X_2, \dots$ which are i.i.d. with unknown distribution function $F(x) = \mathbb{P}(X_1 \le x)$. Let

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}} = \frac{1}{n}\#\{i \in [n] \colon X_i \le x\}.$$

Theorem 5.15.1 (Glivenko-Cantelli Theorem)**.** *With notation as above, we have $F_n \to F$ uniformly as $n \to \infty$; specifically,*

$$\sup_x |F_n(x) - F(x)| \to 0 \quad \text{a.s.}.$$

(Note that for a fixed x, $|F_n(x) - F(x)|$ is a random variable, so it makes sense to talk about almost sure convergence.)

Proof for the case F is continuous. There are various technical details in the general case, but the result is still true.

Let us fix $x \in \mathbb{R}$. Let $Y_n = \mathbb{1}_{\{X_n \le x\}}$. Then

$$\mathbb{E}Y_n = \mathbb{P}(X_n \le x) = \mathbb{P}(X_1 \le x) = F(x).$$

The strong law of large numbers (Theorem 5.14.2), we obtain

$$F_n(x) = \frac{1}{n}(Y_1 + \cdots + Y_n) \to \mathbb{E}Y_1 \quad \text{a.s.}.$$

In other words,

$$\mathbb{P}(\lim_{n\to\infty} F_n(x) = F(x)) = 1.$$

We'd like to conclude that

$$\mathbb{P}(\lim_{n\to\infty} F_n(x) = F(x) \text{ for all } x \in \mathbb{R}) = 1,$$

but unfortunately this is (naively) an uncountable intersection over $x \in \mathbb{R}$. In situations like these, one might hope to get away with regularity/continuity and the observation that the intersection over $x \in \mathbb{Q}$ has probability 1.

Indeed, let us fix $\varepsilon > 0$. Then continuity of F implies there exist $x_0 < x_1 < \cdots < x_m = +\infty$ with $F(x_{j+1}) - F(x_j) < \varepsilon$. Note that if $x \in [x_{j-1}, x_j]$ then

$$F_n(x) - F(x) \le F_n(x_j) - F(x_{j-1}) \le F_n(x_j) - F(x_j) + \varepsilon \le M_n + \varepsilon,$$

where $M_n = \max_{j\in[m]} |F_n(x_j) - F(x_j)|$. Likewise,

$$F_n(x) - F(x) \ge F_n(x_{j-1}) - F(x_j) \ge F_n(x_{j-1} - F(x_{j-1}) - \varepsilon \ge -M_n - \varepsilon.$$

In other words, we've obtained a uniform bound

$$|F_n(x) - F(x)| \le M_n + \varepsilon.$$

Let $A_j = \{\lim_{n\to\infty} F_n(x_j) = F(x_j)\}$. Then $\mathbb{P}(A_j) = 1$ for all j. This implies

$$\mathbb{P}(A_1 \cap \cdots \cap A_m) = 1;$$

because $A_1 \cap \cdots \cap A_m \subseteq \{M_n < \varepsilon \text{ eventually}\}$,

$$\mathbb{P}(M_n \le \varepsilon \text{ eventually}) = 1.$$

Since $|F_n(x) - F(x)| \le M_n = \varepsilon$, we obtain

$$\mathbb{P}(\sup_x |F_n(x) - F(x)| < 2\varepsilon \text{ eventually}) = 1.$$

Equivalently, this says

$$\sup_x |F_n(x) - F(x)| \to 0 \quad \text{a.s.}.$$

□

Durrett has a proof for the general case, where basically the same proof works but to deal with atoms some x_j are set equal to each other.

Let's talk about renewal theory. Let $X_n \geq 0$ be i.i.d., and let

$$T_n = X_1 + \cdots + X_n \quad \text{and} \quad N_t = \sup\{n \colon T_n \leq t\};$$

think of T_n as the time needed to eat n jars of maple syrup, so that N_t is the number of jars eaten by time t. Suppose $\mathbb{E}X_1 = \mu < \infty$, and further $\mathbb{P}(0 < X_1 < \infty) = 1$.

Theorem 5.15.2 (The Renewal Theorem)**.** *With notation and assumptions as above,*

$$\frac{N_t}{t} \to \frac{1}{\mu} \quad \text{a.s..}$$

In HW 6, we'll show that if $X_n \to X$ a.s. and $N_n \to \infty$ a.s., then $X_{N_n} \to X$ a.s..

Remark 5.15.3. This HW problem is slightly nuanced: note that it can fail for convergence in probability! Consider independent events A_n with $\mathbb{P}(A_n) = \frac{1}{n}$, and let $X_n = \mathbb{1}_{A_n}$. Then $X_n \to 0$ in probability, and

$$N_n = \min\{k > N_{n-1} \colon X_k = 1\}.$$

Then $X_{N_n} = 1$ for all n, so $X_{N_n} \to 1$ in probability. In other words, passing to a subsequence does not preserve limits. △

Proof of Theorem 5.15.2. By the strong law of large numbers (Theorem 5.14.2), we have

$$\frac{T_n}{n} \to \mu \quad \text{a.s..} \tag{6}$$

Also, $N_t \to \infty$ a.s. because $\mathbb{P}(0 < X_1 < \infty) = 1$, hence there exists $\varepsilon > 0$ with $\mathbb{P}(X_1 \geq \varepsilon) \geq \frac{1}{2}$. Chasing definitions of $T_\cdot$ and $N_\cdot$, we also see

$$T_{N_t} \leq t < T_{N_t+1}.$$

Dividing both sides by N_t, we have

$$\frac{T_{N_t}}{N_t} \leq \frac{t}{N_t} < \frac{T_{N_t+1}}{N_t}.$$

Equation (6), along with the HW 6 problem, says that the left side converges to μ almost surely, so it suffices to show that the right side goes to μ almost surely. But, also by Equation (6) and the HW 6 problem, we obtain

$$\frac{T_{N_t+1}}{N_t + 1} \to \mu \quad \text{a.s.,}$$

and since

$$\frac{T_{N_t+1}}{N_t} = \frac{T_{N_t+1}}{N_t + 1} \cdot \underbrace{\frac{N_t + 1}{N_t}}_{\to 1 \text{ a.s..}},$$

we're done. □

In the case $\mu = \mathbb{E}X_1$ is infinite, the renewal theorem (Theorem 5.15.2) also holds, but one needs a different proof, and $\frac{1}{\mu}$ should be interpreted as 0.

(Naively speaking, it seems that we'd need a continuous analogue of HW 6 to make the proof work, that is, one might ask whether N_n can be replaced with N_t, $t \in \mathbb{R}$ in the HW problem above.)

6 Central Limit Theorems

6.15 Oct 23, 2019

To make statements in this section precise, we begin with a crucial definition:

Definition 6.15.1. Distribution functions F_n are said to converge to F weakly, written

$$F_n \to F \quad \text{weakly},$$

when

$$F_n(y) \to F(y) \quad \text{as } n \to \infty, \text{ for all } y \text{ such that } F \text{ is continuous at } y. \tag{7}$$

The random variables X_n are said to converge to X weakly when $F_{X_n} \to F_X$. △

Some people say $X_n \to X$ in distribution to mean the same thing, and some people denote this convergence by $X_n \xrightarrow{\mathrm{d}} X$ or $X_n \implies X$.

We make this definition so that $X + \frac{1}{n} \xrightarrow{\mathrm{d}} X$. The caveat "for all y such that F is continuous at y" in Equation (7) is needed since

$$F_{X+\frac{1}{n}}(y) = \mathbb{P}\left(X + \frac{1}{n} \le y\right) = \mathbb{P}\left(X \le y - \frac{1}{n}\right) = F_X\left(y - \frac{1}{n}\right) \uparrow \mathbb{P}(X < y),$$

whereas

$$F_X(y) = \mathbb{P}(X \le y).$$

The two are equal if and only if F_X is continuous at y.

Example 6.15.2. Let $X_p \sim \mathrm{Geom}(p)$, with $0 < p < 1$. We have

$$\mathbb{P}(X_p = n) = (1-p)^{n-1}p \quad \text{and} \quad \mathbb{P}(X_p \ge n) = (1-p)^{n-1},$$

for $n \in \mathbb{N}$. To ask whether X_p has a limit in the sense of Definition 6.15.1, one needs to normalize. Note that $\mathbb{E}X_p = \sum_{n\ge1}(1-p)^{n-1} = \frac{1}{p}$, so we ask: Does $X_p/\mathbb{E}X_p = pX_p$ have a limit as $p \to 0$?

Observe that

$$\mathbb{P}(pX_p \ge y) = \mathbb{P}\left(X_p \ge \frac{y}{p}\right) = (1-p)^{y/p} = (1 - \frac{1}{\mu})^{y\mu} \to e^{-y}.$$

In other words,

$$pX_p \xrightarrow{\mathrm{d}} X, \quad \text{where } X \sim \mathrm{Exp}(1),$$

with $\mathrm{Exp}(1)$ defined by $\mathbb{P}(X \ge y) = e^{-y}$. △

6.16 Oct 28, 2019

Today we'll talk about densities.

Definition 6.16.1 (cf. Definition 5.11.8)**.** A random variable X has density $f\colon \mathbb{R} \to [0,\infty)$ if $\mathbb{P}(X \in B) = \int_B f\,d\lambda$ for all Borel sets $B \subseteq \mathbb{R}$. Here λ denotes the Lebesgue measure on $\mathbb{R}$. △

In particular, we have

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \le x \le b) = \int_a^b f(x)\,dx.$$

Thus not every random variable X has a density. For example, if $\mathbb{P}(X = a) > 0$ for some $a \in \mathbb{R}$, then X does not have a density. Similarly, the uniform random variable on the Cantor set (see Example 2.3.8, item 4) does not have a density.

Example 6.16.2. Consider:

1. The uniform random variable $\text{Unif}(a,b)$ has density given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if } x \notin [a,b] \end{cases}$$

Note that, as with any function defined by the values of its integrals, f is only defined "up to measure zero", that is if $\widetilde{f} = f$ a.e. then $\widetilde{f}$ is also a density for X.

2. The exponential random variable $\text{Exp}(\alpha)$ has density given by

$$f(x) = \begin{cases} \alpha e^{-\alpha x} & \text{if } x \ge 0 \\ 0 & \text{if } x < 0 \end{cases}$$

We've seen this random variable before: if $X_p \sim \text{Geom}(p)$, then

$$\frac{X_p}{\mathbb{E}X_p} = pX_p \xrightarrow{\text{d}} Y$$

as $p \to 0$; here $Y \sim \text{Exp}(1)$. Note that any density must satisfy

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

3. The star player of the chapter we are starting, the normal distribution $N(\mu,\sigma^2)$ with mean μ and variance σ^2 has density given by

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

△

We can generalize this setup from $(\mathbb{R},\mathcal{B},\lambda)$ to $(\mathbb{R}^n,\mathcal{B}^n,\lambda^n)$: a random vector $X = (X_1,\dots,X_n)$ has density $f\colon \mathbb{R}^n \to [0,\infty)$ if

$$\mathbb{P}(X \in B) = \int_B f\,d\lambda^n;$$

here $\lambda^n = \lambda \times \cdots \times \lambda$ is the n-dimensional Lebesgue measure. As a special case of this setup, let $A \subset \mathbb{R}^n$ with $0 < \lambda^n(A) < \infty$. Then we can talk about

Definition 6.16.3. The uniform measure on A is the measure μ given by

$$\mu(B) \stackrel{\text{def}}{=} \frac{\lambda^n(B \cap A)}{\lambda^n(A)}.$$ △

Then $X \sim \text{Unif}(A)$ intuitively means "pick a point at random in A". Its density (with respect to λ^n) is $\mathbb{1}_A/\lambda^n(A)$.

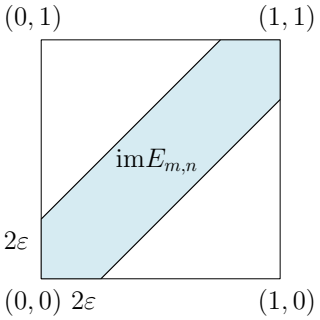
Let's now return to our discussion of weak convergence. We would like to prove the following theorem.

Theorem 6.16.4. *The random variables $X_n \xrightarrow{\text{d}} X$ if and only if $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ for all bounded continuous functions $g\colon \mathbb{R} \to \mathbb{R}$.*

Note that convergence in distribution is a different beast from convergence in probability or convergence almost surely. Let's see some examples illustrating this.

Example 6.16.5. Let $U_1, U_2, \dots$ be i.i.d., uniformly distributed on $(0,1)$. Then $U_n \xrightarrow{\text{d}} U \sim \text{Unif}(0,1)$ trivially, since $F_n = F = \mathbb{1}_{(0,1)}$, so $F_n(x) \to F(x)$ for all x.

But $U_n \not\to U$ in probability, since if $\mathbb{P}(|U_n - U| > \varepsilon) \downarrow 0$ as $n \to \infty$, this would imply, for example, that $\mathbb{P}(|U_n - U_m| > 2\varepsilon) \downarrow 0$ [specifically, the set $E_{m,n} = \{\omega \in \Omega\colon |U_n(\omega) - U_m(\omega)| > 2\varepsilon\}$ will have arbitrarily small measure as long as we pick m, n sufficiently large]. On the other hand, since U_n and U_m are independent random variables, $(U_n(\omega), U_m(\omega))$ is uniform on $(0,1)^2$, so it is not too hard to obtain $\mathbb{P}(E_{m,n}) = (1-2\varepsilon)^2$ from the following geometric picture:



△

Here's a more nontrivial example.

Example 6.16.6. Let's consider a simple random walk on $\mathbb{Z}$, so we have a sequence of random variables S_n, defined by $S_n = X_1 + \cdots + X_n$ with X_i i.i.d. with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = \frac{1}{2}$. Thus

$$\mathbb{E}S_n = \sum_{i=1}^n \mathbb{E}X_i = 0 \quad \text{and} \quad \text{Var}S_n = \sum_{i=1}^n \text{Var}X_i = n.$$

Thus the Strong Law of Large Numbers (Theorem 5.14.2) asserts

$$\frac{S_n}{n} \to 0 \quad \text{a.s..}$$

Usually when we have a limit going to 0, we're not getting as much information as we could (the S_n is getting drowned out by the n). Indeed, note that for $a > \frac{1}{2}$, we have

$$\frac{S_n}{n^a} \to 0 \quad \text{in probability}$$

by Theorem 5.12.1. It turns out that the limit $S_n/n^{1/2}$ will be more interesting. It's possible to compute this limit using our bare hands. We have

$$\mathbb{P}(S_{2n} = 2k) = \mathbb{P}((X_1, \dots, X_{2n}) \text{ has exactly } n+k \text{ many } 1\text{'s}) = \frac{\binom{2n}{n+k}}{2^{2n}} = \frac{(2n)!}{(n+k)!(n-k)!2^{2n}}.$$

Thus Stirling's formula, which says $n! \sim (\frac{n}{e})^n \sqrt{2\pi n}$ (here $f(n) \sim g(n)$ means $\lim_{n\to\infty} f(n)/g(n) = 1$), gives (for k small with respect to n)

$$\begin{aligned}\mathbb{P}(S_{2n} = 2k) &\sim \frac{(2n)^{2n}}{(n+k)^{n+k}(n-k)^{n-k}} \sqrt{\frac{2\pi(2n)}{2\pi(n+k)\cdot 2\pi(n-k)}} \\ &\sim \left(1+\frac{k}{n}\right)^{-(n+k)} \left(1-\frac{k}{n}\right)^{-(n-k)} \sqrt{\frac{1}{\pi n}} \\ &= \left(1-\frac{k^2}{n^2}\right)^{-n} \left(1+\frac{k}{n}\right)^{-k} \left(1-\frac{k}{n}\right)^{k} \frac{1}{\sqrt{\pi n}}.\end{aligned}$$

We know how to compute the limits of each factor above; recall that if $a_n \to 0$, $b_n \to \infty$, and $a_n b_n \to \lambda$, then $(1+a_n)^{b_n} \to e^{\lambda}$. Let x be so that $k = x\sqrt{\frac{n}{2}}$; in other words, so that $k^2/n = x^2/2$. We obtain, when $n \to \infty$ and k is growing so that $x = k/\sqrt{\frac{n}{2}}$ is fixed, the limit

$$\left(1-\frac{k^2}{n^2}\right)^{-n} \left(1+\frac{k}{n}\right)^{-k} \left(1-\frac{k}{n}\right)^{k} \to e^{\frac{x^2}{2}} \cdot e^{-\frac{x^2}{2}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{\pi n}}. \tag{8}$$

Then

$$\begin{aligned}\mathbb{P}\left(a \le \frac{S_{2n}}{\sqrt{2n}} \le b\right) &= \sum_{\substack{\ell \in [a\sqrt{2n}, b\sqrt{2n}], \\ \ell \in 2\mathbb{Z}}} \mathbb{P}(S_{2n} = \ell) \\ &\sim \frac{1}{\sqrt{\pi n}} \sum_{\substack{x \in [a,b], \\ x \in \frac{2}{\sqrt{2\pi}}\mathbb{Z}}} e^{-x^2/2} \\ &\to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx,\end{aligned}$$

which is the integral of the density of $N(0,1)$. (Note that we can apply Equation (8) in the above comptutation since we've scaled S_{2n} by $\sqrt{2n}$.

This shows $\frac{S_{2n}}{\sqrt{2n}} \xrightarrow{\text{d}} N(0,1)$; in the homework we'll prove $\frac{S_{2n+1}}{\sqrt{2n+1}} \xrightarrow{\text{d}} N(0,1)$, and also that $S_n/\sqrt{n}$ does not converge in probability. △

Lemma 6.16.7. *If $F_n \to F$ weakly, then there exist random variables Y_n converging to Y almost surely, with $F_{Y_n} = F_n$ and $F_Y = F$.*

Proof. Let $\Omega = (0,1)$, $\mathcal{B}$ be the Borel sets, and P be the Lebesgue measure. Let

$$Y(\omega) = F^{-1}(\omega) \stackrel{\text{def}}{=} \sup\{y \colon F(y) \le \omega\}.$$

Likewise, we let $Y_n(\omega) = F_n^{-1}(\omega)$.

For $\omega \in (0,1)$, let $a_\omega = \sup\{y \colon F(y) < \omega\}$ and $b_\omega = \inf\{y \colon F(y) > \omega\}$; note that $a_\omega \neq b_\omega$ when the graph of F is horizontal; we have $F(y) = \omega$ on (a_ω, b_ω). Since the intervals (a_ω, b_ω) are disjoint there are only countable many ω such that $a_\omega \neq b_\omega$. Let

$$\Omega_0 = \{\omega \in (0,1) \colon a_\omega = b_\omega\}.$$

We will show $Y_n(\omega) \to Y(\omega)$ for all $\omega \in \Omega_0$. Specifically, we'll first show

$$\liminf_{n\to\infty} F_n^{-1}(\omega) \ge F^{-1}(\omega) \quad \text{and} \quad \limsup_{n\to\infty} F_n^{-1}(\omega) \le F^{-1}(\omega)$$

for all $\omega \in \Omega_0$. Let us choose $y < F^{-1}(\omega)$ such that F is continuous at y. Then $F(y) < \omega$ and $F_n(y) \to F(y)$, so $F_n(y) < \omega$ for sufficiently large n. Then $y \le F_n^{-1}(\omega)$ and taking $\liminf$ we obtain

$$y \le \liminf F_n^{-1}(\omega) \quad \text{for any } y < F^{-1}(\omega), F \text{ continuous at } y.$$

We can take $y \uparrow F^{-1}(\omega)$, avoiding the (at most countably many) discontinuous points along the way. □

6.17 Oct 30, 2019

[OH this week: Jason's OH is in 218 MLT at Wednesday 11:15–1, and Prof Levine's OH is in 438 MLT at Friday 11–12]

Last time we said our goal was to prove Theorem 6.16.4:

Theorem 6.17.1 (cf. Theorem 6.16.4)**.** *The random variables $X_n \xrightarrow{\mathrm{d}} X$ if and only if $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ for all bounded continuous functions $g\colon \mathbb{R} \to \mathbb{R}$.*

Let's prove this now.

Proof. The forward direction follows from Lemma 6.16.7, which guarantees the existence of $Y_n \overset{\mathrm{d}}{=} X_n$ and $Y \overset{\mathrm{d}}{=} X$ (note that this implies $g(Y_n) \overset{\mathrm{d}}{=} g(X_n)$ and $g(Y) \overset{\mathrm{d}}{=} g(X)$), with $Y_n \to Y$ almost surely and $Y_n(\omega) \to Y(\omega)$ for almost all ω. (Here, *almost all* means $\mathbb{P}(\{\omega\colon Y_n(\omega) \to Y(\omega)\}) = 1$.)

The continuity of g implies $Y_n(\omega) \to g(Y(\omega))$ for almost all ω. Hence $\mathbb{E}g(Y_n) \to \mathbb{E}g(Y)$ by Bounded Convergence Theorem (Lemma 3.7.6). Since $\mathbb{E}g(X_n) = \mathbb{E}g(Y_n)$, and $\mathbb{E}g(X) = \mathbb{E}g(Y)$, we obtain the desired $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ for all bounded continuous functions $g\colon \mathbb{R} \to \mathbb{R}$.

For the backward direction, we assume $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ for all bounded and continuous g; we want to show that $\mathbb{P}(X_n \le x) \to \mathbb{P}(X \le x)$ for all continuity points x of $F(x) \overset{\mathrm{def}}{=} \mathbb{P}(X \le x)$.

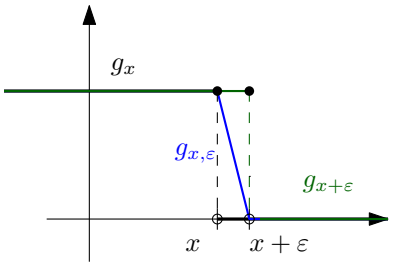
Observe that $\mathbb{P}(X_n \le x) = \mathbb{E}g_x(X_n)$ and $\mathbb{P}(X \le x) = \mathbb{E}g_x(X)$, where

$$g_x(y) = \begin{cases} 1 & \text{if } y \le x \\ 0 & \text{else} \end{cases}$$

which is unfortunately not continuous. So the name of the game is to approximate: we define

$$g_{x,\varepsilon} \overset{\mathrm{def}}{=} \begin{cases} 1 & \text{if } y \le x \\ 1 - \frac{y-x}{\varepsilon} & \text{if } y \in (x, x+\varepsilon) \\ 0 & \text{if } y \ge x + \varepsilon \end{cases}$$

This is indeed bounded and continuous, and is sandwiched in between g_x and $g_{x+\varepsilon}$:



It follows that

$$\mathbb{P}(X_n \le x) = \mathbb{E}g_x(X_n) \le \mathbb{E}g_{x,\varepsilon}(X_n) \to \mathbb{E}g_{x,\varepsilon}(X) \le \mathbb{E}g_{x+\varepsilon}(X) = F(x+\varepsilon).$$

Analogously,

$$\mathbb{P}(X_n \le x) \ge \mathbb{E}g_{x-\varepsilon,\varepsilon}(X_n) \to \mathbb{E}g_{x-\varepsilon,\varepsilon}(X) \ge \mathbb{E}g_{x-\varepsilon}(X) \ge F(x-\varepsilon).$$

If F is continuous at x, then

$$\lim_{\varepsilon \to 0} F(x-\varepsilon) = F(x) = \lim_{\varepsilon \to 0} F(x+\varepsilon).$$

Hence, $\mathbb{P}(X_n \ge x) \to F(x)$, so $X_n \xrightarrow{\mathrm{d}} X$. □

Two remarks:

The theorem can fail for unbounded g, even for $g(x) = x$. In other words, it is possible that $X_n \xrightarrow{\text{d}} X$ but $\mathbb{E}X_n \not\to \mathbb{E}X$. In the homework, we proved that for $X_n = U_1 \cdot \cdots \cdot U_n$, with $U_i \sim \text{Unif}(0, 2.5)$ i.i.d., we have $X_n \to 0$ almost surely (hence $X_n \xrightarrow{\text{d}} 0$, but $\mathbb{E}X_n \to \infty$).

Also, the theorem provides a definition of $X_n \xrightarrow{\text{d}} X$ in the more general setup $\Omega \overset{X_n, X}{\to} S \xrightarrow{g} \mathbb{R}$ for S-valued random variables, where S is any topological space. This will be very important next semester, when we talk about Brownian motion, for example.

Theorem 6.17.2 (Helly Selection)**.** *Let F_n be any sequence of distribution functions. There exists a subsequence $n_1, n_2, \dots$ and an increasing right-continuous F such that $F_n \to F$ weakly.*

(See Definition 6.15.1 for the definition of weak convergence.)

(In HW 2, we showed that the functions which arise as distribution functions of some random variable are precisely the right-continuous increasing functions with $F(-\infty) = 0$ and $F(\infty) = 1$.)

Note that F may not be a distribution function:

Example 6.17.3. Let $X_n \sim \text{Exp}(n)$. The density of X_n is given by

$$f_n(x) = \begin{cases} \frac{1}{n} e^{-nx} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Then the distribution function is given by

$$F_n(x) = \mathbb{P}(X_n \leq x) = \int_{-\infty}^{x} f_n(x)\, dx = \begin{cases} 1 - e^{-nx} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

In this case, we don't need to pass to a subsequence, then as $n \to \infty$ we have $F_n(x)$ is approaching

$$F_n(x) \to \begin{cases} 1 - \lim_{n\to\infty} e^{-nx} = 1 & \text{for all } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

so that $F_n(x) \to \mathbb{1}_{\{x>0\}}$ as $n \to \infty$.

On the other hand, as $n \to 0$ we obtain

$$F_n(x) \to \begin{cases} 1 - \lim_{n\to 0} e^{-nx} = 0 \text{ for all } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

so that $F_n(x) \to 0$ as $n \to 0$. Thus, for the random variables $Y_n \sim \text{Exp}(\frac{1}{n})$, the increasing right-continuous F we approach does not have the correct limits. △

Example 6.17.4. Let $F_n = a\mathbb{1}_{[n,\infty)} + b_{[-n,\infty)} + cG$, where G is any distribution function, and $a, b, c \geq 0$ satisfy $a+b+c = 1$. Since F_n is right increasing, continuous, and has the correct limits, it is the distribution function of some random variable, call it X_n. If G is the distribution of the random variable X, we have

$$X_n = \begin{cases} n & \text{with probability } a \\ -n & \text{with probability } b \\ X & \text{with probability } c \end{cases}$$

Thus as $n \to \infty$ we have $F_n(x) \to b + cG(x)$, and in particular $F_n(\infty) = b + c < 1$ if $a > 0$. △

Proof of Theorem 6.17.2. Let us take an ordering $\mathbb{Q} = \{q_1, q_2, \dots\}$ of the rationals, and let us consider, for each m, the number $F_m(q_1) \in [0, 1]$. Since $[0, 1]$ is compact, there is a convergent subsequence $m(i)$, say $F_{m(i)}(q_1) \to G(q_1)$. There is a convergent subsubsequence $m_2(i)$ so that $F_{m_2(i)}(q_2) \to G(q_2)$; we keep passing to convegent subsequences $m_k(i)$ to obtain values of $G(q_k) = \lim_{i\to\infty} F_{m_k(i)}(q_k)$.

Now let $n_k = m_k(k)$. Then for all $q \in \mathbb{Q}$ we have $F_{n_k}(q) \to G(q)$. Now set, for every $x \in \mathbb{R}$,

$$F(x) = \inf\{G(q) \colon q \in \mathbb{Q}, q > x\}.$$

This function F is right-continuous and increasing.

We want to prove that $F_n \to F$ weakly. Let x be a continuity point of F; for all $\varepsilon > 0$ there exist $r_1, r_2, s \in \mathbb{Q}$ with $r_1 < r_2 < x < s$ such that $F(r_1) \le F(x) \le F(s)$. [to be completed...] □

Definition 6.17.5. A sequence X_n of random variables is tight if for all $\varepsilon > 0$ there exists M_ε such that $\mathbb{P}(|X_n| > M_\varepsilon) < \varepsilon$ for all n. △

Example 6.17.6. The random variables in Example 6.17.4 given by

$$X_n = \begin{cases} n & \text{with probability } a \\ -n & \text{with probability } b \\ X & \text{with probability } c \end{cases}$$

is not tight if $a > 0$ or $b > 0$. The random variables $Z_n \sim N(0, n)$ are not tight, either. △

Definition 6.17.7. A sequence of distribution functions F_n is tight if for all $\varepsilon > 0$ there exists M_ε such that $1 - F_n(M_\varepsilon) + F_n(-M_\varepsilon) < \varepsilon$ for all n. △

Note that if X_n has distribution function F_n, then

$$\begin{aligned} \mathbb{P}(X_n > M) &= 1 - F_n(M) \\ \mathbb{P}(X_n < -M) &= F_n(-M) \end{aligned}$$

imply $\mathbb{P}(|X_n| > M) = 1 - F_n(M) + F_n(-M)$.

Corollary 6.17.8. *Let F_n be any tight sequence of distribution functions. There exists a subsequence $n_1, n_2, \ldots$ and an increasing right-continuous F with $F(\infty) = 1$ and $F(-\infty) = 0$ such that $F_n \to F$ weakly.*

Proof. Fix ε. There exist M_ε so that $\mathbb{P}(|X_n| > M_\varepsilon) < \varepsilon$ for all n, since $F_{n_k} \to F$ weakly by Helly Selection (Theorem 6.17.2). Let $x < -M_\varepsilon$ and $y > +M - \varepsilon$ be contintuity points of F. Then

$$1 - F_{n_k}(y) + F_{n_k}(x) \to 1 - F(y) + F(x)$$

as $k \to \infty$. The quantity $1 - F_{n_k}(y) + F_{n_k}(x)$ is at most

$$1 - F_{n_k}(M_\varepsilon) + F_{n_k}(-M_\varepsilon) < \varepsilon,$$

so $1 - F(y) \to 0$ as $y \to \infty$ and $F(x) \to 0$ as $x \to -\infty$. □

The abstract stuff is out of the way now. Next week we'll study characteristic functions $g(x) = e^{itx}$. This will be useful for proving central limit theorems.

6.18 Nov 4, 2019

Let's talk about characteristic functions today.

Definition 6.18.1. The characteristic function of a real-valued (*not* taking values in $\pm\infty$!) random variable X is the map $\varphi_X : \mathbb{R} \to \mathbb{C}$ defined by

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX) + i\sin(tX)] \stackrel{\text{def}}{=} \mathbb{E}\underbrace{\cos(tX)}_{\in\mathbb{R}} + i\mathbb{E}\underbrace{\sin(tX)}_{\in\mathbb{R}}.$$

In general, for complex-valued random variables $Z : \Omega \to \mathbb{C}$ we may define $\mathbb{E}Z \stackrel{\text{def}}{=} \mathbb{E}(\operatorname{Re} Z) + i\mathbb{E}(\operatorname{Im} Z)$. $\triangle$

As an exercise (see the next HW), show that the following basic properties hold:

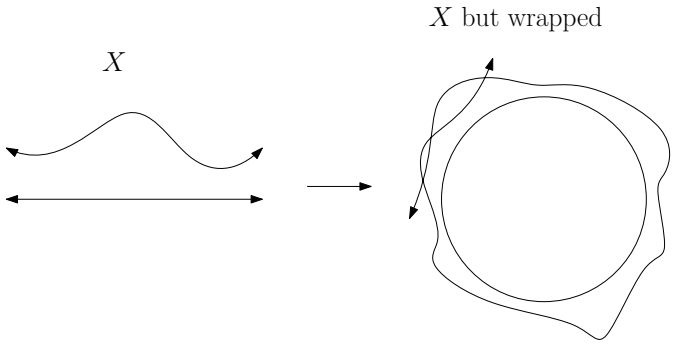
Proposition 6.18.2. *We have*

1. $\mathbb{E}|Z| \geq |\mathbb{E}Z|$
2. $\mathbb{E}(wZ) = w(\mathbb{E}Z)$ *for all* $w \in \mathbb{C}$, *and*
3. *The formula*

$$\int_a^b e^{wx}\,dx = \frac{e^{wb}}{w} - \frac{e^{wa}}{w}$$

holds for all $w \in \mathbb{C} \setminus \{0\}$.

Intuitively, the map $x \mapsto e^{itx}$ wraps the real line around a circle, with t controlling the wrapping rate. Thus the random variable is also wrapped around the circle, and then we take an average:



Proposition 6.18.3. *We have, for* $\varphi = \varphi_X$,

1. $\varphi(0) = 1$
2. $\varphi(-t) = \mathbb{E}[\cos(tX) - i\sin(tX)] = \overline{\varphi(t)}$
3. $|\varphi(t)| \leq 1$ *for all* $t \in \mathbb{R}$.
4. φ *is uniformly continuous on* $\mathbb{R}$.
5. $\varphi_{aX+b}(t) = e^{ibt}\varphi_X(at)$
6. *When* X *and* Y *are independent,* $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$
7. *(Mixtures.) Suppose*

$$X = \begin{cases} Y & \text{with probability } p \\ Z & \text{with probability } 1-p \end{cases}$$

Then $\varphi_X = p\varphi_Y + (1-p)\varphi_Z$.

Proof. For part (3), we may apply Jensen's inequality (Proposition 3.7.2) to estimate

$$|\varphi(t)|^2 = (\mathbb{E}\cos(tX))^2 + (\mathbb{E}\sin(tX))^2 \leq \mathbb{E}(\cos(tX))^2 + \mathbb{E}(\sin(tX))^2 = \mathbb{E}(1) = 1.$$

For part (4), we estimate

$$\begin{aligned} |\varphi(t+h) - \varphi(t)| &= |\mathbb{E}[e^{i(t+h)X} - e^{itX}]| \\ &\leq \mathbb{E}|e^{itX}(e^{ihX} - 1)| \\ &= \mathbb{E}(|e^{itX}| \cdot |e^{ihX} - 1|) \\ &\xrightarrow{\text{BCT}} 0 \qquad \text{as } h \to 0; \end{aligned}$$

here we are crucially using the fact that we're working over $\mathbb{C}$ to use Bounded Convergence Theorem (Lemma 3.7.6).

For part (5), we compute

$$\varphi_{aX+b}(t) = \mathbb{E}e^{it(aX+b)} = \mathbb{E}(e^{i(at)X}e^{ibt}) = e^{ibt}\varphi_X(at),$$

where we are applying Proposition 6.18.2 part (2) to $e^{ibt} \in \mathbb{C}$.

For part (6), we compute

$$\varphi_{X+Y}(t) = \mathbb{E}e^{it(X+Y)} = \mathbb{E}(e^{itX}e^{itY}) = \mathbb{E}e^{itX} \cdot \mathbb{E}e^{itY} = \varphi_X(t)\varphi_Y(t)$$

□

Example 6.18.4. Let us consider $X \sim \text{Exp}(\alpha)$, where $\alpha > 0$ measures the rate; the density of X is given by $d\mu_X = \alpha e^{-\alpha x}\mathbb{1}_{X \geq 0}\, dx$, where dx denotes the Lebesgue measure. Then

$$\begin{aligned} \varphi_X(t) &= \mathbb{E}(e^{itX}) \\ &= \int_{\mathbb{R}} e^{itx} \cdot \alpha e^{-\alpha x}\mathbb{1}_{x \geq 0}\, dx \\ &= \alpha \int_0^\infty e^{x(it-\alpha)}\, dx \\ &= \alpha \left[\frac{e^{(it-\alpha)x}}{it-\alpha}\right]_{x=0}^{x=\infty} \\ &= \alpha \lim_{b\to\infty}\left(\frac{e^{(it-\alpha)b}}{it-\alpha} - \frac{1}{it-\alpha}\right) \\ &= -\frac{\alpha}{it-\alpha}. \end{aligned}$$

△

We'll see how to extract information about a distribution function from a characteristic function. For example, one might ask whether the characteristic function allows you to recover the distribution function:

Theorem 6.18.5 (Inversion Formula)**.** *Let μ be a probability distribution on $\mathbb{R}$. Let*

$$\varphi(t) = \int_{\mathbb{R}} e^{itx}\, d\mu(x).$$

Then:

1. *For all $a < b$, we have*

$$\int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)\, dt = 2\pi\mu((a,b)) + \pi\mu(\{a\}) + \pi\mu(\{b\}).$$

2. *Furthermore,*

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{-ita}\varphi(t)\,dt = \mu(\{a\}).$$

A full proof of this can be found in Durrett.

Corollary 6.18.6. *The characteristic function $\varphi_X(t)$ determines the distribution μ_X.*

Note that if $U \sim \text{Unif}(a,b)$, then its density is given by $\frac{1}{b-a}\mathbb{1}_{(a,b)}$. Then

$$\varphi_U(t) = \int_a^b \frac{e^{itx}}{b-a}\,dx = \left[\frac{e^{itx}}{it(b-a)}\right]_{x=a}^{x-b} = \frac{e^{ibx}-e^{iax}}{it(b-a)}.$$

Suppose μ has a density f. In this case, we may apply Parseval's Theorem from Fourier theory, which says:

Theorem 6.18.7 (Parseval)**.** *If X has density f_X and Y has density f_Y, with both $f_X, f_Y \in L^2$, we have*

$$\int \varphi_X(t)\overline{\varphi_Y(t)}\,dt = \int f_X(t)\overline{f_Y(t)}\,dt.$$

So we take $Y = U \sim \text{Unif}(a,b)$, and part (1) of the Inversion Theorem (Theorem 6.18.5) boils down to Parseval (Theorem 6.18.7) in the case that the latter applies (in which case, μ has no atoms).

Suppose also that X has a density f_X, and f_X is bounded and continuous. Then

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\,dx$$

is the Fourier transform of f_X; the Fourier inversion formula says

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_X(t)\,dt.$$

Here's a crucial theorem.

Theorem 6.18.8. *Let X_n be a sequence of real-valued random variables, and let φ_n denote the characteristic function φ_{X_n}.*

1. *If $X_n \xrightarrow{\text{d}} X$, then $\varphi_n(t) \to \varphi(t)$ for all $t \in \mathbb{R}$, where $\varphi = \varphi_X$;*
2. *If $\varphi_n(t) \to \varphi(t)$ for all $t \in \mathbb{R}$, and φ is continuous at $t = 0$, then there exists a real valued random variable X with $\varphi_X = \varphi$, and $X_n \xrightarrow{\text{d}} X$.*

Often, characteristic functions are easier to work with, e.g. there are no caveats about continuity points since characteristic functions are automatically continuous (Proposition 6.18.3 part 4), and Theorem 6.18.8 provides the translation between characteristic functions and distribution functions.

Proof of Theorem 6.18.8. For part (1), observe first that $g_t(x) = e^{itx}$ is bounded and continuous. So if $X_n \xrightarrow{\text{d}} X$ then $\mathbb{E}g_t(X_n) \to \mathbb{E}g_t(X)$ for all $t \in \mathbb{R}$ by Theorem 6.17.1; it suffices now to observe that $\varphi_n(t) = \mathbb{E}g_t(X_n)$ and $\varphi(t) = \mathbb{E}g_t(X)$.

For part (2), it will be helpful to understand what can go wrong at $t = 0$, so let us consider the following example.

Example 6.18.9. Let $X_n \sim N(0,n)$. We have $X_n \overset{\text{d}}{=} \sqrt{n}X_1$, and $\varphi_{X_n}(t) = \varphi_{X_1}(\sqrt{n}t)$ by part 5 of Proposition 6.18.3. In particular, we have

$$\varphi_{X_n}(t) = e^{-(\sqrt{n}t)^2/2} \to \begin{cases} 0 & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases}$$

Hence the limit of the $\varphi_{X_n}(t)$ is not continuous at 0.

Roughly, the fact that φ is not continuous at 0 is telling us that the mass is escaping to infinity. Formally: △

Let us continue the proof of Theorem 6.18.8. Our claim is that if φ is continuous at $t = 0$, then $\{X_n\}$ is tight in the sense of Definition 6.17.5. If this were the case, Corollary 6.17.8 says that there exists n_k so that $X_{n_k} \xrightarrow{\mathrm{d}} X$. But the subsequential limit X is unique since if $X_{n_k} \to X$ and $X_{m_k} \to Y$, then $\varphi_{n_k} \to \varphi$ and $\varphi_{m_k} \to \varphi$, so $\varphi_X = \varphi = \varphi_Y$, hence $X \overset{\mathrm{d}}{=} Y$.

It is left to show the following lemma:

Lemma 6.18.10. *Suppose for all subsequences n_k, there exists a subsubsequence n_{k_j} satisfying $X_{n_{k_j}} \to X$. Then $X_n \xrightarrow{\mathrm{d}} X$.*

One way to prove this lemma is to show that convergence in distribution comes from a metric; convergence in metric spaces all satisfy Lemma 6.18.10. Another way of proving it is as follows:

Proof of Lemma 6.18.10. We have $\mathbb{E}g(X_{n_{k_j}}) \to \mathbb{E}g(X)$ for all bounded continuous g. Thus $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ for all bounded continuous g, because convergence in real numbers indeed comes from a metric [i.e., we may apply the fact that for real numbers $\{a_n\}$, if for every subsequence a_{n_k} there exists a subsubsequence $a_{n_{k_j}} \to a$, then $a_n \to a$, since convergence in $\mathbb{R}$ comes from a metric]. Then Theorem 6.17.1 implies $X_n \xrightarrow{\mathrm{d}} X$. □

Proof 2 of Lemma 6.18.10. Use the fact that $X_n \xrightarrow{\mathrm{d}} X$ if and only if $W(X_n, X) \to 0$, where where W denotes the Wasserstein metric. □

This completes the proof of Theorem 6.18.8. □

6.19 Nov 6, 2019

We will prove our first central limit theorem today.
Our first task is to understand the moments $\mathbb{E}|X|^n$ in terms of $\varphi_X(t) = \mathbb{E}e^{itX}$.
Suppose $\mathbb{E}|X| < \infty$, and $\varphi'(0)$ exists, and furthermore that φ' is continuous at zero. Recall that

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx}\,d\mu(x) \qquad \text{where } \mu = \mu_X,$$

hence

$$\frac{d}{dt} = \int_{\mathbb{R}} \frac{d}{dt}[e^{itx}]\,d\mu(x) = \int (ix)e^{itx}\,d\mu(x).$$

Hence

$$\varphi_X'(0) = \int_{\mathbb{R}} (ix)\,d\mu(x) = i\mathbb{E}X.$$

Likewise if $\mathbb{E}X^2 < \infty$, and φ is nice enough, we have

$$\varphi_X''(t) = \int_{\mathbb{R}} \frac{d^2}{dt^2}[e^{itx}]\,d\mu(x) = \int_{\mathbb{R}} (ix)^2 e^{itx}\,d\mu(x),$$

hence as above

$$\varphi_X''(0) = -\mathbb{E}X^2.$$

In general, if $\mathbb{E}|X|^n < \infty$ then $\varphi_X^{(n)}$ is continuous at $t = 0$ and $\varphi_X^{(n)}(0) = \mathbb{E}(iX)^n$.

Dictionary 6.19.1. *Strengthening the analogy with Fourier theory,*

$$\begin{aligned}
\text{Decay: } \int |x|^n\,d\mu(x) < \infty &\implies \text{Smoothness: } \varphi^{(n)}(0) \text{ exists and is continuous} \\
\text{Smoothness: there is } f \text{ so that } \mu((a,b)) = \int_a^b f(x)\,dx &\impliedby \text{Decay: } \int |\varphi(t)| < \infty \\
\text{Decay: } \{\mu_n\} \text{ tight} &\impliedby \text{Smoothness: } \varphi_n \to \varphi, \varphi \text{ continuous at } 0 \\
\text{Decay: } \int x^2\,d\mu(x) < \infty &\impliedby \text{Smoothness: } \varphi''(0) > -\infty. \qquad (9)
\end{aligned}$$

Equation (9) will be particularly useful for us, so we prove it here.

Proof of Equation (9). We supposed that

$$\varphi_X''(0) = \lim_{h\to 0} \frac{\varphi(h) - 2\varphi(0) + \varphi(-h)}{h^2} > -\infty.$$

Note that

$$\frac{e^{ihx} - 2 + e^{-ihx}}{h^2} = \frac{-2(1-\cos hx)}{h^2} \to \frac{-2(hx)^2/2}{h^2} = -x^2$$

as $h \to \infty$. (Here we used a Taylor series for cos.)

Now Fatou's Lemma (Lemma 3.7.5) says [Edit, Dec 7: this seems to say that the second moment is sometimes bounded by a negative number.]

$$\int x^2\,d\mu(x) \le \liminf_{h\to 0} \int \left(\frac{2(1-\cos hx)}{h^2}\right) d\mu(x) = -\limsup_{h\to 0} \frac{\varphi(h) - 2\varphi(0) + \varphi(-h)}{h^2} < \infty.$$

□

Sort of a converse to above is

Lemma 6.19.2. *If* $\mathbb{E}|X|^2 < \infty$, *then*

$$\varphi_X(t) = 1 + it(\mathbb{E}X) - t^2\frac{\mathbb{E}X^2}{2} + o(t^2),$$

where $o(t^2)/t^2 \to 0$ *as* $t \to 0$.

Proof. We Taylor expand around $t = 0$:

$$\left|e^{i\theta} - \left(1 + i\theta + \frac{(i\theta)^2}{2}\right)\right| \leq \min\left(\frac{|\theta|}{6}, |\theta|^2\right),$$

note that we should take care not to assume that X has a third moment. Setting $\theta = tX$, we see

$$\mathbb{E}\left|e^{itX} - \left(1 - itX + \frac{(itX)^2}{2}\right)\right| \leq \mathbb{E}\min\left(\frac{|tX|^3}{6}, |tX|^2\right)$$

Let us denote by M_t the minimum $\min(\frac{|tX|^3}{6}, |tX|^2)$. We obtain

$$\frac{|\varphi_X(t) - 1 + it\mathbb{E}X - t^2(\mathbb{E}X^2)|}{t^2} \leq \mathbb{E}\left(\frac{M_t}{t^2}\right);$$

we'd like to show that $\mathbb{E}(M_t/t^2) \to 0$ as $t \to 0$, as the lemma asks for. We have

$$\frac{M_t}{t^2} = \min\left(\frac{|tX^3|}{6}, X^2\right) \leq \frac{|tX^3|}{6} \to 0 \text{ (pointwise, as RVs)}, \qquad \text{as } t \to 0.$$

Since M_t/t^2 is dominated by X^2, and $\mathbb{E}X^2 < \infty$, we may apply the Dominated Convergence Theorem (Theorem 3.8.2) to obtain

$$\mathbb{E}\left(\frac{M_t}{t^2}\right) \to 0. \qquad \square$$

Theorem 6.19.3 (Central Limit Theorem)**.** *Let* $X_1, X_2, \ldots$ *be i.i.d. random variables with* $\mathbb{E}X_1 = \mu$, *and* $\mathrm{Var}(X_1) = \sigma^2$. *Let* $S_n = X_1 + \cdots + X_n$. *Then*

$$\frac{S_n - \mu n}{\sigma\sqrt{n}} \xrightarrow{\mathrm{d}} Z \sim N(0,1).$$

Proof. We may assume $\mu = 0$ since we can apply the $\mu = 0$ case to $X_n' \stackrel{\text{def}}{=} X_n - \mu$. Recall that part 6 of Proposition 6.18.3 says that characteristics multiply, and that convergence of characteristic functions are related to convergence in distribution (Theorem 6.18.8); hence characteristic functions are built to useful for us in this kind of situation.

Specifically, observe that

$$\varphi_{X_1}(t) = 1 + \underbrace{it(\mathbb{E}X_1)}_{=0} - t^2\frac{\mathbb{E}X_1^2}{2} + o(t^2) = 1 - \frac{\sigma^2}{2}t^2 + o(t^2),$$

hence by i.i.d.-ness

$$\varphi_{S_n}(t) = \varphi_{X_1}(t)\ldots\varphi_{X_n}(t) = \varphi_{X_1}(t)^n.$$

In part 5 of Proposition 6.18.3 we showed that $\varphi_{aX}(t) = \varphi_X(at)$, so

$$\begin{aligned}
\varphi_{S_n/(\sigma\sqrt{n})}(t) &= \varphi_{S_n}\left(\frac{t}{\sigma\sqrt{n}}\right) \\
&= \left(1 - \frac{\sigma^2}{2}\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + o\left(\frac{t}{\sigma\sqrt{n}}\right)^2\right)^n \\
&= \left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \\
&= (1 - c_n)^n, \text{ where } nc_n \to t^2/2 \text{ as } n \to \infty
\end{aligned}$$

where in the last equality we used the fact that for fixed t and $n \to \infty$, the quantity $t/(\sigma\sqrt{n}) \to 0$, hence the error goes to zero as well. As $n \to \infty$, we obtain

$$\varphi_{S_n/(\sigma\sqrt{n})}(t) \to e^{-it^2/2} = \varphi_Z(t)$$

as $n \to \infty$. □

Suppose we had a triangular array of random variables $\{X_{ij}\}_{1 \le j \le i \le n}$, perhaps arranged as follows:

$$\begin{array}{ccccc} X_{11} & & & & \\ X_{21} & X_{22} & & & \\ X_{31} & X_{32} & X_{33} & & \\ \vdots & \vdots & \vdots & \ddots & \\ X_{n1} & \dots & \dots & \dots & X_{nn} \end{array}$$

Let $S_n = X_{n1} + \cdots + X_{nn}$. Assume $\{X_{nm}\}_{m=1}^n$ are independent (but not necessarily i.i.d.!), and that $\mathbb{E}(X_{nm}) = 0$. We have

Theorem 6.19.4 (Lindeberg-Feller Central Limit Theorem)**.** *Suppose*

$$\sum_{m=1}^n \mathbb{E}X_{n,m}^2 \to \sigma^2 \quad \text{as } n \to \infty, \tag{10}$$

and

$$\sum_{m=1}^n \mathbb{E}(X_{nm}^2 \mathbb{1}_{\{|X_{nm}|>\varepsilon\}}) \to 0 \quad \text{as } n \to \infty \text{ for all } \varepsilon > 0. \tag{11}$$

Then $S_n \xrightarrow{\mathrm{d}} \sigma Z \sim N(0, \sigma^2)$ *as* $n \to \infty$.

If $Y_1, Y_2, \dots$ are i.i.d. with $\mathbb{E}Y_1 = 0$ and $\mathbb{E}Y_1^2 = \sigma^2$, then we can take $X_{n,m} = Y_m/\sqrt{n}$; this will satisfy the conditions of the theorem. (Check that this satisfies condition (11)!)

Proof of Theorem 6.19.4. Let $\varphi_{n,m}(t) = \mathbb{E}e^{itX_{n,m}}$. It is enough to show that

$$\varphi_{S_n}(t) = \prod_{m=1}^n \varphi_{n,m}(t) \to e^{-t^2\sigma^2/2} = \varphi_{\sigma Z}(t)$$

for all t. To show this, we compute

$$\begin{aligned} \left|\varphi_{nm}(t) - \left(1 - \frac{t^2\sigma_{n,m}^2}{2}\right)\right| &\le \mathbb{E}\min(|tX_{nm}|^3, |tX_{nm}|^2) \\ &\le \mathbb{E}\min(|tX_{nm}|^3 \mathbb{1}_{\{|X_{nm}|\le\varepsilon\}}, |tX_{nm}|^2 \mathbb{1}_{\{|X_{nm}|>\varepsilon\}}) \\ &\le \mathbb{E}(|tX_{nm}|^3 \mathbb{1}_{\{|X_{nm}|\le\varepsilon\}} + |tX_{nm}|^2 \mathbb{1}_{\{|X_{nm}|>\varepsilon\}}) \\ &\le \varepsilon t \mathbb{E}(tX_{nm})^2 + t^2 \mathbb{E}(X_{nm}^2 \mathbb{1}_{\{|X_{nm}|>\varepsilon\}}). \end{aligned}$$

We can sum

$$\sum_{m=1}^n \left|\varphi_{nm}(t) - \left(1 - \frac{t^2\sigma_{nm}^2}{2}\right)\right| \le \underbrace{\varepsilon t^3\sigma^2}_{\text{by (10)}} + t^2 \cdot \underbrace{(\text{something going to zero})}_{\text{by (11)}}.$$

(Proof to be continued.) □

6.20 Nov 11, 2019

[Prof Levine will be gone later this week. There is no class on Wednesday and no OH on Thursday. There will be an extra class on Friday, Nov 22, at 8:40. Room details coming soon. Logistical homework details also coming up soon.]

We were in the middle of the proof of the Lindeberg-Feller CLT (Theorem 6.19.4), rephrased below for convenience:

Theorem 6.20.1 (Lindeberg-Feller Central Limit Theorem; cf. Theorem 6.19.4)**.** *Let $(X_{nm})_{1\le m\le n}$ be a (triangular) array of random variables. Suppose $\mathbb{E}X_{nm}=0$, and that the rows of this array, $\{X_{nm}\}_{m=1}^n$ are independent. Suppose further that*

$$\sum_{m=1}^n \mathbb{E}X_{nm}^2 \to \sigma^2 \quad \text{as } n\to\infty, \tag{12}$$

and

$$\sum_{m=1}^n \mathbb{E}(X_{nm}^2 \mathbb{1}_{\{|X_{nm}|>\varepsilon\}}) \to 0 \quad \text{as } n\to\infty \text{ for all } \varepsilon>0. \tag{13}$$

Let $S_n = X_{n1}+\cdots+X_{nn}$. Then $S_n \xrightarrow{\mathrm{d}} N(0,\sigma^2)$.

Proof. We were considering the characteristic functions $\varphi_{nm}(t)=\mathbb{E}(e^{iX_{nm}t})$ of X_{nm}. Taking Taylor expansions, we saw that

$$\varphi_{nm}(t) = 1-\frac{t^2\sigma_{nm}^2}{2}+o(t^2) \quad \text{as } t\to 0,$$

where $\sigma_{nm}^2=\mathbb{E}(X_{nm}^2)$. We saw that the hypotheses (12) and (13) implied that

$$\sum_{m=1}^n \left|\varphi_{nm}(t)-\left(1-\frac{t^2\sigma_{nm}^2}{2}\right)\right| \to 0 \quad \text{as } n\to\infty.$$

Aside 6.20.2. If $z_i, w_i\in\mathbb{C}$ with $|z_i|,|w_i|\le 1$, then

$$\left|\prod_{i=1}^n z_i - \prod_{i=1}^n w_i\right| \le \sum_{i=1}^n |z_i-w_i|.$$

One can prove this by induction on n and using triangle inequality, since

$$\left|\prod_{i=1}^n z_i - \prod_{i=1}^n w_i\right| = \left|\prod_{i=1}^n z_i - z_1w_2\dots w_n + z_1w_2\dots w_n - \prod_{i=1}^n w_i\right| \le \underbrace{|z_1|}_{\le 1}\cdot\left|\prod_{i=2}^n z_i - \prod_{i=2}^n w_i\right| + |z_1-w_1|\cdot\underbrace{\left|\prod_{i=2}^n w_i\right|}_{\le 1}$$

so induction gives the desired conclusion. △

Let us take $z_m=\varphi_{nm}(t)$ and $w_m = 1-\frac{t^2\sigma_{nm}^2}{2}$ and apply Aside 6.20.2. Note that $|z_m|\le 1$. To see that $|w_m|\le 1$, note that

$$\sigma_{nm}^2 = \mathbb{E}(X_{nm}^2) = \underbrace{\mathbb{E}(X_{nm}^2\mathbb{1}_{\{|X_{nm}|>\varepsilon\}})}_{\to 0} + \underbrace{\mathbb{E}(X_{nm}^2\mathbb{1}_{\{|X_{nm}|\le\varepsilon\}})}_{\le\varepsilon^2},$$

hence

$$\max_{1\le m\le n}(\sigma_{nm}^2)\to 0 \quad \text{as } n\to\infty.$$

So for sufficiently large n, we have $|w_m|\le 1$. By Aside 6.20.2, we obtain

$$\left|\prod_{m=1}^n \varphi_{mn}(t) - \prod_{m=1}^n\left(1-\frac{\sigma_{nm}^2t^2}{2}\right)\right| \le \sum_{m=1}^n\left|\varphi_{nm}(t)-\left(1-\frac{\sigma_{nm}^2t^2}{2}\right)\right| \to 0.$$

We now apply

Lemma 6.20.3. *If $c_{nm} \in \mathbb{R}$, with $\max_{1\le m\le n}(c_{nm}) \to 0$ and $\sum_{m=1}^n c_{nm} \to \lambda$ as $n \to \infty$, then*

$$\prod_{m=1}^n (1 - c_{nm}) \to e^{-\lambda}.$$

(We've seen variations of this lemma before, cf. the computation in Example 6.16.6.) We apply Lemma 6.20.3 with $c_{nm} = \frac{\sigma_{nm}^2 t^2}{2}$; observe that the conditions required to apply Lemma 6.20.3 hold. We obtain

$$\prod_{m=1}^n \left(1 - \frac{t^2\sigma_{nm}^2}{2}\right) \to e^{-t^2\sigma^2/2} = \varphi_Z(t),$$

for $Z \sim N(0, \sigma^2)$. □

One can think of Central Limit Theorems as second-order corrections to laws of large numbers. (This is made precise in the example below.)

Example 6.20.4 (Records in i.i.d. sequences; cf. Example 5.14.5)**.** Consider i.i.d. random variables $\{U_n\}_{n\ge1}$, with $U_i \sim \text{Unif}(0, 1)$. Let

$$Y_n = \mathbb{1}_{\{U_n > \max\{U_1,\dots,U_{n-1}\}\}} = \begin{cases} 1 & \text{with probability } 1/n \\ 0 & \text{with probability } 1 - 1/n \end{cases}.$$

Let $S_n = Y_1 + \cdots + Y_n$ count the number of records up to time n. We saw in Example 5.14.5 that the strong law of large numbers says

$$\frac{S_n}{\log n} \to 1 \quad \text{a.s.},$$

or in other words that $S_n = \log n + R_n$ with $R_n/\log n \to 0$ almost surely. Central limit theorems are second-order corrections to laws of large numbers in the sense that they tell us about the R_n term.

We have

$$\text{var}(S_n) = \sum_{i=1}^n \text{var}(Y_i) = \sum_{m=1}^n (\mathbb{E}Y_m^2 - (\mathbb{E}Y_m)^2) = \sum_{m=1}^n \left(\frac{1}{m} - \left(\frac{1}{m}\right)^2\right) \sim \log n.$$

Let us define, for $1 \le m \le n$, the random variables

$$X_{nm} = \frac{Y_m - \frac{1}{m}}{\sqrt{\log n}}.$$

By construction we have $\mathbb{E}X_{nm} = 0$ and

$$\sum_{m=1}^n \mathbb{E}(X_{nm}^2) = \sum_{m=1}^n \frac{\text{var}(Y_m)}{\log n} \to 1;$$

furthermore, for n sufficiently large, the set $\{|X_{nm}| > \varepsilon\}$ will have measure zero, and

$$\sum_{m=1}^n \mathbb{E}(X_{nm}^2 \mathbb{1}_{\{|X_{nm}|>\varepsilon\}}) = 0$$

for all sufficiently large n. We have verified the conditions to apply Theorem 6.20.1, which says that

$$\sum_{m=1}^n X_{nm} \xrightarrow{\text{d}} N(0, 1).$$

On the other hand, we see that

$$\sum_{m=1}^n X_{nm} = \frac{\sum_{m=1}^n Y_m - \sum_{m=1}^n \frac{1}{m}}{\sqrt{\log n}} \xrightarrow{\text{d}} N(0, 1).$$

We see that the number of records is

$$S_n = \underbrace{\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right)}_{\approx \log n} + R_n,$$

where R_n is asymptotically normal with standard deviation $\sqrt{\log n}$.

This is a pretty typical situation: there is a main term which is deterministic, and there is an error term which is a random variable on the order of the square root of the main term. △

Example 6.20.5 (cf. (HW 6, Ex 2))**.** Let $U_n \sim \text{Unif}(0,1)$ be i.i.d. random variables. Let $X_n = U_1 \dots U_n$ be the product of the U_i, and let

$$S_n = \log X_n = \sum_{m=1}^{n} \log U_m.$$

Set $L_m \stackrel{\text{def}}{=} \log U_m$. We see that

$$\mathbb{E}L_1 = \int_0^b \frac{1}{b} \log x \, dx = \frac{1}{b}[x(\log x - 1)]_0^b = \log b - 1.$$

The strong law of large numbers says that

$$\frac{S_n}{n} \to \log b - 1 \quad \text{a.s..}$$

In particular if $b < e = 2.712...$ then $\log b - 1 < 0$, so

$$S_n \to -\infty \quad \text{a.s..}$$

In particular, $X_n = e^{S_n} \to 0$ almost surely, even though

$$\mathbb{E}(X_n) = (\mathbb{E}U_1) \dots (\mathbb{E}U_n) = \left(\frac{b}{2}\right)^n \to \infty \quad \text{whenever } b > 2.$$

The simultaneous almost sure convergence $X_n \to 0$ and convergence of numbers $\mathbb{E}(X_n) \to \infty$ when $2 < b < e$ is explained by the fact that X_n very rarely takes astronomically large values.

With the central limit theorem, we can get more refined information. Note that

$$\mathbb{E}L_1^2 = \int_0^b \frac{1}{b} (\log x)^2 \, dx = (\log b - 1)^2 + 1,$$

hence $\text{var}(L_1) = 1$. The central limit theorem then says

$$\frac{\log X_n - n(\log b - 1)}{\sqrt{n}} \stackrel{\text{d}}{\to} N(0,1).$$

In the case $b = 2.5$ and $n = 1000$, we are in the setting of Exercise 2(c). We see that $\log b - 1 \approx -0.084$, hence that $n(\log b - 1) \approx -84$.

$$\mathbb{E}(X_{1000}) = \left(\frac{b}{2}\right)^{1000} = (1.25)^{1000},$$

which is huge. But

$$P(X_{1000} > 1000) = P(\log X_{1000} > \log 1000) \approx P\Bigg(\underbrace{\frac{\log X_{1000} - (-84)}{\sqrt{1000}}}_{\approx N(0,1)} > \underbrace{\frac{7 - (-84)}{\sqrt{1000}}}_{\approx 2.88}\Bigg) \approx P(Z > 2.88),$$

where $Z \sim N(0,1)$. This is extremely unlikely!

We will hopefully get to talk about the Berry-Esseen theorem, which makes precise how quickly the CLT random variables converge to the normal distribution. △

In the near future we'll talk about multivariate normal distributions and CLT for random vectors.

6.21 Nov 18, 2019

We'll talk about the multivariate normal distribution today.

Definition 6.21.1. The random variables $\mathbf{X} = (X_1, \dots, X_m)$ are multivariate normal (sometimes called MVN, jointly normal, Gaussian) if for every $\mathbf{t} = (t_1, \dots, t_m) \in \mathbb{R}^m$, the distribution

$$\mathbf{t} \cdot \mathbf{X} = t_1 X_1 + \cdots + t_m X_m$$

has a normal distribution $N(\mu_{\mathbf{t}}, \sigma_{\mathbf{t}}^2)$. △

Note that $\mu_{\mathbf{t}} = \mathbf{t} \cdot \mathbb{E}\mathbf{X}$. We also allow $\sigma_{\mathbf{t}}^2 = 0$.

Example 6.21.2. Let Y, Z be independent $N(0,1)$ random variables. Then $X = (2Y+Z, 5Z-3Y)$ is Gaussian. △

Example 6.21.3. Let $X_2 = X_1 \xi$ with $X_1 \sim N(0,1)$ and

$$\xi = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}$$

Then $X_2 \sim N(0,1)$ as well, and yet (X_1, X_2) is not Gaussian, since

$$\mathbb{P}(X_1 = X_2) = \mathbb{P}(\xi = 1) = \frac{1}{2},$$

hence $\mathbb{P}(X_1 - X_2 = 0) = \frac{1}{2}$. In particular, $X_1 - X_2$ does not have a normal distribution. △

Example 6.21.4. Consider (X_1, X_1), where $X_1 \sim N(0,1)$. We want to call this Gaussian, but $\sigma_{(1,-1)}^2 = 0$. (This is why we allow $\sigma_{\mathbf{t}}^2 = 0$.) △

If $(X_1, \dots, X_m)$ is Gaussian, then so is $(X_1 - \mathbb{E}X_1, \dots, X_m - \mathbb{E}X_m)$. Thus we may assume $\mathbb{E}X_i = 0$ for all i. Now we may write

$$\mathbb{E}((\mathbf{t} \cdot \mathbf{X})(\mathbf{u} \cdot \mathbf{X})) = \mathbb{E}\left(\sum_{i=1}^m (t_i X_i) \sum_{j=1}^m (u_j X_j) \right) = \sum_{i,j=1}^m t_i u_j \mathbb{E}(X_i X_j). \tag{14}$$

This quantity should feel like a quadratic form. Indeed, we can write

$$\sum_{i,j=1}^m t_i u_j \mathbb{E}(X_i X_j) = \begin{bmatrix} t_1 & \dots & t_m \end{bmatrix} \Gamma \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}, \quad \text{where} \quad \begin{cases} \Gamma = (\Gamma_{ij})_{i,j \in [m]} \\ \Gamma_{ij} = \mathbb{E}(X_i X_j) \end{cases}$$

here Γ is called the covariance matrix of X.

Definition 6.21.5. Consider a random vector $\mathbf{X} = (X_1, \dots, X_m) \colon \Omega \to \mathbb{R}^m$. Its characteristic function is defined to be

$$\varphi_{\mathbf{X}}(t_1, \dots, t_m) = \mathbb{E} e^{i(\mathbf{t} \cdot \mathbf{X})}.$$ △

Fact 6.21.6 (Cramér-Wold Theorem)**.** *There is an inversion formula, which implies that if $\varphi_{\mathbf{X}}(\mathbf{t}) = \varphi_{\mathbf{Y}}(\mathbf{t})$ for all $\mathbf{t} \in \mathbb{R}^m$, then $\mathbf{X} \stackrel{\mathrm{d}}{=} \mathbf{Y}$. In particular, if $\mathbf{t} \cdot \mathbf{X} \stackrel{\mathrm{d}}{=} \mathbf{t} \cdot \mathbf{Y}$ for all $\mathbf{t} \in \mathbb{R}^m$, then $\mathbf{X} \stackrel{\mathrm{d}}{=} \mathbf{Y}$.*

(Note that the hypothesis consists of infinitely many equalities of random variables, whereas the conclusion is an equality of joint random variables!)

Lemma 6.21.7. *If $\mathbf{X}$ and $\mathbf{Y}$ are Gaussian, with $\mathbb{E}X_i = \mathbb{E}Y_i$ and $\mathbb{E}(X_i X_j) = \mathbb{E}(Y_i Y_j)$ for all i, j then $\mathbf{X} \stackrel{\mathrm{d}}{=} \mathbf{Y}$.*

Proof. We saw in Equation (14) that $\mathbf{t} \cdot \mathbf{X} \sim N(0, \mathbf{t}^T \Gamma \mathbf{t})$, where Γ is the covariance matrix. Similarly, $\mathbf{t} \cdot \mathbf{Y} \sim N(0, \mathbf{t}^T \Gamma \mathbf{t})$. It follows that $\mathbf{X} \stackrel{\mathrm{d}}{=} \mathbf{Y}$, by Fact 6.21.6. □

Which matrices Γ can arise as covariance matrices? Well, observe that

$$0 \leq \mathbb{E}((\mathbf{t} \cdot \mathbf{X})^2) = \mathbf{t}^T \Gamma \mathbf{t},$$

so Γ is positive-semidefinite. In particular it is symmetric and has nonnegative eigenvalues; we may diagonalize and write

$$\Gamma = U^T V U,$$

with U an orthogonal $m \times m$ matrix and V is a diagonal matrix with nonnegative entries. The nonnegativity of the entries of

$$V = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & \lambda_m \end{bmatrix}$$

allows us to decompose further $V = W^T W$, where $W = W^T$ is a diagonal matrix with $\sqrt{\lambda_i}$ on the ith row and column. Thus can write

$$\Gamma = U^T W^T W U = A^T A.$$

[Everything here is equivalent, i.e. for any matrix A, the matrix $A^T A$ is positive-semidefinite.] This proves the first half of the following result:

Proposition 6.21.8. *The covariance matrix $\Gamma = A^T A$ is positive-semidefinite. Conversely, if $\Gamma = A^T A$ for some A, there exist $(Y_1, \dots, Y_m) \sim N(0, \Gamma)$.*

Proof. Let $\mathbf{X} = (X_1, \dots, X_m)$ be a vector of independent $N(0,1)$ random variables. Let $\mathbf{Y} = \mathbf{X}A$. We can check that

$$\mathbf{t} \cdot \mathbf{Y} = (\mathbf{t} \cdot \mathbf{X}A) = \sum_i t_i \sum_j X_j A_{ji} = \sum_j X_j \sum_{i=1}^m (t_i A_{ji}) \sim N(0, \sigma^2),$$

where $\sigma^2 = \sum_j \left(\sum_{i=1}^m (t_i A_{ji}) \right)^2 = \mathbf{t}^T A^T A \mathbf{t} = \mathbf{t}^T \Gamma \mathbf{t}$. □

We define weak convergence of random vectors as follows. Let $\mathbf{X}_n = (X_n^1, \dots, X_n^m) \colon \Omega \to \mathbb{R}^m$. Define

$$F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(X^1 \leq x^1, \dots, X^m \leq x^m).$$

We have

Theorem 6.21.9. *The following are equivalent:*

1. *$F_{\mathbf{X}_n}(\mathbf{x}) \to F_{\mathbf{X}}(\mathbf{x})$ at all continuity points of $F_{\mathbf{X}}$.*
2. *$\mathbb{E}g(\mathbf{X}_n) \to \mathbb{E}g(\mathbf{X})$ for all bounded continuous $g \colon \mathbb{R}^m \to \mathbb{R}$, and*
3. *$\varphi_{\mathbf{X}_n}(\mathbf{t}) \to \varphi_{\mathbf{X}}(\mathbf{t})$ for all $\mathbf{t} \in \mathbb{R}^m$.*

From this theorem we can obtain a central limit theorem for i.i.d. random vectors.

Theorem 6.21.10 (CLT for i.i.d. random vectors)**.** *Let $\mathbf{X}_1, \mathbf{X}_2, \dots \colon \Omega \to \mathbb{R}^m$, and let $\mathbb{E}\mathbf{X}_1 = \mu \in \mathbb{R}^m$. Assume $\mathbb{E}(X_1^i)^2 < \infty$ for all $i = 1, \dots, m$. Define $\Gamma_{ij} = \mathbb{E}((X_1^i - \mu^i)(X_1^j - \mu^j))$ and $\mathbf{S}_n = \mathbf{X}_1 + \dots + \mathbf{X}_n$. Then*

$$\frac{\mathbf{S}_n - n\mu}{\sqrt{n}} \xrightarrow{\mathrm{d}} N(0, \Gamma).$$

Proof. By replacing $\mathbf{X}_n$ with $\mathbf{X}_n - \mu$ if necessary, we may assume $\mu = \mathbf{0}$. For $\chi \sim N(0, \Gamma)$, we write $\varphi_\chi(\mathbf{t}) = \exp(-\frac{\mathbf{t}^T \Gamma \mathbf{t}}{2})$. It suffices to prove $\mathbf{t} \cdot \mathbf{X}_n \xrightarrow{\mathrm{d}} \mathbf{t} \cdot \chi$ for all $\mathbf{t} \in \mathbb{R}^m$, by Fact 6.21.6. Indeed, let us consider $Y_n = \mathbf{t} \cdot \mathbf{X}_n$. Then $Y_1, Y_2, \dots$ are i.i.d. mean 0 and variance $\mathbf{t}^T \Gamma \mathbf{t}$. The usual 1-dimensional Central Limit Theorem (Theorem 6.19.3) says that

$$\frac{Y_1 + \dots + Y_n}{\sqrt{n}} \xrightarrow{\mathrm{d}} N(0, \mathbf{t}^T \Gamma \mathbf{t}) \sim \mathbf{t} \cdot \chi.$$

□

Example 6.21.11. Let us consider a simple random walk on $\mathbb{Z}^2$ given by

$$S_n = (X_n, Y_n) = \sum_{i=1}^{n} \xi_i,$$

where ξ_i are independent random variables taking the values $(1,0), (-1,0), (0,1), (0,-1)$ each with probability $1/4$. We obtain

$$\mathbb{E}\xi_i = \frac{1}{4}(1,00) + \frac{1}{4}(-1,0) + \frac{1}{4}(0,1) + \frac{1}{4}(0,-1) = (0,0),$$

and

$$\Gamma_{22} = \Gamma_{11} = \mathbb{E}(\xi_i^1)^2 = \frac{1}{4}1^2 + \frac{1}{4}(-1)^2 + \frac{1}{2}(0)^2 = \frac{1}{2},$$

and furthermore

$$\Gamma_{12} = \mathbb{E}(\xi_i^1 \xi_i^2) = 0.$$

Hence

$$\Gamma = \frac{1}{2}I = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

The central limit theorem (Theorem 6.21.10) says that $\mathbf{S}_n/\sqrt{n} \xrightarrow{\mathrm{d}} N(0, \frac{1}{2}I)$. Note that the limit is rotationally invariant, even though our grid is not rotationally invariant. [(!)] Indeed, the density of $N(0, \frac{1}{2}I)$ is equal to

$$\frac{e^{-x^2} \cdot e^{-y^2}}{(\sqrt{2}\pi\sigma^2)^2} = \frac{e^{-(x^2+y^2)}}{2\pi\sigma^2},$$

where for us $\sigma = \frac{1}{2}$. This density depends only on the distance to the origin, so it is rotationally invariant. △

The normal distribution is an extremely pervasive limiting distribution. Roughly, the CLT says that if you add together a large number of small contributions and normalize accordingly, the limiting distribution is normal.

An important non-normal limit distribution is the Poisson distribution. Roughly, the Poisson distribution is obtained by adding together a small number of large contributions. We say $Z \sim \mathrm{Pois}(\lambda)$ if $\mathbb{P}(Z = k) = \frac{\lambda^k}{e^\lambda k!}$.

Theorem 6.21.12. *Given a triangular array $\{X_{nm}\}_{1 \le m \le n}$ with each row $\{X_{nm}\}_{m=1}^n$ independent, write*

$$X_{nm} = \begin{cases} 1 & \text{with probability } p_{nm} \\ 0 & \text{with probability } 1 - p_{nm} \end{cases}$$

and define $S_n = X_{n1} + \cdots + X_{nn}$. Suppose that

1. *$\mathbb{E}S_n = p_{n1} + \cdots + p_{nn} \to \lambda \in (0, \infty)$ as $n \to \infty$,*
2. *$\max\{p_{n1}, \ldots, p_{nn}\} \to 0$.*

Then $S_n \xrightarrow{\mathrm{d}} Z \sim \mathrm{Pois}(\lambda)$.

Example 6.21.13 (Balls in boxes)**.** Let us consider dropping n balls in b boxes, with $b = n/\lambda$. Let

$$X_{nm} = \mathbb{1}_{\{\text{ball } m \text{ lands in box } 1\}}.$$

Define

$$S_n = \sum_{m=1}^{n} X_{nm} = \#\{\text{balls in box } 1\}.$$

Then observe that

$$\mathbb{P}(S_n = k)\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k} \sim \mathrm{Bin}\left(n, \frac{\lambda}{n}\right).$$

Since $\mathrm{Bin}(n, \frac{\lambda}{n}) \xrightarrow{\mathrm{d}} \mathrm{Pois}(\lambda)$ as $n \to \infty$. △

6.22 Nov 20, 2019

Last time we were talking about the Poisson distribution. By definition, a random variable $Z \sim \text{Pois}(\lambda)$ satisfies $\mathbb{P}(Z = k) = \frac{\lambda^k}{k!e^\lambda}$ for $k \in \mathbb{N} = \{0, 1, 2, \dots\}$.

Its characteristic function is

$$\varphi_Z(t) = \mathbb{E}e^{itZ} = \sum_{k\geq 0} \mathbb{P}(Z = k)e^{itk} = \sum_{k\geq 0} \frac{\lambda^k}{k!e^\lambda} = e^{-\lambda} \sum_{k\geq 0} \frac{(\lambda e^{it})^k}{k!} = e^{-\lambda} e^{\lambda it} = e^{\lambda(e^{it}-1)}.$$

We had stated Theorem 6.21.12, restated here for convenience:

Theorem 6.22.1 (cf. Theorem 6.21.12)**.** *Given a triangular array $\{X_{nm}\}_{1\leq m\leq n}$ with each row $\{X_{nm}\}_{m=1}^n$ independent, write*

$$X_{nm} = \begin{cases} 1 & \text{with probability } p_{nm} \\ 0 & \text{with probability } 1 - p_{nm} \end{cases}$$

and define $S_n = X_{n1} + \cdots + X_{nn}$. Suppose that

$$\mathbb{E}S_n = p_{n1} + \cdots + p_{nn} \to \lambda \in (0, \infty) \text{ as } n \to \infty, \tag{15}$$

and that

$$\max\{p_{n1}, \dots, p_{nn}\} \to 0. \tag{16}$$

Then $S_n \xrightarrow{\text{d}} Z \sim \text{Pois}(\lambda)$.

This theorem formalizes the fact that the Poisson distribution is obtained by adding together a small number of large contributions. (We're adding a bunch of random variables, most of which are zero)

Proof of Theorem 6.22.1. Let us compute

$$\varphi_{S_n}(t) = \prod_{m=1}^n ((1 - p_{nm})e^0 + p_{nm}e^{it}) = \prod_{m=1}^n (1 + p_{nm}(e^{it} - 1)).$$

Note that if $|w| \leq 1$ then

$$|e^w - (1 + w)| < |w|^2, \tag{17}$$

since

$$\left|\frac{w^2}{2} + \frac{w^3}{6} + \frac{w^4}{24} + \dots\right| \leq \left|\frac{w^2}{2}\right| + \left|\frac{w^3}{6}\right| + \cdots \leq \left(\frac{1}{2} + \frac{1}{6} + \dots\right)|w|^2.$$

We can use Equation (17) to estimate the difference

$$|\varphi_{S_n}(t) - e^{(\mathbb{E}S_n)(e^{it}-1)}| = \left|\prod_{m=1}^n (1 + p_{nm}(e^{it} - 1)) - \prod_{m=1}^n e^{p_{nm}(e^{it}-1)}\right| \leq \sum_{m=1}^n |1 + p_{nm}(e^{it} - 1) - e^{p_{nm}(e^{it}-1)}|,$$

where the inequality is Aside 6.20.2 from the proof of the central limit theorem. Setting $w = p_{nm}(e^{it} - 1)$ in Equation (17) we obtain

$$\sum_{m=1}^n |1 + p_{nm}(e^{it} - 1) - e^{p_{nm}(e^{it}-1)}| \leq \sum_{m=1}^n |p_{nm}(e^{it} - 1)|^2 \leq 4 \sum_{m=1}^n p_{nm}^2 \leq 4(\max_m(p_{nm})) \sum_{m=1}^n p_{nm} \to 0$$

as $n \to \infty$ by condition (16). (To apply Equation (17) we need $|p_{nm}(e^{it} - 1)| \leq 1$, but if we take n sufficiently large that $\max_m p_{nm} \leq \frac{1}{2}$ then we're good.)

With the estimation

$$|\varphi_{S_n}(t) - e^{(\mathbb{E}S_n)(e^{it}-1)}| \to 0,$$

we can write

$$|\varphi_{S_n}(t) - e^{\lambda(e^{it}-1)}| \leq |\varphi_{S_n}(t) - e^{(\mathbb{E}S_n)(e^{it}-1)}| + \underbrace{|e^{(\mathbb{E}S_n)(e^{it}-1)} - e^{\lambda(e^{it}-1)}|}_{\to 0 \text{ by condition (15)}}.$$

□

Example 6.22.2. Let $U_1, \dots, U_n \sim \text{Unif}(-n, n)$ be independent random variables. Let $X_{nm} = \mathbb{1}_{U_m \in (a,b)}$ for fixed a, b. Observe that $p_{nm} = \frac{b-a}{2n} \to 0$. Let $S_n = X_{n1} + \cdots + X_{nn} = \#(\{U_1, \dots, U_n\} \cap (a, b))$. Theorem 6.22.1 says that $S_n \xrightarrow{\text{d}} \text{Pois}(p_{n1} + \cdots + p_{nn}) = \text{Pois}(\frac{b-a}{2})$. Let's give a conceptual explanation for this. We begin with

Lemma 6.22.3. *Let $N_1 \sim \text{Pois}(\lambda_1)$ and $N_2 \sim \text{Pois}(\lambda_2)$ be independent random variables. Then $N_1 + N_2 \sim \text{Pois}(\lambda_1 + \lambda_2)$.*

Proof. Note that

$$\varphi_{N_1+N_2} = \varphi_{N_1}(t)\varphi_{N_2}(t) = e^{\lambda_1(e^{it}-1)}e^{\lambda_2(e^{it}-1)} = e^{(\lambda_1+\lambda_2)(e^{it}-1)},$$

hence that $N_1 + N_2 \sim \text{Pois}(\lambda_1 + \lambda_2)$. △

This should not be surprising: if you want to count the number $\#(\{U_1, \dots, U_n\} \cap (a, b))$ and you had a partition $[n] = S_1 \sqcup S_2$, then you can count $\#(\{U_i\}_{i \in S_1} \cap (a, b))$ and $\#(\{U_i\}_{i \in S_2} \cap (a, b))$ and add them.

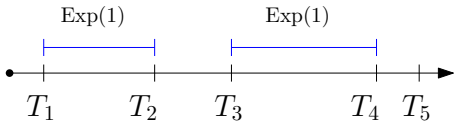
We'll show in the next homework more generally that if $N_1, N_2, \dots$ are independent and $N_i \sim \text{Pois}(\lambda_i)$ with $\sum_{i \geq 1} \lambda_i < \infty$, then $N_1 + N_2 + \cdots \sim \text{Pois}(\sum_{i \geq 1} \lambda_i)$.

Thus to conceptualize this example, one can try to make a statement that the Poisson distribution is the unique distribution supported on the integers having this additivity property. △

Theorem-Definition 6.22.4. *A Poisson point process on a σ-finite measure space $(S, \mathcal{F}, \mu)$ is a collection of random variables $(N(A))_{A \in \mathcal{F}}$ such that $N(A) \sim \text{Pois}(\mu(A))$ and if $A_1, \dots, A_k$ are disjoint then $N(A_1), \dots, N(A_k)$ are independent.*

It's not clear at all that these exist (that's the "theorem" part of this).

Here's how to construct a Poisson point process (P.P.P.) of intensity μ (Lebesgue measure) on $\mathbb{R}_{>0} = (0, \infty)$. Let $X_1, X_2, \dots$ be independent $\text{Exp}(1)$ random variables and let $T_n = X_1 + \cdots + X_n$. Think of T_i as living on $\mathbb{R}_{>0}$ where their differences are independent $\text{Exp}(1)$ random variables, as below:

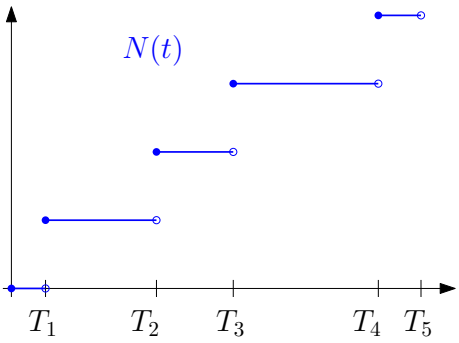


Let $N(A) = \#(A \cap \{T_1, T_2, \dots\})$. The claim is that these $N(A)$'s are a P.P.P. This is not so easy to prove, but the key to proving this is the memoryless property of the $\text{Exp}(1)$ distribution, that says $\mathbb{P}(X_i > s + t | X_i > s) = \mathbb{P}(X_i > t)$ for all $s, t > 0$.

There is a related notion of a Poisson process defined by

$$N(t) \stackrel{\text{def}}{=} N((0, t]) = \#\{n \colon T_n \leq t\}. \tag{18}$$

Thus the graph of $N(t)$ may look like



This process has the Markov property, which says that $(N(s+t) - N(s))_{t \geq 0} \stackrel{\text{d}}{=} (N(t))_{t \geq 0}$ and is independent of $(N(u))_{u < s}$. This also uses the fact that $\text{Exp}(1)$ is memoryless.

The density of $T_1 \sim \text{Exp}(1)$ is $f_1(t) = e^{-t}$ for $t \geq 0$. The density of $T_2 = X_1 + X_2$, with $X_i \sim \text{Exp}(1)$ independent, is

$$f_2(t) = \int_{-\infty}^{\infty} f_{X_1}(s) f_{X_2}(t-s)\, ds = \int_0^t e^{-s} e^{-(t-s)}\, ds = te^{-t}.$$

Similarly, the density of $T_n = X_1 + \cdots + X_n$, with each $X_i \sim \text{Exp}(1)$ independent, is $f_n(t) = \frac{t^{n-1}}{(n-1)!} e^{-t}$; this is called a Gamma$(n, 1)$ density. With the observation that $T_{n+1} = T_n + X_{n+1}$, we can compute

$$\begin{aligned}\mathbb{P}(N(t) = n) = \mathbb{P}(T_n \leq t < T_{n+1}) &= \int_0^t f_{T_n}(s) \mathbb{P}(X_{n+1} > t - s)\, ds \\ &= \int_0^t \frac{s^{n-1}}{(n-1)!} e^{-s} e^{-(t-s)}\, ds = \frac{e^{-t}}{(n-1)!} \int_0^t s^{n-1}\, ds = \frac{t^n}{n! e^t},\end{aligned}$$

hence $N(t) \sim \text{Pois}(t)$.

Note also that if $X_1, X_2, \cdots \sim \text{Exp}(\lambda)$, then $X_n = \frac{Y_n}{\lambda}$ with $Y_n \sim \text{Exp}(1)$. Hence $N(t) \sim \text{Pois}(\lambda t)$.

Let us briefly sketch how to construct a Poisson point process on a general σ-finite measure space $(S, \mathcal{F}, \mu)$. Take $S = \mathbb{R}^2$ with the Lebesgue measure, for example. Roughly, we're going to throw points randomly into $S = \mathbb{R}^2$ and for $A \subseteq S$ the random variable $N(A)$ will count the number of points there. More precisely, let

$$S = \bigsqcup_{n \geq 0} S_n, \qquad \mu(S_n) < \infty.$$

Pick $N_n \sim \text{Pois}(\mu(S_n))$ independent, and pick N_n points $X_{n,1}, \ldots, X_{n,N_n}$ uniformly at random from S_n. Then we set

$$N(A) = \#\left(A \cap \bigcup_n \{X_{n,1}, \ldots, X_{n,N_n}\} \right).$$

The claim is that this works. We'd need to prove that $N(A) \sim \text{Pois}(\mu(A))$ and that if $A_1, \ldots A_k$ are disjoint then $N(A_i)$ are independent. Ultimately, though, the proof of this result follows from Lemma 6.22.3 and the following lemma:

Lemma 6.22.5 (Thinning lemma)**.** *Take a Poisson point process of intensity μ, and cross out each point independently with probability p. The result is again a Poisson point process, but with intensity $(1-p)\mu$.*

There are some miracles happening, e.g. that this construction doesn't depend on the decomposition $S = \sqcup S_n$.

6.23 Nov 22, 2019

[We'll have one more homework, due Monday, Dec 9. We'll get takehome finals that Monday Dec 9, and we'll have 48 hours to do it (i.e. it will be due Wednesday, Dec 11).]

We've seen normal distributions as sums of i.i.d.s (Theorem 6.19.3), and Poisson distributions as sums of Bernoullis (Theorem 6.22.1).

Specifically, the Central Limit Theorem (Theorem 6.19.3) says that if X_i are i.i.d. random variables with *finite variance* and $S_n = X_1 + \cdots + X_n$, then

$$\frac{X_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathrm{d}} N(0,1).$$

We will be able to say something about the case $\mathbb{E}X_1^2 = \infty$. Thus our goal is to find hypotheses on X_i and sequences a_n, b_n so that

$$\frac{S_n - b_n}{a_n} \xrightarrow{\mathrm{d}} Y,$$

for some nontrivial limit Y (i.e., not ∞, and not an a.s. constant).

Definition 6.23.1. The random variable Y is called stable if there exists a sequence a_n, b_n such that for all n,

$$Y \stackrel{\mathrm{d}}{=} \frac{Y_1 + \cdots + Y_n - b_n}{a_n}$$

for all n, where $Y_1, Y_2, \ldots$ are independent and identically distributed to Y. △

Lemma 6.23.2. *The random variable Y arises as a limit of $\frac{X_1+\cdots+X_n-a_n}{b_n}$ for i.i.d. X_i if and only if Y is stable.*

If Y is stable we can pick X_i to be i.i.d. to Y. The backward direction follows from grouping $X_1+\cdots+X_n$ into

$$X_1 + \cdots + X_n = (X_1 + \cdots + X_k) + (X_{k+1} + \cdots + X_{2k}) + \ldots$$

and the sending n and k to infinity, but tactfully.

The simplest case of this is the symmetric α-stable case:

Definition 6.23.3. The stable random variable Y is called symmetric α-stable if $b_n = 0$ and $a_n = n^{1/\alpha}$. In other words,

$$\frac{Y_1 + \cdots + Y_n}{n^{1/\alpha}} \stackrel{\mathrm{d}}{=} Y_1.$$ △

For $\alpha = 2$, the normal distribution $N(0, \sigma^2)$ is a familiar example.

For $\alpha = 1$, we have

$$\frac{Y_1 + \cdots + Y_n}{n} \stackrel{\mathrm{d}}{=} Y_1.$$

On the characteristic function side, we see that any such random variable satisfies

$$\varphi(t/n)^n = \varphi(t). \tag{19}$$

Thus, characteristic function $\varphi(t) = e^{-c|t|}$ satisfies Equation (19); if it comes from a random variable Y then Y is symmetric 1-stable. Luckily, we have the inverse Fourier transform (cf. Theorem 6.18.5 – but not exactly): if φ is the characteristic function of a random variable with density f, then

$$f(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi(t)e^{-ity}\,dt = \frac{1}{2\pi}\left(\int_0^{\infty} e^{-ct-ity}\,dt + \int_{-\infty}^{0} e^{+ct-ity}\,dt\right). \tag{20}$$

We can take $c = 1$ now, since the general case will follow from scaling considerations. Thus we simplify Equation (20) to

$$\frac{1}{2\pi}\int_0^{\infty} e^{-t(1+iy)}\,dt + \frac{1}{2\pi}\int_0^{\infty} e^{-t(1-iy)}\,dt = \frac{1}{2\pi}\left(\frac{1}{1+iy} + \frac{1}{1-iy}\right) = \frac{1}{\pi}\frac{1}{1+y^2}.$$

Thus we've obtained the standard symmetric Cauchy density. Note that this is heavy-tailed (e.g., the random variable corresponding to this density doesn't have a well defined mean). Regardless, there is still a random variable having the Cauchy density as its distribution, and if Y is such a random variable then it satisfies $\varphi_Y = e^{-|t|}$, hence $\varphi_{cY} = e^{-c|t|}$.

A symmetric α-stable random variable Y having characteristic function

$$\varphi_Y(t) = e^{-c|t|^\alpha}$$

exists for any $0 < \alpha \leq 2$. Specifically, for a fixed $0 < \alpha < 2$, we have:

Theorem 6.23.4. *Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{P}(X_1 > x) = \mathbb{P}(X_1 < -x) = \frac{1}{2x^\alpha}$ for $x \geq 1$. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n^{1/\alpha}} \xrightarrow{\mathrm{d}} Y,$$

where Y is symmetric α-stable. Furthermore, $\varphi_Y(t) = e^{-C|t|^\alpha}$ where C is the constant so that

$$\varphi_{X_1}(t) = 1 - C|t|^\alpha + o(|t|^\alpha)$$

as $t \to 0$.

Proof idea. Only a few big X_i values matter (the X_i are heavy tailed, so there will be a few large values); we can drop the small X_i values without much loss. Here, "big" means x so that $\frac{1}{2x^\alpha} = \frac{1}{n}$, i.e. $x \sim n^{1/\alpha}$. We compute

$$\mathbb{P}(X_1 > an^{1/\alpha}) = \frac{1}{2}a^{-\alpha} \cdot \frac{1}{n}$$

and hence

$$\mathbb{P}\left(a < \frac{X_1}{n^{1/\alpha}} < b\right) = \frac{1}{2}(b^{-\alpha} - a^{-\alpha}) \cdot \frac{1}{n}.$$

We are in the situation of the balls in boxes example (Example 6.21.13). Specifically, the random variables $N_n(a,b)$ defined by

$$N_n(a,b) \stackrel{\mathrm{def}}{=} \#\left\{1 \leq m \leq n \colon a < \frac{X_m}{n^{1/\alpha}} < b\right\}$$

have a limit

$$N_n(a,b) \xrightarrow{\mathrm{d}} \mathrm{Pois}\left(\frac{1}{2}(b^{-\alpha} - a^{-\alpha})\right).$$

Fix an ε and throw away all X_i such that $|X_i| < \varepsilon \leq \varepsilon n^{1/\alpha}$. The rest are distributed like a Poisson Point Process with mean measure

$$\mu(a,b) = \frac{1}{2}(b^{-\alpha} - a^{-\alpha}).$$

So there are two things to check. Writing $S_n = \widetilde{S_n} + \overline{S_n}$, where $\widetilde{S_n}$ is the sum of the big X_i and $\overline{S_n}$ is the sum of the small X_i, we need to check that

$$\frac{\overline{S_n}}{n^{1/\alpha}} \xrightarrow{\mathrm{d}} 0 \quad \text{and that} \quad \frac{\widetilde{S_n}}{n^{1/\alpha}} \xrightarrow{\mathrm{d}} Y. \qquad \square$$

Definition 6.23.5. A function $L \colon \mathbb{R}_+ \to \mathbb{R}_+$ is slowly varying if for all $t > 0$ we have

$$\frac{L(tx)}{L(x)} \to 1 \quad \text{as } x \to \infty. \qquad \triangle$$

The classic example is $L(x) = \log x$. The following result, which we'll state without proof, holds:

Theorem 6.23.6. *If $X_1, X_2, \dots$ are i.i.d., satisfying*

$$\lim_{x\to\infty} \frac{\mathbb{P}(X_1 > x)}{\mathbb{P}(|X_1| > x)} = \theta$$

and

$$\mathbb{P}(|X_i| > x) = \frac{L(x)}{x^\alpha},$$

for a slowly varying L and some $0 < \alpha < 2$. Then

$$\frac{S_n - b_n}{a_n} \xrightarrow{\mathrm{d}} Y$$

where Y is a nontrivial stable random variable. Here,

$$a_n = \inf\left\{x \colon \mathbb{P}(|X_i| > x) \le \frac{1}{n}\right\} \quad \text{and} \quad b_n = n\mathbb{E}(X_1 \mathbb{1}_{\{|X_1| \le a_n\}}).$$

(Above, θ measures how asymmetric the X_i is.)

Definition 6.23.7. A random variable Z is called infinitely divisible if for all n, there exist i.i.d. $Y_1, \dots, Y_n$ so that $Z \overset{\mathrm{d}}{=} Y_1 + \cdots + Y_n$. △

Example 6.23.8. We have:

1. Any stable random variable is infinitely divisible.
2. If $Z \sim \text{Pois}(\lambda)$, then $Z = Y_1 + \cdots + Y_n$, where $Y_i \sim \text{Pois}(\lambda/n)$ are independent.
3. Let $\xi_1, \xi_2, \dots$ be any i.i.d. sequence of random variables, and let $Z(t) = \xi_1 + \cdots + \xi_{N(t)}$, where $(N(t))_{t\ge 0}$ is a Poisson process (see Equation (18) for a definition) with rate 1 that is independent of $(\xi_n)_{n=1}^\infty$. After writing
$$Z(t) = (\xi_1 + \cdots + \xi_{N(t/n)}) + (\xi_{N(t/n)+1} + \cdots + \xi_{N(2t/n)}) + \dots,$$
we see that $Z(t)$ is infinitely divisible. Furthermore,

$$\begin{aligned}\varphi_{Z(\lambda)}(t) = \mathbb{E}\exp(itZ(\lambda)) &= \sum_{n=0}^\infty \mathbb{P}(N(\lambda) = n)\mathbb{E}\exp(it(\xi_1 + \cdots + \xi_n)) \\ &= \sum_{n=0}^\infty \frac{\lambda^n}{e^\lambda n!}\varphi_{\xi_1}(t)^n = e^{-\lambda}\sum_{n=0}^\infty \frac{(\lambda\varphi(t))^n}{n!} = e^{-\lambda}e^{\lambda\varphi(t)} = e^{-\lambda(1-\varphi(t))}.\end{aligned}$$ △

Example 6.23.9 (Cauchy densities)**.** Cauchy densities show up:

1. Let's consider a random ray $\mathbb{R}^2$ starting at the origin with angle θ from the x-axis; pick $\theta \sim \text{Unif}(0, \pi)$. To such a ray consider the value of x so that $(x, 1)$ is on the ray. Then x has the standard symmetric Cauchy distribution.
2. If Y and Z are independent $N(0, 1)$, then Y/Z has the standard symmetric Cauchy density. △

7 Additional Topics

7.24 Nov 25, 2019 (Large Deviations)

[It's Thanksigiving soon! Prof Levine's office hours will be moved to Tuesday, 1-2pm at 438. HW 10 will probably be due two Mondays from now. We'll receive a poll soon for 48 hour slots during which we'll have the takehome final.]

Let's talk about large deviations. This is covered in Durrett chapter 2.7, because the topic doesn't require lots of prerequisites. But the results are contexualized by central limit theorems, so we'll talk about them now.

Let $X_1, X_2, \dots$ be i.i.d. with mean μ. Let $S_n = X_1 + \cdots + X_n$. The weak law of large numbers (Theorem 5.11.4) says that for a fixed $a \in \mathbb{R}$,

$$\pi_n \stackrel{\text{def}}{=} \mathbb{P}(S_n \geq na) \to \begin{cases} 0 & \text{if } a < \mu \\ 1 & \text{if } a > \mu \end{cases}$$

as $n \to \infty$. How fast is this convergence? Well,

$$\pi_{n+m} = \mathbb{P}(S_{n+m} \geq (n+m)a) \geq \mathbb{P}(S_n \geq na, S_{n+m} - S_n \geq ma) = \pi_n \pi_m. \tag{21}$$

Let us define $\ell_n \stackrel{\text{def}}{=} \log \pi_n \in [-\infty, 0]$. Equation (21) precisely says that ℓ_n is subadditive, i.e., $\ell_m + \ell_n \leq \ell_{n+m}$. As with any subadditive sequence, we have

$$\frac{\ell_n}{n} \to \ell \stackrel{\text{def}}{=} \sup_m \frac{\ell_m}{m}.$$

Definition 7.24.1. The rate function is

$$\gamma(a) \stackrel{\text{def}}{=} \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq na). \qquad \triangle$$

Note that if $a > \mu$ and $\gamma(a) < 0$ then

$$\frac{1}{n} \log \mathbb{P}(S_n \geq na) \to \gamma(a) < 0$$

is saying $\mathbb{P}(S_n \geq a) \approx e^{n\gamma(a)}$. Thus we get exponentially fast convergence. (!)

Let's contrast this with the central limit theorem.

Example 7.24.2. Consider a simple random walk in $\mathbb{Z}$: define random variables X_i which are equal to 1 or -1 with probability $\frac{1}{2}$. The central limit theorem says that $S_n/\sqrt{n} \stackrel{\text{d}}{\to} N(0,1)$, so in particular $\mathbb{P}(S_n \geq b\sqrt{n}) \to \mathbb{P}(Z > b)$ for $Z \sim N(0,1)$. Contrast this to the study of large deviations, which studies how quickly we get $\mathbb{P}(S_n \geq na) \to 0$ or $\mathbb{P}(S_n \geq na) \to 1$. $\triangle$

When can we expect exponential decay of $\mathbb{P}(S_n > na)$? We'd at least need $a > \mu$. We also need to impose tail bounds on S_n: if the S_n is heavy-tailed then some summands in S_n mess the estimates.

An elegant way of expressing the tail bounds is in terms of the Laplace transform.

Definition 7.24.3. The Laplace transform of a random variable X is the function

$$\psi_X(\theta) = \mathbb{E}(e^{\theta X}). \qquad \triangle$$

This is the characteristic function without the i in the exponent. We've been studying characteristic functions because the random variables $e^{i\theta X}$ were automatically bounded; without the i this is no longer true. Thus, the assumption $\psi_X(\theta) < \infty$ for all θ is effectively a tail bound on X. (At the very least, X should decay exponentially.)

Example 7.24.4.

- Let $X \sim N(0,1)$. The normal distribution has thin-enough tail bounds to make $\psi_X(\theta) < \infty$: we have $\psi_X(\theta) = e^{\theta^2/2} < \infty$ for each θ.
- Let $X \sim \text{Exp}(\lambda)$. Then $\psi_X(\theta)$ is finite in an interval around 0, but not for all θ. △

Markov's inequality (Lemma 3.8.4) says

$$e^{\theta na}\mathbb{P}(S_n \geq na) \leq \mathbb{E}(e^{\theta S_n}) = \mathbb{E}e^{\theta(X_1+\cdots+X_n)} = \mathbb{E}(e^{\theta X_1}\dots e^{\theta X_n}) = \mathbb{E}(e^{\theta X_1})\dots\mathbb{E}(e^{\theta X_n}) = \psi_{X_1}(\theta)^n.$$

In other words, we obtain

$$\mathbb{P}(S_n \geq na) \leq \frac{e^{\theta na}}{\psi_{X_1}(\theta)^n} = \exp(-(\log\psi(\theta) - a\theta)n). \tag{22}$$

Theorem 7.24.5 (Cramer's Theorem)**.** *For all $a > \mathbb{E}X_1$, we have*

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(S_n \geq na) = -I(a),$$

where $I(a) = \sup_{\theta\in\mathbb{R}}(a\theta - \log\psi(\theta))$.

Note that the limit is necessarily negative, and that $I(a)$ is positive (for convenience's sake).

Example 7.24.6. Let $X_1 \sim N(0,1)$ and

$$I(a) = \sup_{\theta\in\mathbb{R}}\left(a\theta - \frac{\theta^2}{2}\right) = \frac{a^2}{2},$$

which is telling us something we already knew, namely, that the sum of i.i.d. Gaussians is again Gaussian. [... right?] △

Example 7.24.7. Let X_1 be equal to 1 or -1 with probability $\frac{1}{2}$. We have

$$\psi_{X_1}(\theta) = \frac{e^\theta}{2} + \frac{e^{-\theta}}{2},$$

so we need to find

$$\sup_{\theta\in\mathbb{R}}\Bigg(\underbrace{a\theta - \log\left(\frac{e^\theta}{2} + \frac{e^{-\theta}}{2}\right)}_{\stackrel{\text{def}}{=} f(\theta)}\Bigg).$$

This is a calculus exercise:

$$f'(\theta) = a - \frac{e^\theta - e^{-\theta}}{e^\theta + e^{-\theta}}.$$

So we'd need to set $\theta = \tanh^{-1} a$, and

$$I(a) = \sup_{\theta\in\mathbb{R}} f(\theta) = a\tanh^{-1}a - \log(\cosh\tanh^{-1}a).$$ △

We proved the upper bound part of Cramer's Theorem (Theorem 7.24.5) when we wrote down a bunch of Markov inequalities Equation (22). The idea of the proof of the lower bound is as follows. On the unlikely event that $S_n \geq na$, the summands $X_1, X_2, \dots, X_n$ behave like i.i.d. random variables with a "tilted" distribution, namely:

Definition 7.24.8. If X_1 has distribution F, then define a distribution F_λ by

$$F_\lambda(x) \stackrel{\text{def}}{=} \frac{1}{\psi(\lambda)}\int_{-\infty}^{x} e^{\lambda y}\,dF(y).$$ △

Returning to our discussion, the summands $X_1, \dots, X_n$ behave like i.i.d. random variables with the tilted distribution F_{θ_*}, where θ_* is the maximiser of the function $\log\psi(\theta) - a\theta$. [It turns out that θ_* actually exists, because $\log\psi$ is convex.]

In the next (last!) HW, we'll show that if $\psi_X(\theta) < \infty$ in an interval around 0, then $\mathbb{E}|X|^n < \infty$ for all n and furthermore that

$$\psi_X(\theta) = \sum_{n\geq 0}\frac{\theta^n}{n!}\mathbb{E}X^n.$$

7.25 Dec 4, 2019 (Random Series)

Let's begin with an example. Let's consider the harmonic series with random signs, i.e., let us define the partial sums

$$S_n \stackrel{\text{def}}{=} \sum_{k=1}^{n} X_k, \quad \text{with} \quad X_k = \begin{cases} \frac{1}{k} & \text{with probability } \frac{1}{2} \\ -\frac{1}{k} & \text{with probability } \frac{1}{2} \end{cases}$$

and ask for the quantity $\mathbb{P}(\lim_{n\to\infty} S_n \text{ exists})$. We have $\text{Var}(X_n) = \mathbb{E}X_n^2 = \frac{1}{n^2}$. Hence

$$\text{Var}(S_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = \frac{1}{1^2} + \cdots + \frac{1}{n^2} \to \frac{\pi^2}{6} \quad \text{as } n \to \infty.$$

Hence, the following theorem tells us $\mathbb{P}(\lim_{n\to\infty} S_n \text{ exists})$:

Theorem 7.25.1. *If $X_1, X_2, \dots$ are independent, with $\mathbb{E}X_n = 0$ and $\sum_{n\geq 1} \text{Var}(X_n) < \infty$, then $S_n = X_1 + \cdots + X_n$ converges a.s..*

Note that unlike the strong law of large numbers, the limit is random: the variance of the limiting random variable S_∞ has variance $\text{Var}(S_\infty) = \sum_{n\geq 1} \text{Var}(X_n)$.

Let's talk about the Kolmogorov 0-1 law. To do this we begin with a definition:

Definition 7.25.2. The tail σ-field of a sequence of random variables $X_1, X_2, \dots : \Omega \to \mathbb{R}$ is

$$\mathcal{T} = \bigcap_{n\geq 1} \sigma(X_{n+1}, X_{n+2}, \dots).$$

Intuitively, $A \in \mathcal{T}$ if changing finitely many X_n values doesn't change whether A occurs. △

Example 7.25.3. We have $\{X_n > 0 \text{ i.o.}\} \in \mathcal{T}$ and $\{X_n > 5 \text{ eventually}\} \in \mathcal{T}$. More relevantly, if $S_n = X_1 + \cdots + X_n$, then $\{\lim S_n \text{ exists in } \mathbb{R}\} \in \mathcal{T}$.

Note, though, that $\{\lim S_n = \pi\} \notin \mathcal{T}$ (since changing the value of X_1, for example, changes the value of $\lim S_n$). On the other hand, $\{\lim \frac{S_n}{n} = \pi\} \in \mathcal{T}$, since now changing the value of X_1 doesn't change the value of the limit. △

Theorem 7.25.4 (Kolmogorov 0-1 law)**.** *If $X_1, X_2, \dots$ are independent, then for all $A \in \mathcal{T}$ either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.*

Note that the following proof gives us no information about whether probability is 0 or 1. (There are open problems which ask, for specific A, whether it has probability 0 or 1!)

Proof. The idea is to show that A is independent of itself (!). Then

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

so $\mathbb{P}(A) \in \{0, 1\}$.

Recall that for all $k, j \geq 1$ the σ-fields

$$\sigma(X_1, \dots, X_k) \perp \sigma(X_{k+1}, \dots, X_{k+j}).$$

(The notation $\mathcal{A} \perp \mathcal{B}$ means that $\mathcal{A}$ and $\mathcal{B}$ are independent, i.e., $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$.)

Recall also that if $\mathcal{A}$ and $\mathcal{B}$ are π-systems, then $\mathcal{A} \perp \mathcal{B}$ implies $\sigma(\mathcal{A}) \perp \sigma(\mathcal{B})$.

So let

$$\mathcal{A} = \sigma(X_1, \dots, X_k)$$
$$\mathcal{B} = \bigcup_{j\geq 1} \sigma(X_{k+1}, \dots, X_{k+j}).$$

Then $\mathcal{A} \perp \mathcal{B}$ and hence $\mathcal{A} \perp \sigma(\mathcal{B}) = \sigma(X_{k+1}, X_{k+2}, \dots)$.

Now we let

$$\mathcal{A} = \bigcup_{k \geq 1} \sigma(X_1, \ldots, X_k)$$
$$\mathcal{B} = \mathcal{T}$$

and note that if $A \in \mathcal{A}$ then there exists k so that $A \in \sigma(X_1, \ldots, X_k)$. Since

$$\sigma(X_1, \ldots, X_k) \perp \sigma(X_{k+1}, \ldots, X_{k+2}, \ldots) \supseteq \mathcal{T}$$

we get $\mathcal{A} \perp \mathcal{T}$. Then $\sigma(\mathcal{A}) \perp \mathcal{T}$, and $\mathcal{T} \subseteq \sigma(\mathcal{A})$ gives $\mathcal{T} \perp \mathcal{T}$. □

Theorem 7.25.5 (Kolmogorov maximal inequality)**.** *Let $X_1, \ldots, X_n$ be independent with mean 0 and finite variance, and let $S_n = X_1 + \cdots + X_n$. Then*

$$\mathbb{P}(\max_{k=1}^{n} |S_k| \geq y) \leq \frac{\mathrm{Var}(S_n)}{y^2}.$$

(Compare this to Chebyshev's inequality (special case of Markov; Lemma 3.8.4), which would say $\mathbb{P}(|S_n| \geq y) \leq \frac{\mathrm{Var}(S_n)}{y^2}$.)

Proof. Let

$$\mathrm{Var}(S_n) = \mathbb{E}S_n^2 = \int_\Omega S_n^2 \, d\mathbb{P} \geq \int_{A_1 \sqcup \cdots \sqcup A_n} S_n^2 \, d\mathbb{P},$$

where

$$A_k = \{\omega \in \Omega \colon |S_k| \geq y, \text{ and } |S_j| < y \text{ for all } j \in [k-1]\}.$$

We obtain

$$\begin{aligned}
\mathrm{Var}(S_n) &\geq \sum_{k=1}^{n} \int_{A_k} S_n^2 \, d\mathbb{P} \\
&= \sum_{k=1}^{n} \int_{A_k} (S_k + (S_n - S_k))^2 \, d\mathbb{P} \\
&\geq \sum_{k=1}^{n} \left(\int_{A_k} S_k^2 \, d\mathbb{P} + 2 \int_{A_k} S_k (S_n - S_k) \, d\mathbb{P} \right) \\
&= \sum_{k=1}^{n} \int_{A_k} S_k^2 \, d\mathbb{P} + 2 \int_\Omega (S_k \mathbb{1}_{A_k})(S_n - S_k) \, d\mathbb{P}.
\end{aligned}$$

Now the random variable $S_k \mathbb{1}_{A_k}$ is in $\sigma(X_1, \ldots, X_k)$ and $S_n - S_k$ is in $\sigma(X_{k+1}, \ldots, X_n)$, so they are independent. Furthermore $\mathbb{E}(S_n - S_k) = 0$, so we have

$$2 \int_\Omega (S_k \mathbb{1}_{A_k})(S_n - S_k) \, d\mathbb{P} = 0.$$

In other words,

$$\mathrm{Var}(S_n) \geq \sum_{k=1}^{n} \int_{A_k} S_k^2 \, d\mathbb{P} \geq \sum_{k=1}^{n} \mathbb{P}(A_k) y^2.$$

Then

$$\mathbb{P}(\max_{k=1}^{n} |S_n| \geq y) = \mathbb{P}(A_1 \sqcup \cdots \sqcup A_n) = \sum_{k=1}^{n} \mathbb{P}(A_k) \leq \frac{\mathrm{Var}(S_n)}{y^2}.$$

□

Proof of Theorem 7.25.1. We'll use the following fact:

Exercise: If W_n is decreasing and $W_n \to 0$ in probability, then $W_n \to 0$ almost surely.

To prove Theorem 7.25.1 we need to show that $\mathbb{P}(S_n \text{ is a Cauchy sequence }) = 1$. Let us fix $M, N \in \mathbb{N}$. We have

$$\max_{k=M+1}^{N} |S_k - S_M| \uparrow \sup_{k \geq M+1} |S_k - S_m| \quad \text{as } N \to \infty.$$

By Kolmogorov's maximal inequality (Theorem 7.25.5) we have

$$\mathbb{P}(\max_{k=M+1}^{N} |S_k - S_M| > \varepsilon) \leq \frac{1}{\varepsilon^2} \sum_{k=M+1}^{N} \text{Var}(X_k).$$

Let

$$W_M = \sup_{k,\ell > M} |S_k - S_\ell| \leq \sup_{k,\ell > M} (|S_k - S_M| + |S_\ell - S_M|) = 2 \sup_{k \geq M} |S_k - S_M|.$$

Note that W_M is decreasing. We want to show $W_M \to 0$ almost surely, and the exercise says it suffices to show that it goes to 0 in probability. We estimate

$$\mathbb{P}(W_M > 2\varepsilon) \leq \mathbb{P}(\sup_{k \geq M} |S_k - S_M > \varepsilon) = \lim_{N \to \infty} \mathbb{P}(\max_{k=M+}^{N} |S_k - S_M| > \varepsilon) \leq \sum_{k=M+1}^{\infty} \text{Var}(X_k) \downarrow 0 \text{ as } M \to \infty,$$

since $\sum \text{Var}(X_k) < \infty$. We're done:

$$\mathbb{P}(S_n \text{ Cauchy}) = \mathbb{P}(W_M \to 0 \text{ as } M \to \infty) = 1. \qquad \square$$

Let's go back to our original random series. We were considering the harmonic series with random signs, with partial sums $S_n = \sum_{k=1}^{n} X_k$ and $X_k = \pm \frac{1}{k}$ with probability $\frac{1}{2}$, and we observed that the sum of the variances is $\sum_k \frac{1}{k^2} < \infty$, so Theorem 7.25.1 says the limit converges almost surely.

In fact, we may consider $S_n = \sum_{k=1}^{n} X_k$ with $X_k = \pm \frac{1}{k^s}$ with probability $\frac{1}{2}$, for any $s > \frac{1}{2}$. In this case, the sum of the variances is $\sum_k \frac{1}{k^{2s}} < \infty$, and Theorem 7.25.1 still says that the limit converges almost surely.

What happens when $s \leq \frac{1}{2}$? In this case $\sum \text{Var}(X_n) = \infty$. One can show, using Lindeberg-Feller CLT (Theorem 6.20.1) that

$$\frac{S_n}{\sqrt{\text{Var}(S_n)}} \xrightarrow{\text{d}} N(0,1).$$

Since $\text{Var}(S_n)$ is going to infinity, it follows that S_n does not converge almost surely.

Let's end by stating Kolmogorov's strengthening of Theorem 7.25.1

Theorem 7.25.6 (Kolmogorov 3 series theorem)**.** *Let $X_1, X_2, \ldots$ be independent, and let $S_n = X_1 + \cdots + X_n$. Fix $A > 0$ and let $Y_n = X_n \mathbb{1}_{\{|X_n| < A\}}$. Then S_n converges if and only if all three of the following converge:*

1. $\sum_{n \geq 1} \mathbb{P}(|X_n| > A)$
2. $\sum_{n \geq 1} \mathbb{E}(Y_n)$
3. $\sum_{n \geq 1} \text{Var}(X_n)$

7.26 Dec 9, 2019 (Moments and convergence in distribution)

We're going to talk about moments and see how they might help us find limit laws. Often these are useful to prove convergence in distribution when the characteristic function is hard to compute. An example of this is Wigner's theorem, which we'll talk about in a bit.

Let's talk about the moment problem (in the bounded case): let μ and ν be probability measures on $\mathbb{R}$ with ν supported on an interval $[-A, A]$. Suppose

$$\int x^k \, d\mu(x) = \underbrace{\int x^k \, d\nu(x)}_{<\infty}$$

for every positive integer $k \geq 1$.

Theorem 7.26.1. *These assumptions imply $\mu = \nu$.*

Proof. We want to show

$$\int f \, d\mu = \int f \, d\nu$$

for all bounded continuous $f : \mathbb{R} \to \mathbb{R}$; we have this equality for the (nonbounded!) monomials x^k. The idea is to use Weierstrass approximation theorem: for fixed $B > A$ and $\delta > 0$, there exists a polynomial $P(x) = a_0 + a_1 x + \cdots + a_k x^k$ such that $|p - f| \leq \delta$ on $[-B, B]$. For brevity of notation we're going to denote by

$$\langle f, \mu \rangle \stackrel{\text{def}}{=} \int f \, d\mu.$$

Because p is a polynomial, $\langle p, \mu \rangle = \langle p, \nu \rangle$, and furthermore

$$|\langle f, \mu \rangle - \langle f, \nu \rangle| \leq |\langle f, \mu \rangle - \langle p, \mu \rangle| + \underbrace{|\langle p, \mu \rangle - \langle p, \nu \rangle|}_{=0} + \underbrace{|\langle p, \nu \rangle - \langle f, \nu \rangle|}_{\leq 2\delta A}.$$

Thus we need to estimate $|\langle f, \mu \rangle - \langle p, \mu \rangle|$. To do this we split

$$f = f \mathbb{1}_{\{|x| < B\}} + f \mathbb{1}_{\{|x| > B\}};$$

since $|\langle f \mathbb{1}_{\{|x|<B\}}, \mu \rangle - \langle p \mathbb{1}_{\{|x|<B\}} \rangle| \leq 2\delta A$, it suffices to estimate $|\langle f \mathbb{1}_{\{|x|>B\}}, \mu \rangle - \langle p \mathbb{1}_{\{|x|>B\}} \rangle|$. To do this we'll use Chebyshev, who says

$$B^k \int x^k \mathbb{1}_{\{|x|>B\}} \, d\mu(x) \leq \int x^{2k} \, d\mu(x) = \int x^{2k} \, d\nu(x) \leq A^{2k}.$$

In other words,

$$\int x^k \mathbb{1}_{\{|x|>B\}} \, d\mu(x) \leq \left(\frac{A^2}{B} \right)^k.$$

Thus taking $B = \max(2A^2, 1)$, the left side is increasing in k and the right side is decreasing (to 0 (!)) in k, so the inequality can only hold if the left side is 0 for every k. We've proven $\langle p \mathbb{1}_{\{|x|>B\}}, \mu \rangle = 0$. Finally,

$$|\langle f \mathbb{1}_{\{|x|>B\}}, \mu \rangle - \langle p \mathbb{1}_{\{|x|>B\}} \, \mu \rangle| \leq \underbrace{|\langle f \mathbb{1}_{\{|x|>B\}}, \mu \rangle|}_{\to 0 \text{ as } B \to \infty} + \underbrace{|\langle p \mathbb{1}_{\{|x|>B\}} \, \mu \rangle|}_{=0}$$

This completes the proof. □

Example 7.26.2. Theorem 7.26.1 may fail when ν does not have bounded support. For example, we may consider the lognormal distribution $X = e^Z$ where $Z \sim N(0, 1)$ has moments

$$\mathbb{E}X^k = \mathbb{E}(e^{kZ} = \int_{-\infty}^{\infty} e^{kx} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx = \int_{-\infty}^{\infty} e^{-(x-k)^2/2} \sqrt{2\pi} e^{k^2/2} \, dx = e^{k^2/2}.$$

There are many random variables which have the same moments as X. For example, the discrete random variable Y with

$$\mathbb{P}(Y = e^n) = \frac{e^{-n^2/2}}{C}, \quad n \in \mathbb{Z}$$

with $C = \sum_{m\in\mathbb{Z}} e^{-m^2/2}$ the normalizing factor. Now clearly $X \overset{\mathrm{d}}{\neq} Y$ and it's not so bad to check that $\mathbb{E}Y^k = e^{k^2/2} = \mathbb{E}X^k$ for all $k \geq 1$ integer. △

The following can be found in Durrett:

Theorem 7.26.3. *If* $\limsup_{k\to\infty} \frac{(m_{2k})^{1/2k}}{2k} < \infty$ *then there is at most one distribution function* F *with*

$$\int_{-\infty}^{\infty} x^k \, dF(x) = m_k$$

for all $k \geq 1$.

Note that if ν is supported on $[-A, A]$, then

$$\mu_k = \int x^k \, d\nu(x) \leq A^k,$$

and $m_k^{1/k} \leq A$. So the theorem says that $m_k^{1/k}$ doesn't have to be bounded uniformly (in this case, by A), but can grow (as long as it's less than linearly in k).

Corollary 7.26.4. *If* Z *satisfies*

$$\mathbb{E}Z^k = \begin{cases} 0 & k \text{ odd} \\ (k-1)!! & k \text{ even} \end{cases}$$

then $Z \overset{\mathrm{d}}{=} N(0,1)$.

Note that in the lognormal example (Example 7.26.2), we have $m_k = e^{k^2/2}$ so $m_k^{1/k} = e^{k/2}$ grows too fast, and Theorem 7.26.3 doesn't apply.

Proof idea for Theorem 7.26.3. Use the fact that

$$\mathbb{E}e^{itX} = \sum \mathbb{E}\frac{(itx)^k}{k!} = \sum_{k\geq 0} \frac{(it)^k}{k!} m_k$$

converges in a neighborhood of 0. □

Lemma 7.26.5. *If* X *is determined by its moments, and*

$$\mathbb{E}(X_n)^k \to \mathbb{E}X^k$$

for all $k = 1, 2, \ldots$, *then* $X_n \overset{\mathrm{d}}{\to} X$.

Proof sketch. First we show that $\{X_n\}$ is a tight sequence: $B^2\mathbb{P}(|X_n| > B) \leq \mathbb{E}X_N^2 \to \mathbb{E}X^2$. This implies there exists a subsequence X_{n_j} converging in distribution [I think this is Corollary 6.17.8]. Then we show that for any subsequence $X_{n_j} \overset{\mathrm{d}}{\to} Y$, we have $\mathbb{E}X_{n_j}^k \to \mathbb{E}Y^k$ for $k = 1, 2, \ldots$, and hence $Y \overset{\mathrm{d}}{=} X$. □

We've seen in Theorem 7.26.1 and in Corollary 7.26.4 that when X is bounded or $N(0,1)$ that X is determined by its moments, so we obtain

Corollary 7.26.6. *If* X *has bounded support or* $X \sim N(0,1)$, *and* $\mathbb{E}X_n^k \to \mathbb{E}X^k$, *then* $X_n \overset{\mathrm{d}}{\to} X$.

Let's now talk about random matrices and Wigner's semicircle law.

Definition 7.26.7. An $N \times N$ matrix

$$M = \frac{1}{\sqrt{N}} \begin{bmatrix} M_{11} & M_{12} & \dots & M_{1N} \\ M_{21} & M_{22} & \dots & M_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ M_{1N} & M_{2N} & \dots & M_{NN} \end{bmatrix}$$

is called a Wigner matrix if

1. For $i \neq j$ we have $M_{ij} = M_{ji}$ and $\{M_{ij}\}_{i>j}$ are i.i.d., and furthermore $\mathbb{E}M_{ij} = 0$, $\mathbb{E}M_{ij}^2 = 1$,
2. The diagonal entries M_{ii} are also i.i.d. (but possibly different from $\{M_{ij}\}_{i>j}$) with $\mathbb{E}M_{ii} = 0$ and $\mathbb{E}M_{ii}^2 < \infty$.

These are going to be the Wigner matrices for which we can say some sort of limit law. But for simplicity of proofs for today, we're going to further assume that $\mathbb{E}M_{ij}^k < \infty$ for all k, that $M_{ii} = 0$, and $M_{ij} \stackrel{\mathrm{d}}{=} -M_{ij}$. $\triangle$

Wigner studied these because he was interested in Hamiltonian systems from quantum mechanics. But they come up in many contexts; one can think of them as a multiplicative analogue of what we've been thinking about. Random matrices even show up in the study of the distributions of zeros of the Riemann zeta function.

Since a Wigner matrix M is symmetric, it has N real eigenvalues, say $\lambda_1 \leq \cdots \leq \lambda_N$. Let us define the empirical distribution to be

$$\mu_N = \frac{1}{N} \sum_{i=1}^{n} \delta_{\lambda_i}.$$

This is a random measure; for $A \subseteq \mathbb{R}$ Borel we have

$$\mu_N(A) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{\lambda_j \in A\}}.$$

Define

$$\overline{\mu_N}(A) = \mathbb{E}\mu_N(A).$$

This is just a regular old measure.

Theorem 7.26.8 (Wigner's semicircle law, 1955)**.** *Let $\sigma(x)\,dx$ denote the measure corresponding to integration (with respect to Lebesgue) against $\sigma(x) = \frac{1}{2\pi}\sqrt{4-x^2}$. Then*

$$\mathbb{P}(\mu_N \stackrel{\mathrm{d}}{\to} \sigma(x)\,dx) = 1.$$

Proof outline. One first computes the moments of $\sigma(x)\,dx$. For k even, they turn out to be the Catalan numbers:

$$\begin{aligned} m_k &= \int_{-2}^{2} x^k \sigma(x)\,dx \\ &= \int_0^{\pi} (2\cos\theta)^k \left(\frac{2\sin\theta}{2\pi}\right)(2\sin\theta\,d\theta) \\ &= \frac{2^{2k+1}}{\pi} \int_0^{\pi} \cos\theta^k (1-\cos^2\theta)\,d\theta \\ &= \begin{cases} \frac{1}{k/2+1}\binom{k}{k/2} & k \text{ even} \\ 0 & k \text{ odd} \end{cases} \end{aligned}$$

with the change of variables $x = 2\cos\theta$. The Catalan numbers count a lot of combinatorial things, but importantly for us, they count the number of paths from $(0,0)$ to $(2k,0)$ using steps $(1,1)$ or $(1,-1)$ staying weakly above the x-axis.

The idea of the proof is to show that

$$\overline{\mu_N} \xrightarrow{\mathrm{d}} \sigma(x)\,dx \quad \text{as } N \to \infty$$

using the moments

$$m_k = \int x^k \, d\overline{\mu_N} = \frac{1}{N}\mathbb{E}\left(\sum_{i=1}^{N}(\lambda_i)^k\right) = \frac{1}{N}\mathbb{E}(\mathrm{Tr}(M^k)) = \frac{1}{N}$$
$$= \sum_{\{i_1,\ldots,i_k,i_{k+1}=i_1\}} \mathbb{E}[M_{i_1,i_2}M_{i_2,i_3}\ldots Mi_k, i_{k+1}].$$

Since the $\mathbb{E}M_{i,j} = 0$, the nonzero terms in the sum are

$$\mathbb{E}M_{e_1}^2\,\mathbb{E}M_{e_2}^2\ldots\mathbb{E}M_{e_{k/2}}^2,$$

where the $e_k = \{i_j, i_{j+1}\}$ are thought of as "edges". These are encoded using labelled trees on $k/2+1$ vertices, of which there are (constant multiple of) Catalan many. The scaling by $\sqrt{N}$ cancels out the constant multiplication and everything works out.

It turns out $\overline{\mu_N}$ is close to μ_N, and will complete the proof. □