Probability Theory

Daniele Zago

November 6, 2021

Contents

Lecture 0: Probability review

2021-10-11

References Çinlar (2011, §1-2)
Paolella (2007)

In this section we summarize a (hopefully useful) review of concepts which can serve as a basis for the following lectures.

0.1 Probability spaces

Let E be a set, we want to define some useful quantities to build the notion of a probability space, that is, a space onto which a probability measure can be defined.

Def. (Sigma-algebra)
A non-empty collection $\mathcal{E}$ of subsets of E is called a σ-algebra on E if

a) $E \in \mathcal{E}$

b) (Closure under ^{c}) $A \in \mathcal{E} \implies A^c \in \mathcal{E}$

c) (Closure under $\cap$) $A_1, A_2, \ldots \in \mathcal{E} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{E}$

Remarks

› Every σ-algebra on E includes E and $\emptyset$ at least, indeed $\mathcal{E} = \{\emptyset, \mathcal{E}\}$ is called the *trivial* σ-algebra.

› Conversely, the maximal sigma algebra on E is given by the power set of E denoted by $\mathscr{P}(E)$.

› A countable (or uncountable) intersection of σ-algebras on E is again a σ-algebra on E. Given a collection $\mathcal{C}$ of subsets of E, we define the σ-algebra *generated by* $\mathcal{C}$ as the intersection of all σ-algebras $\mathcal{E}$ on E which contain $\mathcal{C}$,

$$\sigma(\mathcal{C}) = \bigcap_{\mathcal{E} \,:\, \mathcal{C} \subseteq \mathcal{E}} \mathcal{E}.$$

› If E is a topological space, then the σ-algebra generated by the collection of all open subsets of E is called the *Borel σ-algebra* and is denoted by $\mathcal{B}(E)$. $B \in \mathcal{B}(E)$ is called a *Borel set*.

› Given two sets E and F with σ-algebras $\mathcal{E}$ and $\mathcal{F}$, we can define the *σ-algebra generated by the rectangles* on $E \times F$ as

$$\mathcal{E} \otimes \mathcal{F} = \sigma\big(\{A \times B : A \subseteq \mathcal{E}, B \subseteq \mathcal{F}\}\big).$$

Moreover, if $\mathcal{E}$ and $\mathcal{F}$ are the Borel σ-algebra on $\mathbb{R}$, we have

$$\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^2).$$

With the above definition of a σ-algebra, we can now define the basic type of space onto which a probability measure can be constructed.

Def. (Measurable space)

A *measurable space* is a pair $(E, \mathcal{E})$ where E is a set and $\mathcal{E}$ a σ-algebra on E. Elements of $\mathcal{E}$ are accordingly called *measurable sets*.

Let E and F be sets. A *function* $f : E \longrightarrow F$ is a rule that assigns an element $f(x) \in F$ to each $x \in E$. We are interested in a particular class of functions, namely those which are related to the sigma algebra defined on the spaces E and F.

Def. (Measurable function)

Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. A mapping $f : E \longrightarrow F$ is said to be *measurable* wrt to $\mathcal{E}$ and $\mathcal{F}$ if for every $B \in \mathcal{F}$,

$$f^{-1}(B) \in \mathcal{E}.$$

Prop. 1 (Measurable functions of measurable functions are measurable)

If f is measurable relative to $\mathcal{E}$ and $\mathcal{F}$ and g is measurable relative to $\mathcal{F}$ and $\mathcal{G}$, then $g \circ f : E \longrightarrow G$ given by $g \circ f(x) = g\big(f(x)\big)$ is measurable relative to $\mathcal{E}$ and $\mathcal{G}$.

Proof.
For $C \in \mathcal{G}$, we have that $(g \circ f)^{-1}(C) = f^{-1}\big(g^{-1}(C)\big)$. Now, $g^{-1}(C) \in \mathcal{F}$ since g is measurable, and therefore $f^{-1}\big(g^{-1}(C)\big) \in \mathcal{E}$ by the measurability of f.

□

Remark

› If μ is a measure on $\mathcal{E}$ and $f : E \to F$ is measurable wrt to E and F, then f induces a measure $\widehat{\mu}$ on $\mathcal{F}$ given by

$$\widehat{\mu}(B) = \mu\big(f^{-1}(B)\big), \quad B \in \mathcal{F}.$$

A probability space is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is a set (set of *outcomes*), $\mathcal{F}$ is a σ-algebra on Ω (set of *events*), and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$. Mathematically, a probability space is a measure space where the measure has a total mass of one.

The probability measure has the following properties, which are verified for all finite measures:

$$
\begin{aligned}
&(\textit{Norming}) && \mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1, \mathbb{P}(H) = 1 - \mathbb{P}(H^c) \\
&(\textit{Monotonicity}) && H \subset K \implies \mathbb{P}(H) \leq \mathbb{P}(K) \\
&(\textit{Finite additivity}) && H \cap K = \emptyset \implies \mathbb{P}(H \cup K) = \mathbb{P}(H) + \mathbb{P}(K) \\
&(\textit{Countable additivity}) && (H_n)_{n\in\mathbb{N}} \text{ disjoint} \implies \mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} H_n\Big) = \sum_{n\in\mathbb{N}} \mathbb{P}(H_n) \\
&(\textit{Sequential continuity}) && H_n \nearrow H \implies \mathbb{P}(H_n) \nearrow \mathbb{P}(H) \\
& && H_n \searrow H \implies \mathbb{P}(H_n) \searrow \mathbb{P}(H) \\
&(\textit{Boole's inequality}) && \mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} H_n\Big) \leq \sum_{n\in\mathbb{N}} \mathbb{P}(H_n).
\end{aligned}
$$

0.2 Random variables

Def. (Random variable)
Let $(E, \mathcal{E})$ be a measurable space. A mapping $X : \Omega \longrightarrow E$ is called a *random variable* provided that it be measurable relative to $\mathcal{F}$ and $\mathcal{E}$, that is, if for every $A \in \mathcal{E}$,

$$X^{-1}(A) = \{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}.$$

In general, we say that X is E-valued with the σ-algebra $\mathcal{E}$ that is understood from context.

Def. (Distribution of a random variable)
Let X be a random variable on $(E, \mathcal{E})$, then we define the *distribution of* X as the image of μ of $\mathbb{P}$ under X,

$$\mu(A) = \mathbb{P}\big(X^{-1}(A)\big) = \mathbb{P}\big(X \in A\big), \quad A \in \mathcal{E}.$$

Let X be a r.v. in $(E, \mathcal{E})$ and let $(F, \mathcal{F})$ be another measurable space. Let now $f : E \longrightarrow F$ a measurable function relative to $\mathcal{E}$ and $\mathcal{F}$, then the composition $Y = f \circ E$

$$Y(\omega) = f \circ X(\omega) = f\big(X(\omega)\big), \quad \omega \in \Omega$$

is a random variable taking values in $(F, \mathcal{F})$ (Prop 1). If μ is the distribution of X, then the distribution ν of Y is $\nu = \mu \circ f^{-1}$:

$$\nu(B) = \mathbb{P}\big(Y \in B\big) = \mathbb{P}\big(X \in f^{-1}(B)\big) = \mu\big(f^{-1}(B)\big), \quad B \in \mathcal{F}.$$

Def. (Joint distribution)

If X and Y are random variables on $(E, \mathcal{E})$ and $(F, \mathcal{F})$ respectively, then $Z = (X, Y)$ is random variable on $(E \times F, \mathcal{E} \otimes \mathcal{F})$ and the distribution of Z is called the *joint distribution of X and Y*, which is fully specified by

$$\pi(A \times B) = \mathbb{P}\big(X \in A, Y \in B\big), \quad \text{for all } A \in \mathcal{E}, B \in \mathcal{F}.$$

Def. (Marginal distribution)

If $Z = (X, Y)$ is a r.v. on $(E \times F, \mathcal{E} \otimes \mathcal{F})$ that has joint distribution π, then the *marginal distributions* of X and Y are, respectively,

$$\mu(A) = \pi(A \times F) \quad \text{and} \quad \nu(B) = \pi(E \times B).$$

Def. (Independence)

With the previous assumptions, X and Y are said to be *independent* if their joint distribution is

$$\mathbb{P}\big(X \in A, Y \in B\big) = \mathbb{P}\big(X \in A\big)\mathbb{P}\big(Y \in B\big), \quad A \in \mathcal{E}, B \in \mathcal{F}.$$

Remark

An arbitrary collection (countable or uncountable) of random variables is said to be *independent* if every finite subcollection $(X_{i_1}, \ldots, X_{i_n})$ is independent.
If X is a random variable, then its integral w.r.t. the measure $\mathbb{P}$ makes sense to talk about, since by definition it is $\mathcal{F}$-measurable.

Def. (Expected value)

The integral of X w.r.t the measure $\mathbb{P}$ is called the *expected value of X*,

$$\mathbb{E}\big[X\big] = \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_\Omega X d\mathbb{P}.$$

If $\mathbb{E}\big[X\big] < \infty$ then X is said to be integrable.

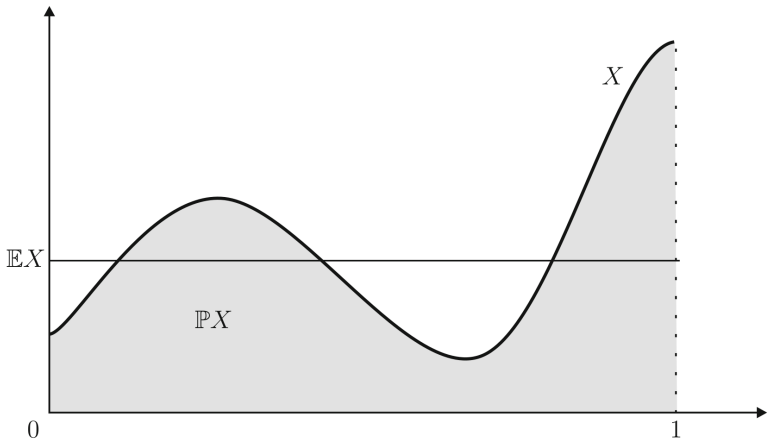

Figure 1: The integral $\mathbb{P}(X)$ is the area under X, the expected value $\mathbb{E}(X)$ is the constant "closest" to X.

Thm. 1 (Law of the unconscious statistician)
If X is a r.v. on $(E, \mathcal{E})$ and f is $\mathcal{E}$-measurable, then

$$\mathbb{E}\big[f(X)\big] = \int_\Omega f\big(X(\omega)\big)\mathbb{P}(d\omega)$$

Remark

Choosing $f(X) = \mathbb{1}_A$, we find that $\mathbb{E}\big[\mathbb{1}_A(X)\big] = \mathbb{P}\big(X \in A\big)$.

0.3 L^p spaces

Def. (p-norm)
For $p \in [1, \infty)$ we define the *p-norm of X* to be

$$\|X\|_p = \mathbb{E}\big[|X|^p\big]^{1/p},$$

and for $p = \infty$ we define it as the *essential supremum* of X

$$\|X\|_\infty = \inf_{b \in \mathbb{R}^+} \{|X| \leq b \text{ almost surely}\}.$$

Remarks

- $\|X\|_p = 0 \implies X \equiv 0$ almost surely.
- $\|cX\|_p = c\|X\|_p$ for $x \geq 0$.

We have a very famous theorem which defines the relationship between different random variable norms.

Thm. 2 (Hölder's inequality)
For $p, q, r \in [1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$,

$$\|XY\|_r \leq \|X\|_p \|Y\|_q,$$

in particular for $r = 1, p = 2, q = 2$ we have Schwartz's inequality

$$\|XY\|_1 \leq \|X\|_2 \|Y\|_2.$$

Thm. 3 (Minkowski's inequality)
For $p \in [1, \infty]$,

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$

Lemma 1 (Jensen's inequality)
Let D be a convex subset of $\mathbb{R}^d$ and $f : D \longrightarrow \mathbb{R}$ be continuous and concave. If $X_1, \ldots, X_d$ are integrable r.v. and $(X_1, \ldots, X_d) \in D$ almost surely. Then,

$$\mathbb{E}\big[f(X_1, \ldots, X_d)\big] \leq f\big(\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d]\big).$$

0.4 Generating functions

References Paolella (2007, §1)

Various integrals of interest are obtained by choosing an appropriate function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of two variables, (t, X), and are usually viewed as a function of t after integration wrt to X,

$$\mathbb{E}\big[g(t, X)\big] = \int_{-\infty}^{\infty} g(t, x) \, \mathrm{d}F_X(x).$$

Some notable examples of these functions include the following:

- *n-th moment*: $g(n, x) = x^n \implies \mathbb{E}\big[X^n\big]$
- *n-th abs. moment*: $g(n, x) = |x|^n \implies \mathbb{E}\big[|X|^n\big]$
- *Probability-generating function*: $g(t, x) = t^x \implies G(t) = \mathbb{E}\big[t^X\big]$. This function is useful for discrete random variables, since
  - $p(k) = \mathbb{P}\big(X = k\big) = \dfrac{1}{k!} \cdot \dfrac{\partial}{\partial t} G(t)\Big|_{t=0}$
  - $G_X = G_Y \implies p_X = p_Y$.
  - The k^{th} factorial moment is

$$\mathbb{E}\left[\frac{X!}{(X-k)!}\right] = \frac{\partial}{\partial t} G(t)\Big|_{t=1^-}$$

– If $M_X(t)$ is the moment-generating function of X, then

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

– If $N \sim \mathbb{P}_N$ and $S_N = \sum_{i=1}^N X_i$, with $X_i \overset{\text{iid}}{\sim} \mathbb{P}_X$ and $N \perp\!\!\!\perp X_i$, then using the law of total expectation we have

$$G_{S_N}(t) = \mathbb{E}_{\mathbb{P}_N}\Big[\mathbb{E}_{\mathbb{P}_X}\big[t^{\sum_{i=1}^N X_i}|N\big]\Big] = \mathbb{E}_{\mathbb{P}_N}\big[G_X(t)^N\big] = G_N\big(G_X(t)\big).$$

0.4.1 Moment-generating function

Def. (Moment-generating function)
The *moment-generating function* (mgf) of a random variable X is the function $t \mapsto e^{tX}$ and is said to *exist* if there is an $h > 0$ such that

$$\text{For all } t \in (-h, h), \quad M_X(t) < \infty.$$

Remark

› If $M_X(t)$ exists, then the *convergence strip* of $M_X(t)$ is the largest open interval such that $M_X(t) < \infty$,

$$\sup_h \big\{(-h, h) : M_X(t) < \infty \quad \forall t \in (-h, h)\big\}.$$

› For a location-scale family, if $Z = \mu + \sigma X$ we have that

$$M_Z(t) = \mathbb{E}\big[e^{t(\mu+\sigma X)}\big] = e^{\mu t} M_X(\sigma t).$$

› If $N \sim \mathbb{P}_N$ and $S_N = \sum_{i=1}^N X_i$, with $X_i \overset{\text{iid}}{\sim} \mathbb{P}_X$ and $N \perp\!\!\!\perp X_i$, then again by using the law of total expectation we have

$$M_{S_N}(t) = \mathbb{E}_{\mathbb{P}_N}\Big[\mathbb{E}_{\mathbb{P}_X}\big[e^{t\sum_{i=1}^N X_i}|N\big]\Big] = \mathbb{E}_{\mathbb{P}_N}\big[M_X(t)^N\big] = G_N\big(M_X(t)\big).$$

Thm. 4 (Existence of absoute moments)
If $M_X(t)$ exists, then for all $r \in (0, +\infty)$ we have that

$$\mathbb{E}\big[|X|^r\big] < \infty.$$

It can be shown that the derivative operator can be moved inside the expectation, and the moment-generating function can be used to compute the k^{th} moment of X.

Thm. 5 (Generation of moments)
If $M_X(t)$ exists, then we can write

$$\frac{\partial}{\partial t}M_X(t) = \frac{\partial}{\partial t}\mathbb{E}\big[e^{tX}\big] = \mathbb{E}\left[\frac{\partial}{\partial t}e^{tX}\right] = \mathbb{E}\big[X^j e^{tX}\big],$$

and therefore $\mathbb{E}\big[X^j\big] = \dfrac{\partial}{\partial t}M_X(t)\Big|_{t=0}$.

Example (mgf of DUnif(ϑ))
Let $X \sim \text{DUnif}(\vartheta)$, i.e. X is discrete with pmf

$$p_X(x;\vartheta) = \frac{1}{\vartheta}\mathbb{1}_{\{1,2,\dots,\vartheta\}}(x).$$

Then, the mgf of X is

$$M_X(t) = \mathbb{E}\big[e^{tX}\big] = \frac{1}{\vartheta}\sum_{i=1}^{\vartheta} e^{tj}.$$

From this, we can easily calculate $\mathbb{E}\big[X\big]$ simply by deriving wrt to t

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{\vartheta}\frac{\partial}{\partial t}\sum_{j=1}^{\vartheta} e^{tj}\Big|_{t=0} \\
&= \frac{1}{\vartheta}\sum_{j=1}^{\vartheta} j e^{tj}\Big|_{t=0} \\
&= \frac{1}{\vartheta}\sum_{j=1}^{\vartheta} j \\
&= \frac{1}{\vartheta}\frac{\vartheta(\vartheta+1)}{2} \\
&= \frac{\vartheta+1}{2}.
\end{aligned}$$

Example (mgf of Unif(0, 1))
Let $X \sim \text{Unif}(0,1)$, then we find that the mgf of X is

$$M_X(t) = \int_0^1 e^{tx}\,\mathrm{d}x = \frac{1}{t}(e^t - 1),$$

which exists finite for all $t \in (0,1)$. Since the Taylor expansion of $M_X(t)$ around zero is

$$\frac{e^t - 1}{t} \overset{t\approx 0}{=} \frac{1}{t}\left(t + \frac{t^2}{2} + \frac{t^3}{6} + \frac{t^4}{24} + \dots\right) = 1 + \frac{t}{2} + \frac{t^2}{6} + \dots = \sum_{j=0}^{\infty}\frac{t^j}{(j+1)!},$$

we have that the r^{th} derivative has only the r^{th} term constantly equal to 1 in t at the numerator, and therefore

$$\mathbb{E}\big[X^r\big] = \frac{1}{r+1}.$$

For the multivariate case, we have a straightforward generalization of the mgf using vector notation.

Def. (Multivariate moment-generating function)
Let X be a multivariate r.v, then its *moment-generating function* is

$$M_X(t) = \mathbb{E}\big[e^{t^\top X}\big].$$

Thm. 6 (Sawa)
Let X_1, X_2 be r.v.s such that $\mathbb{P}\big(X_1 > 0\big) = 1$ with joint mgf $M_{X_1,X_2}(t_1, t_2)$ which exists for $t_1 < \varepsilon$ and $|t_2| < \varepsilon$, $\varepsilon > 0$. Then, we have that

$$\mathbb{E}\Big[\Big(\frac{X_2}{X_1}\Big)^k\Big] = \frac{1}{\Gamma(k)} \int_{-\infty}^{0} (-t_1)^{k-1} \Big[\frac{\partial^k}{\partial t_2^k} M_{x_1,x_2}(t_1, t_2)\Big]_{t_2=0} \mathrm{d}t_1.$$

0.4.2 Cumulant-generating function

Def. (Cumulant-generating function)
Let $M_X(t)$ be the moment-generating function of a r.v. X. Then, the *cumulant-generating function* $K_X(t)$ of X is

$$K_X(t) = \log M_X(t).$$

Remarks

› If $S_n = \sum_{i=1}^n X_i$ with X_i i.i.d, then

$$K_{S_n}(t) = nK_X(t).$$

› The j^{th} derivative of K_X evaluated at $t = 0$ is the j^{th} cumulant of X,

$$\kappa_j = \frac{\partial^j}{\partial t^j} K_X(t)\Big|_{t=0},$$

where if $\mu_j = \mathbb{E}\big[X^j\big]$, the first four cumulants are given by (Pace and Salvan, 1997):

$$\kappa_1 = \mu_1,$$

$$\kappa_2 = \mu_2 - \mu_1^2,$$

$$\kappa_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^2,$$

$$\kappa_4 = \mu_4 - 3\mu_2^2 - 4\mu_1\mu_3 + 12\mu_1^2\mu_2 + 6\mu_1^4.$$

Example (cgf of a $\mathcal{N}(\mu, \sigma^2)$)
For $X \sim \mathcal{N}(\mu, \sigma^2)$ we have that the moment-generating function is

$$M_X(t) = e^{\mu t + \sigma^2 \frac{t^2}{2}} \implies K_X(t) = \log M_X(t) = \mu t + \sigma^2 \frac{t^2}{2}.$$

Therefore, the first two cumulants are

$$\begin{cases} \kappa_1 = \frac{\partial}{\partial t}\left(\mu t + \sigma^2 \frac{t^2}{2}\right)\Big|_{t=0} & = \mu, \\ \kappa_2 = \frac{\partial^2}{\partial t^2}\left(\mu t + \sigma^2 \frac{t^2}{2}\right)\Big|_{t=0} & = \sigma^2. \end{cases}$$

Other examples of cgf's can be found in (Paolella, 2007, pp. 8–10).

Lecture 1: Convergence and limit theorems

2021-10-14

References Gut (2009), first portion of the course

Email: stefano.pagliarani9@unibo.it

The course will be focussed on the stochastic processes portion of probability theory, after a brief reminder of limit theorems, conditional probability, and measure theory.

1.1 Convergence of random variables

Convergence of random variables is a little bit trickier than just real numbers.

Notation: AC is the set of absolutely continuous probability measures wrt the Lebesgue measure.

› *Absolute continuity*: if $\mu \in \text{AC}$ is absolutely continuous, we write

$$\mu(dx) = f(x)dx$$

› *Integration in measure spaces:* Let $X \sim \mu$, then by a theorem we have

$$\mathbb{E}\big[f(X)\big] = \int_{\mathbb{R}^d} f(x)\mu(dx), \tag{1}$$

and we can differentiate between two types of distribution:

a) μ discrete $\implies \mathbb{E}\big[X\big] = \sum_n xp(x)$

b) $\mu \in \text{AC} \implies \mathbb{E}\big[X\big] = \int_{\mathbb{R}^d} x \cdot f(x)dx$

Example (Intuition of convergence)
Consider $\mu_n = \text{Unif}_{[0,\frac{1}{n}]}$ for $n \in \mathbb{N}$, and it is absolutely continuous w.r.t. Lebesgue measure. This means that it admits a probability density which is defined by

$$\mu_n(dx) = \left(\begin{cases} n & \text{if } x \in [0, \frac{1}{n}] \\ 0 & \text{if } x \notin [0, \frac{1}{n}] \end{cases} \right) dx$$

It is intuitive to think that the measure is converging to a spike in zero, i.e.

$$\mu_n \xrightarrow{n\to\infty} \delta_0,$$

where δ_x denotes the Dirac delta distribution centered in x, such that $\delta_x(\{x\}) = 1$. We need to mathematically characterize this type of convergence in a more formal way than by intuition.

Maybe it could be that for any Borel set $A \subseteq \mathscr{B}(\mathbb{R})$,

$$\mu_n(A) \xrightarrow{n\to\infty} \delta_0(A),$$

but unfortunately this is wrong since we can see that, for $A = \{0\}$ and for all $n \in \mathbb{N}$:

$$\mu_n(\{0\}) = 0 \neq 1 = \delta_0(\{0\}).$$

So we can either throw out the idea that the uniform converges to a Dirac delta, or change the definition of convergence to accommodate for the behaviour in Figure 2.

Moreover, assume now that $X_n \sim \mu_n$ such that $\mu_n \xrightarrow{n\to\infty} \delta_0$, what can we say about the properties of X_n? In general (as we will see afterwards), this depends on the specific type of convergence that we assume.

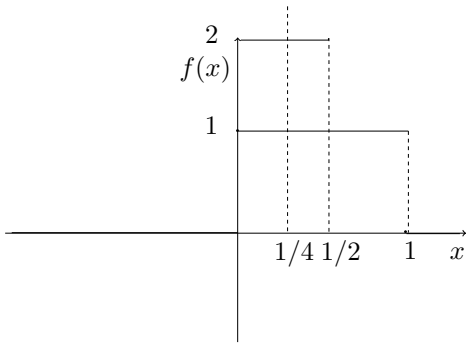

Figure 2: Convergence of the sequence of uniform distributions to the Dirac measure in zero.

Def. (Convergence in distribution)
Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of distributions on $(\mathbb{R}^d, \mathscr{B})$. We say that μ_n *converges in distribution* to another distribution μ,

$$\mu_n \xrightarrow{d} \mu,$$

if, for any possible choice of *test function* $f \in C_b(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} f(x)\mu_n(dx) \xrightarrow{n\to\infty} \int_{\mathbb{R}^d} f(x)\mu(dx).$$

This convergence is in the sense of standard real analysis.

Notation: $C_b(\mathbb{R}^d)$ is the set of continuous bounded functions

Remark

All test functions f define a measure when integrated wrt to $\mu_n(dx)$, and when all said measures are equal to those obtained by integrating against another distribution μ, then we obtain the convergence in distribution.

Example (Uniform distribution)

Consider $\mu_n = \text{Unif}_{[0,\frac{1}{n}]}$ and $\mu = \delta_0$, take any function $f \in C_b(\mathbb{R})$ and compute

$$\begin{aligned}\int_{\mathbb{R}} f(x)\mu_n(dx) &= \int_0^{\frac{1}{n}} f(x) \cdot n \cdot dx \\ &= n \cdot \underbrace{\int_{[0,\frac{1}{n}]} f(x)dx}_{\approx \frac{1}{n}\cdot f(0)} \\ &\xrightarrow{n\to\infty} f(0).\end{aligned}$$

The last equality holds since f is continuous, and by the mean value theorem we can approximate it by the left extrema. However, by definition of the abstract integral wrt the Dirac delta function we have that

$$f(0) = \int_{\mathbb{R}} f(x)\delta_0(dx),$$

which proves that $\mu_n \xrightarrow{d} \mu$.

Remark

If $A \in \mathscr{B}(\mathbb{R}^d)$ is an event and μ is a distribution, then

$$\mu(A) = \int_{\mathbb{R}^d} \mathbb{1}_A(x)dx,$$

where $\mathbb{1}_A$ is the indicator function such that

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

Had we used $f \notin C_b(\mathbb{R}^d)$ instead, then we could have chosen $f = \mathbb{1}_{\{0\}}$ and convergence in distribution would not have been satisfied. The example below shows another case in which another type of convergence is useful in order to characterize a common-sense behaviour of random variables.

Example (Sequence of Dirac functions)

Consider $\mu_n = \delta_{1/n}$ and $\mu = \delta_0$, then it is clear that this is a discrete measure that in some intuitive sense converges to zero. If we choose $f(x) = \mathbb{1}_{\{0\}}$, then we find that

$$\int_{\mathbb{R}} f(x)\mu_n(dx) = \int_{\mathbb{R}} \mathbb{1}_{\{0\}}(x)\delta_{\frac{1}{n}}(dx) = \mathbb{1}_{\{0\}}(1/n) = 0 \quad \forall n,$$

and therefore does not converges to δ_0.

Recall: A random variable is such that the event $(X_n \in A) \in \mathcal{F}_n$, which means that the function is measurable.

Def. (Weak convergence of random variables)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables, $X_n : (\Omega_n, \mathcal{F}_n, \mathbb{P}_n) \longrightarrow (\mathbb{R}^d, \mathscr{B})$. Let now X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Then, we say that X_n *converges weakly/in distribution/in law*, $X_n \xrightarrow{d} X$, if their measures are such that

$$\mu_{X_n} \xrightarrow{d} \mu_X.$$

Remark

By the definition of expected value in Equation (1), a family of random variables $(X_n)_{n\in\mathbb{N}}$ is such that, for any $f \in C_b(\mathbb{R}^d)$

$$X_n \xrightarrow{d} X \iff \mathbb{E}\big[f(X_n)\big] \xrightarrow{n\to\infty} \mathbb{E}\big[f(X)\big].$$

This is however the weakest type of convergence out of all those that we will consider, since in other cases the probability spaces might be different.

Def. (Stronger definitions of convergence)

$(X_n)_{n\in\mathbb{N}}$ sequence of random variables and X a r.v., all defined on the same probability space

$$X_n, X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (\mathbb{R}^d, \mathscr{B}).$$

Then we say that

a) If $X_n, X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ for $p \geq 1$, where

$$L^p = \big\{\text{r.v. on } (\Omega, \mathcal{F}, \mathbb{P}) \text{ such that } \mathbb{E}\big[|X|^p\big] < \infty\big\}.$$

then $X_n \xrightarrow{L^p} X$ if

$$\|X_n - X\|_{L^p} \xrightarrow{n\to\infty} 0,$$

where $\|X\|_{L^p} = \mathbb{E}\big[|X|^p\big]^{\frac{1}{p}}$.

b) X_n *converges in probability* to X, $X_n \xrightarrow{P} X$ if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(|X_n - X| \geq \varepsilon\big) = 0.$$

c) X_n *converges almost surely* to X, $X_n \xrightarrow{\text{a.s.}} X$ if

$$\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) = 1,$$

where the event inside $\mathbb{P}$ is in the sense of real analysis,

$$\Big\{w \in \Omega : X_n(\omega) \xrightarrow{n\to\infty} X(\omega)\Big\},$$

which can be proven to be a measurable set and therefore a valid event.

Remark

The L^p norm of the difference induces a *distance between functions* in the sense of functional analysis.

Example (Difference in interpretation)
Consider a Bernoulli game where we equally bet on an outcome ± 1. The second type of convergence does not tell us that almost surely our gain will converge to zero, but rather that we can set a small tolerance and find some n such that our gain will be smaller than that.

The following inequality is a basic tool for probability, which will be useful later on.

Thm. 7 (Markov's inequality)
Let X be a r.v. and $\lambda > 0$, then

$$\mathbb{P}\big(|X| > \lambda\big) \leq \frac{\mathbb{E}\big[|X|^p\big]}{\lambda^p}, \quad p \geq 0.$$

Proof.
If $\mathbb{E}\big[|X|^p\big] = \infty$, then there is nothing to prove. If instead $\mathbb{E}\big[|X|^p\big] < \infty$, then since $\mathbb{1}_A$ is either 1 or 0 we have

$$\begin{aligned}
\mathbb{E}\big[|X|^p\big] &\geq \mathbb{E}\big[|X|^p \cdot \mathbb{1}_{|X|>\lambda}\big] \\
&\geq \mathbb{E}\big[\lambda^p \cdot \mathbb{1}_{|X|>\lambda}\big] \qquad \text{(since } |X| \geq \lambda\text{)} \\
&= \lambda^p \cdot \mathbb{P}\big(|X| > \lambda\big).
\end{aligned}$$

□

Corollary 1 (Chebyshev's inequality)
By choosing $p = 2$ and considering the random variable $X - \mathbb{E}[X]$, Markov's inequality states that

$$\mathbb{P}\big[|X - \mathbb{E}[X]| > \lambda\big] \leq \frac{\mathbb{E}\big[|X - \mathbb{E}[X]|\big]^2}{\lambda^2} = \frac{\mathbb{V}\big[X\big]}{\lambda^2}.$$

Thm. 8
Under the according assumptions for X_n, X we have the following set of implications:

1. $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X$.
2. $X_n \xrightarrow{P} X \implies$ *there is a subsequence* X_{k_n} *such that* $X_{k_n} \xrightarrow{a.s.} X$.
3. $X_n \xrightarrow{d} X \implies X_n \xrightarrow{P} X$ *iff* $\mu_X = \delta_{x_0}$
4. $X_n \xrightarrow{L^1} X \implies X_n \xrightarrow{P} X$
5. $X_n \xrightarrow{P} X \implies X_n \xrightarrow{L^1} X$ *iff* $|X_n| \leq Y \in L^p$

Proof.

1. $\boxed{\text{a.s.} \implies \text{p}}$: $\mathbb{P}\big(|X_n - X| \geq \varepsilon\big) = \mathbb{E}\big[\mathbb{1}_{|X_n - X| \geq \varepsilon}\big]$ and the indicator function converges to zero as $n \to \infty$ by assumption. Since $\mathbb{1}_A$ is bounded, by the dominated convergence theorem the integral (expectation) also converges to zero.

4. $\boxed{L^p \implies p}$: Follows as a consequence of Markov's property, since we can majorize the probability by the expected value

$$\mathbb{P}\big(|X_n - X| \geq \varepsilon\big) \overset{\text{Thm.7}}{\leq} \frac{\mathbb{E}\big[|X_n - X|^p\big]}{\varepsilon^p} = \frac{\|X_n - X\|_{L^p}^p}{\varepsilon^p} \xrightarrow{n\to\infty} 0.$$

where the convergence to 0 is a consequence of the L^p convergence assumption.

□

Example (A.s. does not imply L^p)

Let $m \in \mathbb{R}$ and $X_n = n^m \mathbb{1}_{[0,\frac{1}{n}]}$ on the probability space $\big([0,1], \mathscr{B}([0,1]), \lambda_{[0,1]}\big) \to \mathbb{R}$, and let's try to establish some convergence for the random variable X_n.

› If $\omega > 0$, then we can find some $\bar{n}$ such that X_n is equal to zero:

$$X_n(\omega) = n^m \mathbb{1}_{[0,\frac{1}{n}]}(\omega) \xrightarrow{n\to\infty} 0.$$

› If $\omega = 0$, then

$$X_n(0) = n^m \xrightarrow{n\to\infty} +\infty, \quad \text{for } m > 0,$$

however the event $\{0\}$ has null probability since we have a uniform distribution on $\big[0, \frac{1}{n}\big]$ at all steps of the limit, and as such we have

$$\mathbb{P}_{\mu_n}\big(\{0\}\big) = 0 \quad \text{for all } n \in \mathbb{N}.$$

Therefore, the set of limit elements for absolute convergence is

$$\Big\{\omega \in \Omega : X_n(\omega) \xrightarrow{n\to\infty} X(\omega)\Big\} = \Omega \setminus \{0\}.$$

Since $\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) = \mathbb{P}\big(\Omega \setminus \{0\}\big) = 1$, we have that

$$X_n \xrightarrow{\text{a.s.}} X \equiv 0 \qquad (\implies X \xrightarrow{P} X).$$

On the other hand for L^p convergence we have that

$$\begin{aligned}\mathbb{E}\big[|X_n - X|^p\big] &= \mathbb{E}\big[|X_n|^p\big] \\ &= \int_{[0,1]} n^{mp} \cdot \mathbb{1}_{[0,\frac{1}{n}]}(x)dx \\ &= n^{mp} \cdot \frac{1}{n} \\ &= n^{mp-1}.\end{aligned}$$

We conclude that $X_n \xrightarrow{L^p} X \iff mp - 1 < 0 \iff m < 1/p$, but we always have almost-sure convergence for any $m > 0$.

Example (Gaussian distribution)

Consider $\mathcal{N}_{\mu,\sigma^2} = \varphi_{\mu,\sigma^2}(x)dx$, with

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}.$$

Consider now a sequence of real numbers $\mu_n \to \mu$ and a sequence of real numbers $\sigma_n \to 0$.

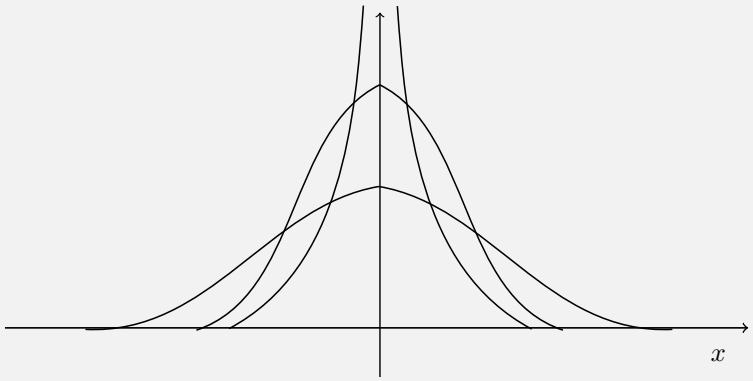

Figure 3: Convergence of the normal distribution to the Dirac delta function.

So we can expect that $\mathcal{N}_{\mu_n,\sigma_n} \xrightarrow{d} \delta_\mu$. As an exercise, prove this convergence (use a simple change of variables).

However, for the Gaussian case we can prove something stronger: if $X_n \sim \mathcal{N}_{\mu_n,\sigma_n}$ and $X \equiv \mu$ we can prove convergence in L^2. Using the triangle inequality, we can write

$$\mathbb{E}\big[|X_n - \mu|^2\big] \leq \mathbb{E}\big[|X_n - \mu_n|^2 + \underbrace{|\mu_n - \mu|^2}_{\to 0}\big],$$

and since $\mathbb{E}\big[|X_n - \mu_n|^2\big] = \mathbb{V}\big[X_n\big] = \sigma_n^2 \xrightarrow{n\to\infty} 0$, we also have L^2 convergence.

Exercise: prove that $\mathcal{N}_{\mu_n,\sigma_n} \xrightarrow{d} \delta_\mu$ if $\mu_n \to \mu$ and $\sigma_n \to 0$.

Proof.
Consider any test function $f \in \mathrm{C}_b(\mathbb{R})$, then if $\varphi(t)$ is the pdf of a $\mathcal{N}_{0,1}$ distribution we have that

$$\begin{aligned}\int_{\mathbb{R}} f(x)\mathcal{N}_{\mu_n,\sigma_n}(dx) &= \int_{\mathbb{R}} f(x)\cdot\frac{1}{\sigma_n}\cdot\varphi\left(\frac{x-\mu_n}{\sigma_n}\right)dx \qquad \text{(abs. continuity)}\\ &= \int_{\mathbb{R}} f\left(\sigma_n y+\mu_n\right)\cdot\cancel{\sigma_n}\frac{1}{\cancel{\sigma_n}}\varphi(y)dy \qquad \text{(change of var.)}.\end{aligned}$$

Since both f and φ are bounded the function $t \mapsto f(t)\varphi(t)$ is bounded by $g(t) = \max_{t'} f(t')\cdot\varphi(t)$, which is Lebesgue integrable and the dominated convergence theorem can be therefore applied to obtain the following equivalence

$$\lim_{n\to\infty}\int_{\mathbb{R}} f\left(\sigma_n y+\mu_n\right)\varphi(y)dy = \int_{\mathbb{R}}\lim_{n\to\infty} f(\sigma_n y+\mu_n)\varphi(y)dy = f(\mu)\int_{\mathbb{R}}\varphi(y)dy = f(\mu).$$

Therefore we have convergence in distribution to δ_μ by definition of the abstract integral wrt the Dirac measure.

□

Def. (C.d.f. of a distribution)
Given a distribution μ on $\mathbb{R}$, the cdf of μ is the function $F_\mu : \mathbb{R} \longrightarrow [0,1]$ defined by

$$F_\mu(x) = \mu\big((-\infty, x]\big).$$

Remark

Among all known properties such as monotonicity, boundedness, etc, the most important for what follows is the property of *right-continuity*.

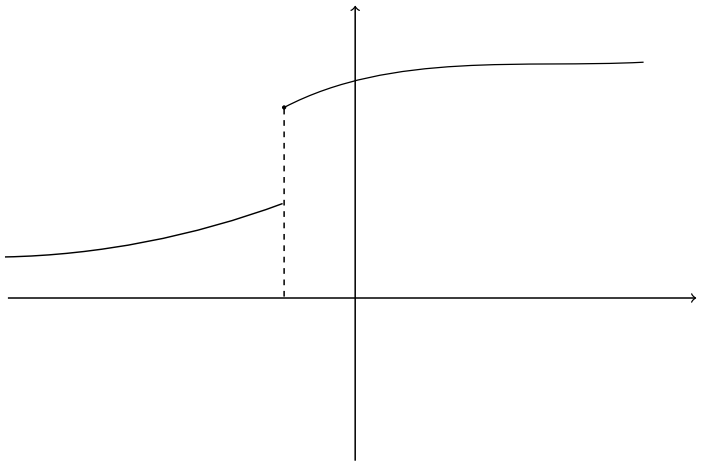

Figure 4: Right-continuity of the cumulative distribution function.

Def. (Cumulative distribution function)

Let X be a real-valued random variable, then the *cumulative distribution function* (CDF) of X is the function $F_X : \mathbb{R} \longrightarrow [0,1]$ defined by

$$F_X(x) = F_{\mu_X}(x) = \mathbb{P}\big(X \leq x\big)$$

Since the property of convergence in distribution is quite hard to prove for any bounded test function f, we want to characterize this property with respect to something else in order to make it easier to check it.

Example (Cdf of a uniform distribution)

Let $\mu_n = \text{Unif}_{[0,\frac{1}{n}]}$, then the cdf is

$$F_n(x) = \begin{cases} 0 & \text{if } x < 0 \\ nx & \text{if } 0 < x < \frac{1}{n} \\ 1 & \text{if } x \geq \frac{1}{n} \end{cases}$$

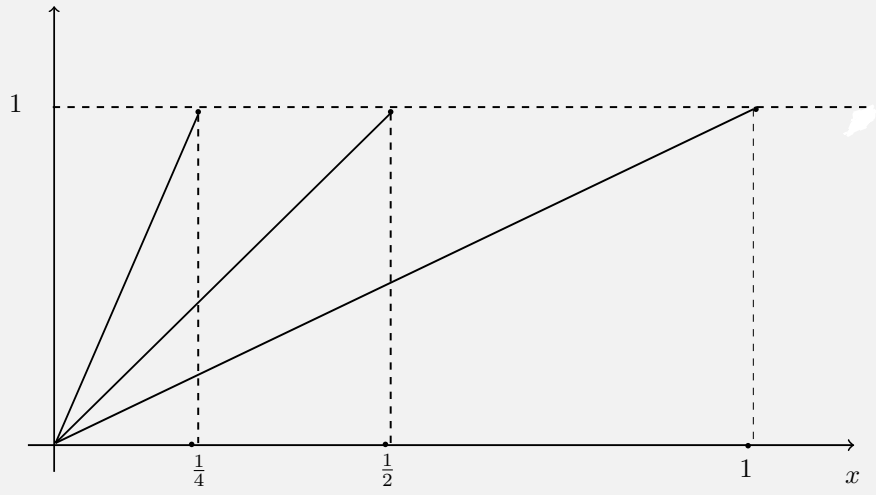

Figure 5: Convergence of the cdf of the uniform distribution to the unit step function.

The Dirac delta measure has a very simple cdf given by the unit step function,

$$F(x) = \mathbb{1}_{[0,\infty)}(x),$$

and in this example we have convergence of $F_n(x) \to F(x)$ in all points $x \in \mathbb{R}$ except for $x = 0$, since $F_n(0) = 0$ for all $n \in \mathbb{N}$.

Thm. 9 (Characterization of $\xrightarrow{d}$ using the cdf)
Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of distributions and μ be a distribution, then we have that

$$\mu_n \xrightarrow{d} \mu \iff F_{\mu_n}(x) \xrightarrow{n\to\infty} F_\mu(x),$$

for all x that are points of continuity of F_μ.

Proof.
No.

□

Remark

There can also be convergence in points of discontinuity, but it is not guaranteed in general.

Example (of convergence in the points of discontinuity)
$\mu_n = \delta_{-\frac{1}{n}}$, then it is clear that in this case also $\mu_n \to \delta_0$, and continuity is guaranteed for all points $x > 0$. However, in this case the cdf is such that

$$F_{\mu_n}(0) = F_{\delta_{-\frac{1}{n}}}(0) = 1 \quad \text{for all } n \in \mathbb{N},$$

therefore $\lim_{n\to\infty} F_{\mu_n}(0) = 1$ and convergence is satisfied both in the points of continuity as well as in the point of discontinuity of F.

Let us now discuss another important function when dealing with real-valued random variables, which also allows a convenient characterization of $\xrightarrow{d}$.

Def. (Characteristic function of a distribution)
Let μ be a distribution, then we say that the *characteristic function* (CHF) of μ is the function $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined by

$$\varphi(\eta) = \int_{\mathbb{R}^d} e^{i\langle \eta, x\rangle} \mu(dx).$$

Def. (Characteristic function of a random variable)
Let X be a random variable with distribution μ on $\mathbb{R}^d$, then the *characteristic function of X* is the function $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined by

$$\varphi_X(\eta) = \varphi_{\mu_X}(\eta) = \mathbb{E}\big[e^{i\langle X, \eta\rangle}\big].$$

Remark

If $\mu \in \text{AC}$ has density f, then we can write it exactly as a Lebesgue integral and it equals to a scaled and "slowed" version of the Fourier transform,

$$\varphi(\eta) = \int_{\mathbb{R}^d} e^{i\langle \eta, x\rangle} f(x) dx.$$

Thm. 10 (Lévy, characterization of $\xrightarrow{d}$ using the CHF)
Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of distributions and μ be a distribution, then

a) $\mu_n \xrightarrow{d} \mu \implies \varphi_n(\eta) \xrightarrow{n\to\infty} \varphi(\eta)$ *for any* $\eta \in \mathbb{R}^d$.

b) $\varphi \xrightarrow{n\to\infty} \varphi$ *everywhere, with* φ *continuous in* $\eta = 0$, *then* φ *is a CHF of a distribution* μ *and* $\mu_n \xrightarrow{d} \mu$.

Remarks

CHF's have some interesting properties, most notably

1. $\varphi(0) = 1$ since $\mathbb{E}\big[e^{i\langle 0, x\rangle}\big] = \mathbb{E}\big[1\big] = 1$.
2. φ_X is continuous in $\nu = 0$, which we can check by the limiting procedure

$$\lim_{\eta\to 0} \varphi_X(\eta) \overset{?}{=} \varphi_X(0) = 1.$$

Since $e^{i\vartheta} = \cos\vartheta + i\sin\vartheta$ is always equal in norm to 1 (Euler's formula), we can apply the dominated convergence theorem

$$\lim_{\eta\to 0} \mathbb{E}\big[e^{i\langle X,\eta\rangle}\big] \overset{\text{DCT}}{=} \mathbb{E}\big[\lim_{\eta\to 0} e^{i\langle X,\eta\rangle}\big] = \mathbb{E}\big[1\big] = 1.$$

1.2 Limit theorems

Notation: If $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables, we define the partial sums and partial means by

$$S_n = X_1 + X_2 + \ldots + X_n,$$

$$M_n = S_n/n.$$

Thm. 11 (Law of large numbers)
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables in $L^1(\Omega, \mathbb{P})$ that are i.i.d with mean $\mathbb{E}\big[X_n\big] = \mu$, then

- *(Weak L.L.N.)* $M_n \xrightarrow{d} \mu$ *and therefore* $M_n \xrightarrow{P} \mu$ *since* μ *is a constant.*
- *(Strong L.L.N.)* $M_n \xrightarrow{a.s.} \mu$

Proof.
We only prove the weak form since the strong one is very difficult. However, even for the weak form we would have to prove Lévy's theorem, which is also quite difficult. We will use the following lemma for proving the weak law of large numbers:

Lemma 2 (First derivative of the CHF)
For the CHF of a random variable X we can

$$\begin{aligned}\frac{\partial \varphi_X(\eta)}{\partial \eta} &= \frac{\partial}{\partial \eta}\mathbb{E}\big[e^{i\eta X}\big] \\ &= \mathbb{E}\big[\frac{\partial}{\partial \eta}e^{i\eta X}\big] && (DCT\ since) \\ &= \mathbb{E}\big[iXe^{i\eta X}\big]\end{aligned}$$

And computing this value in $\eta = 0$, we have that

$$\frac{\partial}{\partial \eta}\varphi_X(\eta)\Big|_{\eta=0} = i\mathbb{E}\big[X\big].$$

We want to prove that the CHF of M_n converges to that of δ_μ and then use Lévy's theorem:

$$\lim_{n\to\infty}\varphi_{M_n}(\eta) \overset{?}{=} e^{i\eta\mu} = \mathbb{E}\big[e^{i\eta\mu}\big].$$

Start by explicitly writing the CHF of M_n:

$$\begin{aligned}\varphi_{M_n}(\eta) &= \mathbb{E}\big[e^{i\eta\frac{1}{n}\sum_{j=1}^n X_j}\big] \\ &= \mathbb{E}\big[\prod_{j=1}^n e^{i\frac{\eta}{n}X_j}\big] \\ &= \mathbb{E}\big[e^{i\frac{\eta}{n}X_1}\big]^n && (\text{i.i.d}) \\ &= \varphi\big(\frac{\eta}{n}\big)^n.\end{aligned}$$

Using Lemma 2 we can apply a Taylor expansion of φ_{M_n} around $\eta = 0$:

$$\begin{aligned}\varphi_{M_n}(\eta) &= \left(1 + \frac{\eta}{n}i\mu + o\big(\frac{1}{n}\big)\right)^n \\ &= \Big(1 + \frac{\eta i\mu + \overbrace{n\cdot o\big(\frac{1}{n}\big)}^{\xrightarrow{n\to\infty}0}}{n}\Big)^n \\ &= e^{i\eta\mu} && (\text{standard limit})\end{aligned}$$

□

Remark

Had we also assumed that $X_n \in L^2(\Omega, \mathbb{P})$ with $\mathbb{V}\big[X_n\big] = \sigma^2$, then this would've become a one-line proof since

$$\mathbb{P}\big(|M_n - \mu| > \varepsilon\big) \leq \frac{\mathbb{E}\big[\overbrace{|M_n - \mu|^2}^{L^2\text{converg.}}\big]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\to\infty} 0.$$

Using this, we have convergence in L^2 which implies $\xrightarrow{P}$ and $\xrightarrow{d}$. These inequalities are useful as a very basic estimate of the speed of convergence for Monte Carlo simulations and confidence regions, in order to provide error bounds. However, proper estimates are more refined and will be discussed later on.

Lecture 2: Central limit theorems

2021-10-21

One could already be satisfied with the LLN, which describes the behaviour of the empirical average M_n. However, this doesn't tell us what the distribution of M_n will look like as $n \longrightarrow \infty$.

Given the ways we saw in the examples, how does the law μ_{M_n} approach μ?

We can first compute some quantities related to M_n:

- $\mathbb{E}\big[M_n\big] = \mu$
- $\mathbb{V}\big[M_n\big] = \frac{\sigma^2}{n}$

We will try now to normalize the empirical average and see what we obtain as a result:

$$\tilde{M}_n = \frac{M_n - \mu}{\mathrm{sd}(M_n)} = \frac{\sqrt{n}(M_n - \mu)}{\sigma}.$$

Thm. 12 (Central limit theorem)
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d r.v. in $L^2(\Omega, \mathbb{P})$, i.e. with finite variance, then we have that the normalized empirical average $\tilde{M}_n$ is such that

$$\tilde{M}_n \xrightarrow{d} \mathcal{N}_{0,1}.$$

Proof.
We use the following lemma for proving the central limit theorem:

Lemma 3 (Second derivative of the CHF)
We have that if $X \in L^2(\Omega, \mathbb{P})$,

$$\begin{aligned}\frac{\partial^2}{\partial\eta^2}\varphi_X(\eta) &= \frac{\partial}{\partial\eta}\mathbb{E}\big[iXe^{i\eta X}\big] \\ &= -\mathbb{E}\big[X^2 e^{i\eta X}\big] \qquad\qquad DCT \text{ if } \mathbb{E}[X^2] < \infty\end{aligned}$$

And by computing the derivative in $\eta = 0$,

$$\frac{\partial^2}{\partial\eta^2}\varphi_X(\eta)\Big|_{\eta=0} = -\mathbb{E}\big[X^2\big].$$

Consider $\mu = 0, \sigma^2 = 1$ which is not restrictive by the properties of the normal distribution.

$$\frac{M_n - \mu}{\sigma} = \frac{\frac{1}{n}\sum_{j=1}^n X_j - \mu}{\sigma} = \frac{1}{n}\sum_{j=1}^n \underbrace{\left(\frac{X_j - \mu}{\sigma}\right)}_{Z_j},$$

and the Z_j are such that $\mathbb{E}\big[Z_j\big] = 0, \mathbb{V}\big[Z_j\big] = 1$.

Now, the CHF of $\tilde{M}_n = S_n/\sqrt{n}$ can be written as

$$\begin{aligned}
\varphi_{\tilde{M}_n}(\eta) &= \varphi_{\frac{S_n}{\sqrt{n}}}(\eta) \\
&= \mathbb{E}\big[e^{i\eta \sum_{j=1}^{X_j}/\sqrt{n}}\big] \\
&= \mathbb{E}\big[e^{i\eta X_j/\sqrt{n}}\big]^n && \text{(i.i.d)} \\
&= \varphi_{X_1}(\eta/\sqrt{n})^n \\
&= \left(1 + \frac{1}{2}\frac{\eta^2}{n}\cdot(-1) + o\Big(\frac{1}{n}\Big)\right) && \text{(Taylor + Lemma 3)} \\
&= \left(1 + \frac{-\frac{1}{2}\eta + n\cdot o\big(\frac{1}{n}\big)}{n}\right)^n \\
&\xrightarrow{n\to\infty} e^{-\frac{\eta^2}{2}},
\end{aligned}$$

which is the characteristic function of a $\mathcal{N}_{0,1}$ random variable. The second-order expansion of φ_{X_1} does not contain the first term since the $\mathbb{E}\big[Z_j\big] = 0$, and we use Lemma 3 for the variance.

□

Remark

We can think of the CLT as telling us that for large enough n,

$$\frac{\sqrt{n}(M_n - \mu)}{\sigma} \sim \mathcal{N}_{0,1} \implies M_n \sim \mathcal{N}_{\mu,\frac{\sigma^2}{n}} \xrightarrow{d} \delta_\mu.$$

We had already computed the expected value and variance,.r the CLT also tells us the shape of the distribution. Moreover, since $S_n = n \cdot M_n$ we also know that the partial summations behave as a normal distribution,

$$S_n \sim \mathcal{N}_{n\mu, n\sigma^2},$$

which however *does not* weakly converge to any probability distribution.

Example (Bernoulli game)

We consider a Bernoulli sequence of random variables: let $(E_n)_{n\in\mathbb{N}}$ be a sequence of independent events, such that $\mathbb{P}\big(E_n\big) = p$ for all n. Set $X_n := \mathbb{1}_{E_n}$ and consider the sequence of partial sums $S_n = \sum_{j=1}^n X_j \sim \text{Binom}(n,p)$.

Since $\mu = \mathbb{E}\big[X_n\big] = p$ and $\sigma^2 = \mathbb{V}\big[X_n\big] = p(1-p)$, the CLT tells us that the empirical average is such that

$$\frac{\sqrt{n}(M_n - p)}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}_{0,1},$$

and therefore $S_n \xrightarrow{d} \mathcal{N}_{np,np(1-p)}$, which is called the De Moivre-Laplace approximation.

Example

Let $(Y_n)_{n\in\mathbb{N}}$ be a random sample of a random variable X, which means $Y_n \overset{\text{i.i.d}}{\sim} X$. We fix a real number $x \in \mathbb{R}$ and we consider the *empirical cumulative distribution function* of X,

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(Y_i).$$

Intuitively we expect that $F_n \xrightarrow{n\to\infty} F_X$, which is actually a consequence of the CLT. By defining

$$X_j = \mathbb{1}_{(-\infty,x]}(Y_j),$$

then we find that

- X_j are independent (transformation of i.i.d r.v.)
- $\mathbb{E}\big[X_j\big] = \mathbb{P}\big(Y_j \le x\big) = \mathbb{P}\big(X \le x\big) = F(x)$.

Therefore, we have that

i. LLN $\implies F_n(x) \xrightarrow{\text{a.s.}} F_X(x)$

ii. CLT $\implies \sqrt{n}\big(F_n(x) - F_X(x)\big) \xrightarrow{d} \mathcal{N}_{0,F_X(x)(1-F_X(x))}$.

However, we can also prove a convergence result which is stronger than the pointwise convergence.

Thm. 13 (Glivenko-Cantelli)

With the assumptions defined above, the empirical cdf of X is such that

$$\sup_x \|F_n(x) - F_X(x)\| \xrightarrow{a.s.} 0.$$

Proof.

No.

□

Unfortunately, with the above theorem we don't have an estimate for the number of observations needed for an asymptotical normal behaviour. However we can state the following result, which holds for *any* random variable X:

Thm. 14 (Berry-Essen)

If X_n is such that $\mathbb{E}\big[|X_n|^3\big] < \infty$, then if $\Phi(\cdot)$ is the cdf of a $\mathcal{N}_{0,1}$ random variable we have that

$$\sup_x |F_{\tilde{M}_n}(x) - \Phi(x)| \le c\frac{\mathbb{E}\big[|X|^3\big]}{\sigma^3\sqrt{n}},$$

with $c \approx 0.79\ldots$

Proof.
No.

□

Remark

The result holds for all possible choice of distributions, and although this convergence can be considered slow – $o(n^{-1/2})$ – we usually observe a faster convergence behaviour when using common distributions.

Example (Counterexample when $\mathbb{E}(X_n)$ is not defined)
Let $\mu_{X_n}(dx) = \frac{1}{\pi} \cdot \frac{1}{1+x^2} dx$. If we were in a convergence situation, then we would expect $\mu_{M_n} \to \delta_0$. However, this random variable is such that

$$\begin{aligned}
\varphi_{M_n}(\eta) &\overset{\text{iid}}{=} \varphi_{X_1}\left(\frac{\eta}{n}\right)^n \\
&= e^{-\left|\frac{\eta}{n}\right| \cdot n} && \text{(CHF of Cauchy distrib.)} \\
&= e^{-|\eta|} \\
&= \varphi_{X_1}(\eta).
\end{aligned}$$

Therefore, we see that $M_n \sim X_1$ for all n and thus it does not converge to 0. This is a consequence of the fact that X does not have a finite integral, $\mathbb{E}\big[|X_1|\big] = +\infty$.

We now state some useful generalizations of the central limit theorem, which extend its applicability to the non-identically distributed case.

Thm. 15 (Lyapunov's CLT)
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of r.v. such that

i. $\mu_n = \mathbb{E}\big[X_n\big]$, $\sigma_n^2 = \mathbb{V}\big[X_n\big] < \infty$

ii. X_n *are independent*

iii. There exists $\delta > 0$ *such that*

$$\lim_{n\to\infty} \frac{1}{\vartheta^{2+\delta}} \sum_{j=1}^{n} \mathbb{E}\big[|X_j - \mu_j|^{2+\delta}\big] = 0,$$

where $\vartheta_n^2 = \sum_{j=1}^n \sigma_j^2$.

Then, we have that

$$\frac{1}{\vartheta_n} \sum_{j=1}^{n} (X_j - \mu_j) \xrightarrow{d} \mathcal{N}_{0,1}.$$

Proof.
No.

□

Thm. 16 (Lindeberg's CLT)

Same as Lyapunov's CLT but with the third condition replaced by

iii′. For all $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{\vartheta_n^2} \sum_{j=1}^{n} \mathbb{E}\Big[(X_j - \mu_j)^2 \mathbb{1}_{[\varepsilon\vartheta_n,\infty)}\big(|X_j - \mu_j|\big)\Big] = 0$$

Proof.
No.

□

Exercises

1. Prove that Lindeberg's CLT $\implies$ Lyapunov's CLT.
2. Starting from p. 176 of Gut (2009): Ex. 2, 19, 21, 24, 32.

Lecture 3: Simulations and conditional probability

2021-10-28

3.1 Monte Carlo simulation

In this lecture we start by considering some applications of the convergence theorems we discussed earlier, in particular under the context of *Monte Carlo simulation*.

Consider a sequence of i.i.d random variables $X_1, X_2, \ldots, X_n, \ldots$ of a given r.v. X. For the sake of simplicity we will assume that $X \in L^2(\Omega, \mathbb{P})$, i.e.

$$\begin{cases} \mathbb{E}[X] = \mu < \infty \\ \mathbb{V}[X] = \sigma^2 < \infty \end{cases}$$

The goal of Monte Carlo simulation is to use the observed sample to approximate the expected value μ using the LLN and/or CLT. In particular, we will use the fact that by the LLN,

$$M_n = \frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow{\text{a.s.}} \mu.$$

Remark

i. If we consider f measurable and such that $f(X) \in L^2(\Omega, \mathbb{P})$, then the transformed sequence $f(X_1), f(X_2), \ldots, f(X_n)$ is a sample from $f(X)$. Therefore,

$$M_n^{(f)} = \frac{1}{n}\sum_{j=1}^{n} f(X_j) \xrightarrow{\text{a.s.}} f(\mu).$$

An interesting question to pose is the following one:

What is a good choice of n in order to obtain a good accuracy for the simulation?

We will try to answer this question by considering two approaches. Firstly, using the fact that $X \in L^2(\Omega, \mathbb{P})$, we can apply Chebyshev's inequality (corollary 1) and assert that for any fixed tolerance $\varepsilon > 0$:

$$\mathbb{P}\big(|M_n - \mu| \geq \varepsilon\big) \leq \frac{\sigma^2}{n\varepsilon^2} \implies \mathbb{P}\big(|M_n - \mu| < \varepsilon\big) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

We then find the minimum number of observations $\bar{n} \in \mathbb{N}$ such that, for some specified probability p, we have $\mathbb{P}\big(|M_n - \mu| < \varepsilon\big) \geq p$ for all $n \geq \bar{n}$:

$$1 - \frac{\sigma^2}{n\varepsilon^2} \geq p \iff n \geq \frac{\sigma^2}{\varepsilon^2(1-p)} = \bar{n}. \tag{2}$$

Remark

We have that the limit $\bar{n} = \bar{n}(\sigma^2, \varepsilon, p)$ is a function of three quantities, of which σ^2 is not known and is usually estimated either from the sampled data or from previous simulations. Moreover, from

Equation (2) we notice that the minimum number of samples is such that

$$\bar{n}(\sigma^2, \varepsilon, p) \longrightarrow +\infty \quad \text{if either} \quad \begin{cases} \sigma^2 & \longrightarrow \infty \\ \varepsilon & \longrightarrow 0^+ \\ p & \longrightarrow 1^- \end{cases}$$

The convergence however is quite slow and can be refined in terms of p by using the Central Limit Theorem. If n is large enough, we know by the CLT that

$$M_n - \mu \sim \mathcal{N}_{0, \frac{\sigma^2}{n}},$$

therefore we compute the approximate coverage probability

$$\mathbb{P}\big(|M_n - \mu| < \varepsilon\big) \overset{n \gg 1}{\approx} \mathbb{P}\Big(\Big| \underbrace{\frac{(M_n - \mu)\sqrt{n}}{\sigma}}_{\xrightarrow{d} \mathcal{N}_{0,1}} \Big| < \frac{\sqrt{n}\varepsilon}{\sigma}\Big) \overset{\text{sym.}}{=} 2\left(\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) - \frac{1}{2}\right) = 2\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) - 1,$$

where the last equalities come from the symmetry of the Gaussian density function. Now we want to solve the inequality

$$\begin{aligned} 2\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) - 1 \geq p &\overset{\text{monot.}}{\Longrightarrow} \frac{\sqrt{n}\varepsilon}{\sigma} \geq \Phi^{-1}\left(\frac{1+p}{2}\right) \\ &\iff n \geq \frac{\sigma^2}{\varepsilon^2} \cdot \left[\Phi^{-1}\left(\frac{1-p}{2}\right)\right]^2. \end{aligned}$$

What we claim is that the factor is sharper than the previous result $\frac{1}{1-p}$ in Equation (2), hence it is what is used in practice when computing the confidence interval.

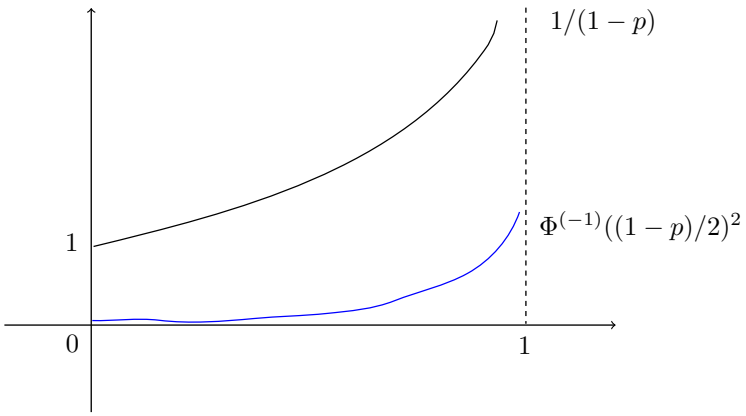

Figure 6: Sharpness of the bounds when using the two approximations.

3.2 Conditioning

Conditional probability and conditional random variables become an extremely hard topic when dealing with events that have probability zero, i.e. for continuous distribution and continuous-time stochastic processes.

Example (Dice roll)
Consider the rolling of two dice, we are interested in the outcome. We consider the sample space $\Omega = \{(i,j) : i,j = 1,\dots,6\}$. Since we have discrete events there is no problem in considering the σ-algebra given by the power set $\mathcal{F} = \mathscr{P}(\Omega)$. As for the probability measure on the measurable space we use the uniform probability $\mathbb{P} = \text{Unif}_\Omega$ on Ω:

$$\mathbb{P}\big(\{(i,j)\}\big) = \frac{1}{36}.$$

Define two variables X_1, X_2 such that X_j is the result of the j^{th} throw,

$$X_1(\omega) = X_1\big((\omega_1,\omega_2)\big) = \omega_1.$$

$$X_2(\omega) = X_2\big((\omega_1,\omega_2)\big) = \omega_2.$$

Consider the event $A =$ "the sum of the two die is smaller or equal than 6" and suppose that we win when this event occurs,

$$A = \{X_1 + X_2 \leq 6\} \implies Y = \mathbb{1}_A - \mathbb{1}_{A^c} \quad \text{is the expected win.}$$

Therefore, $\mathbb{E}[Y] = \mathbb{P}(A) - \mathbb{P}(A^c) = (15-21)/36 = -1/6$.

Assume now that the dice are instead thrown sequentially, i.e. we observe at $t = 1$ the outcome $X_1 = 5$. No one would think now that the chances of winning would be the same as before, so the observer should *update their belief* about their probabilities. Since now we can only win if the next throw is $X_2 = 1$, it's immediate to find that

$$\mathbb{P}\big(A|X_1 = 5\big) = \mathbb{P}\big(X_2 = 1\big) = \frac{1}{6}.$$

Remarks

› To calculate $\mathbb{P}(A|X_1 = 5)$ we assumed some sort of independence structure, i.e.

$$A \cap \{X_1 = 5\} = \{X_1 = 5\} \cap \{X_2 = 1\}.$$

› How do we update our belief if these random variables are not independent?

› What does it mean for two random variables to be independent in the first place?

To answer these questions we need a good definition of conditional probability, from which we will derive a notion of conditional expected value

$$\mathbb{E}[Y] \rightsquigarrow \mathbb{E}[Y|X_1 = 5] = \frac{1}{6} - \frac{5}{6}.$$

Def. (Conditional proability)
Let $A, B \in \mathcal{F}$ be events with $\mathbb{P}(B) > 0$, then we say that the *conditional probability of A given B* is given by

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

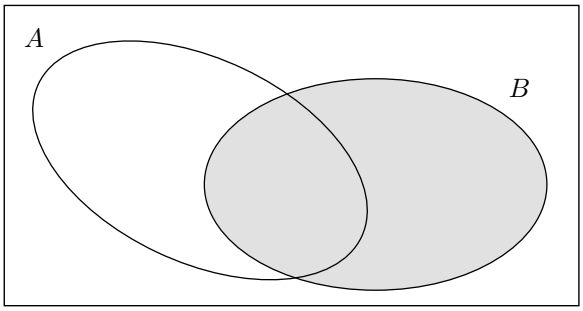

Figure 7: In some sense B takes place of the event space Ω when calculating the probability of the event $A|B$. In some sense, the admissible σ-algebra is *updated* upon observing B.

Remark

Observe that the function that maps $A \mapsto \mathbb{P}\big(A|B\big) = \mathbb{P}_{|B}(A)$ *is a new probability measure* for any $Ain\mathcal{F}$, since it satisfies the Kolmogorov axioms:

$$\mathbb{P}_{|B}(\Omega) = 1$$

$$\mathbb{P}_{|B}(A^c) = 1 - \mathbb{P}_{|B}(A)$$

$$\mathbb{P}_{|B}\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) \overset{\text{disj.}}{=} \sum_{n\in\mathbb{N}} \mathbb{P}_{|B}(A_n)$$

Def. (Conditional expectation)
For any $Y \in L^1(\Omega, \mathbb{P})$ we define the *conditional expectation* of Y given event B as the expected value w.r. to the newly-defined conditional probability measure,

$$\mathbb{E}\big[Y|B\big] = \mathbb{E}_{\mathbb{P}_{|B}}\big[Y\big]$$

Remark

We can prove that $\mathbb{E}[Y|B] = \frac{1}{\mathbb{P}(B)}\mathbb{E}\big[Y \cdot \mathbb{1}_B\big]$, which yields a convenient way of calculating the conditional probability only by using the *a priori* probability measure $\mathbb{P}$.

In general, we can define any conditional quantity that we already defined for standard random variables, such as conditional variances, etc.

If $X \in L^2(\Omega, \mathbb{P})$ is a random vector, $X : \Omega \longrightarrow \mathbb{R}^n$ then its *covariance matrix* is

$$\mathrm{Cov}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\big].$$

When conditioning w.r. to an event, we have the conditional covariance matrix

$$\mathrm{Cov}(X|B) = \mathbb{E}_{\mathbb{P}_{|B}}\big[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\big].$$

3.3 Independence

Since we have a satisfactory notion of conditional probability, by intuition we could define two events A and B to be independent if

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

Remark

If $\mathbb{P}(A), \mathbb{P}(B) > 0$ then we have that by Bayes' formula,

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

According to our intuitive definition, then we obtain that $\mathbb{P}(B) = \mathbb{P}(B|A)$. The main problem however is when $\mathbb{P}(B) = 0$, which is when the theory of probability diverges into different approaches.

3.3.1 Kolmogorov's approach

If we take for granted the definition of independence as $\mathbb{P}(A|B) = \mathbb{P}(A)$, then we obtain the following identity:

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

Therefore, we can always go back in the other direction by using this as a definition of independence and re-discovering that $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Def. (Independence of events)
Two events $A, B \in \mathcal{F}$ are said to be *independent events* if

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

Remark

If $\mathbb{P}(A) = 0$ and/or $\mathbb{P}(B) = 0$, then we have that

$$\mathbb{P}(A) \cdot \mathbb{P}(B) = 0 \overset{\text{moton.}}{=} \mathbb{P}(A \cap B).$$

Example
If $A \cap B = \emptyset$ with $\mathbb{P}(A) = 0$ then according to the definition that we gave this would mean that A and B are independent. However,

$$A \cap B = \emptyset \implies A, B \text{ ARE LOGICALLY DEPENDENT.}$$

With Kolmogorov's approach we can just say that we ignore these philosophical subtleties and work with events that are meaningful in practice.

Instead, in the approach of de Finetti we define $\mathbb{P}(A|B) = \mathbb{P}(A)$ and consider the logical coherence of the events, recovering as a *theorem* the relationship

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

With Kolmogorov's approach it is possible to define $\mathbb{P}(A|B)$ even for events that have probability zero.

Def. (Correlated events)
Two events $A, B \in \mathcal{F}$ are *positively correlated* if $\mathbb{P}(A|B) = \mathbb{P}(A)$, which can be seen to be true if and only if also $\mathbb{P}(B|A) > \mathbb{P}(B)$ (using Bayes' theorem).

Sometimes we observe one and only one event out of a set of events, i.e. we have a partition, and we would like to define how the probability measures get updated.

Example (Dice roll (cont.))
Consider all events $E_i = \{X_1 = i\}$, $i = 1, \dots, 6$, then the family of events

$$\mathcal{E} = (E_i)_{i=1,\dots,6}$$

is a partition and we can define the family of conditional measures given the partition of events whose conditional value is still unknown to us

$$\mathbb{P}(A|\mathcal{E}) = \sum_{i=1}^{6} \underbrace{\mathbb{P}(A|E_i)}_{\text{numbers}} \cdot \underbrace{\mathbb{1}_{E_i}}_{\text{r.v.}}$$

which is a *random measure*. Therefore, we can compute the conditional expected value given the partition $\mathcal{E}$

$$\mathbb{E}[Y|\mathcal{E}] = \sum_{i=1}^{6} \underbrace{\mathbb{E}[Y|E_i]}_{\text{numbers}} \cdot \underbrace{\mathbb{1}_{E_i}}_{\text{r.v.}}.$$

This type of structure is useful to model stochastic processes that evolve over time as the new partitions are observable and (eventually) observed.

Def. (Conditional probability w.r. to a partition)
Let $\mathcal{E}$ be a countable partition of events with positive probability,

- $\mathcal{E} = (E_n)_{n\in\mathbb{N}}, \quad \mathbb{P}(E_n) > 0$ for all $n \in \mathbb{N}$.
- $E_n \cap E_m = \emptyset$ for all $n \neq m$.
- $\bigcup_{n\in\mathbb{N}} E_n = \Omega$.

Given $A \in \mathcal{F}$, we define the *conditional probability w.r. to the partition* $\mathcal{E}$ as the random measure given by

$$\mathbb{P}(A|\mathcal{E}) = \sum_{n\in\mathbb{N}} \mathbb{P}(A|E_n) \cdot \mathbb{1}_{E_n}$$

Remark

Consider the function $A \mapsto \mathbb{P}(A|\mathcal{E}) :=$, then this is a *random probability measure*, i.e. by letting A vary over all possible events we have a function

$$\mathbb{P}_{|\mathcal{E}} : \mathcal{F} \longrightarrow [0,1].$$

Def. (Conditional expectation w.r. to a partition)
For any $Y \in L^1(\Omega, \mathbb{P})$ we can define the *conditional expectation given the partition* as the expected value under the random probability measure $\mathbb{P}_{|\mathcal{E}}$,

$$\mathbb{E}\big[Y|\mathcal{E}\big] = \mathbb{E}_{\mathbb{P}_{|\mathcal{E}}}\big[Y\big] = \sum_{n\in\mathbb{N}} \mathbb{E}\big[Y|E_n\big] \cdot \mathbb{1}_{E_n}$$

This could be the end of the story, unless we also want to consider *a)* uncountable partitions and *b)* events with zero probabilities, which is the case for absolutely continuous probability measures and continuous-time stochastic processes.

Example (Dice rolls (cont. ii))
This time we consider two *continuous dice*, where the probability space is now $\Omega = [0,6] \times [0,6]$, $\mathcal{F} = \mathcal{B}$, $\mathbb{P} = \text{Unif}_\Omega = \text{Unif}_{[0,6]} \otimes \text{Unif}_{[0,6]}$. We consider the same variables,

$$X_1(\omega) = \omega_1$$

$$X_2(\omega) = \omega_2$$

$$A = \{X_1 + X_2 \leq 6\}$$

$$Y = \mathbb{1}_A - \mathbb{1}_{A^c}$$

Since we have a uniform distribution, $\mathbb{P}(A) = \frac{1}{2}$ and $\mathbb{E}[Y] = \frac{1}{2} - \frac{1}{2} = 0$. Let us now assume that we observe the event $\{X_1 = 5\}$, again we have the intuition to change our probabilities

and expected value to

$$\mathbb{P}(A|\underbrace{X_1 = 5}_{\mathbb{P}(\cdot)=0}) = \mathbb{P}(X_2 \leq 1) = \frac{1}{6}.$$

$$\mathbb{E}\big[Y|X_1 = 5\big] = \frac{1}{6} - \frac{5}{6} = -\frac{2}{3}.$$

However both these quantities and the notion of independence are not defined by means of the previous definitions, since the conditioning event has probability zero.

In this case, $\mathcal{E} = (\{X_1 = k\})_{x\in[0,6]}$.

As it turns out, in order to obtain a formal definition of conditional probability we have to work the other way around: first by defining a good notion of $\mathbb{E}[Y|\mathcal{E}]$ and subsequently deduce a value for $\mathbb{P}(A|\mathcal{E})$.

Lecture 4: Conditional expectation and probability

2021-11-04

Last lecture we considered a countable partition $\mathscr{P} = \{E_i : \bigcup_{i\in\mathbb{N}} E_i = \Omega, \mathbb{P}(E_i) > 0\}$ and we defined the conditional probability w.r. to $\mathscr{P}$:

$$\mathbb{P}(A|\mathscr{P}) = \sum_{i\in\mathbb{N}} \mathbb{P}(A|E_i)\cdot \mathbb{1}_{E_i}.$$

We discuss now the process of conditioning w.r. to more general uncountable partitions and start this from the following observation.

Observation Consider a random variable $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, then we have that the random variable defined by the expected value w.r. to the countable partition $\mathscr{P}$,

$$\mathbb{E}[X|\mathscr{P}] = \sum_{i\in\mathbb{N}} \mathbb{E}[X|E_i]\cdot \mathbb{1}_{E_i},$$

satisfies the following properties:

i. This r.v. is $\sigma(\mathscr{P})$-measurable (i.e. it is observable) since $\mathscr{P}$ is a countable partition and therefore we simply have that

$$\sigma(\mathscr{P}) = \{\text{all possible unions of elements of } \mathscr{P}\}.$$

ii. For any $A \in \sigma(\mathscr{P})$, we have that

$$\mathbb{E}[X|A] = \mathbb{E}\big[\underbrace{\mathbb{E}[X|\mathscr{P}]}_{\text{r.v.}}|A\big],$$

Proof.

$A \in \sigma(\mathscr{P}) \iff A = \bigcup_{j\in J} E_j$ with J countable, therefore we can write the following chain of equations

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}[X|\mathscr{P}]\big|A\big] &= \frac{1}{\mathbb{P}(A)}\int_A \mathbb{E}[X|\mathscr{P}]\,\mathrm{d}\mathbb{P} \\
&= \frac{1}{\mathbb{P}(A)}\mathbb{E}\big[\sum_{i\in\mathbb{N}} \mathbb{E}[X|E_i]\cdot \overbrace{\mathbb{1}_{E_i}\cdot\mathbb{1}_A}^{\mathbb{1}_{E_i\cap A}}\big] \qquad (A \text{ is union of } E_j's) \\
&= \frac{1}{\mathbb{P}(A)}\mathbb{E}\big[\sum_{j\in J} \overbrace{\mathbb{E}[X|E_j]}^{\text{constant}}\cdot\mathbb{1}_{E_j}\big] \\
&= \frac{1}{\mathbb{P}(A)}\cdot\sum_{j\in J}\frac{1}{\cancel{\mathbb{P}(E_j)}}\cdot\mathbb{E}[X|E_j]\cdot\cancel{\mathbb{P}(E_j)} \\
&= \frac{1}{\mathbb{P}(A)}\mathbb{E}\Big[\sum_{j\in J} X\cdot\mathbb{1}_{E_j}\Big] \\
&= \mathbb{E}[X|A].
\end{aligned}$$

□

4.1 General case

We start from defining the conditional expectation of X and work our way up to the definition of conditional probability. Let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra of events.

Def. (Version of the conditional expectation of X given $\mathcal{G}$)
Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, we say that a random variable Z is a *version of the conditional expectation of X given $\mathcal{G}$* if Z satisfies the following properties:

i. Z is $\mathcal{G}$-measurable

ii. For any $A \in \mathcal{G}$ such that $P(A) > 0$,

$$\mathbb{E}[X|A] = \mathbb{E}[Z|A] \iff \frac{1}{\mathbb{P}(A)}\mathbb{E}[X \cdot \mathbb{1}_A] = \frac{1}{\mathbb{P}(A)}\mathbb{E}[Z \cdot \mathbb{1}_A]$$

We denote by $\mathbb{E}[X|\mathcal{G}]$ the set of all such r.v.'s.

Remark

› Sometimes Z is unique, but in general there might be equivalent random variables up to a set of measure zero, therefore $\mathbb{E}[X|A]$ defines an ***equivalence class***.

› Given $Z \in \mathbb{E}[X|\mathcal{G}]$ and $Z' \overset{\text{a.s.}}{=} Z$ it is not sufficient to guarantee that $Z' \in \mathbb{E}[X|\mathcal{G}]$ since Z' might not be measurable w.r. to $\mathcal{G}$. For the simplest example of such Z', consider if $\mathcal{G}^c$ is a set of measure zero and

$$Z' = \begin{cases} Z & \text{if } \omega \in \mathcal{G} \\ 1 & \text{if } \omega \in \mathcal{G}^c \end{cases}$$

Clearly, Z' is not measurable w.r. to $\mathcal{G}$ since $Z' = 1$ if $\omega \in \mathcal{G}^c$. However, since $\mathcal{G}^c$ has measure zero, we also have $Z \overset{\text{a.s.}}{=} Z'$.

Prop. 2 (Almost sure equality of versions)
If $Z, Z' \in \mathbb{E}[X|\mathcal{G}]$, then $Z' \overset{a.s.}{=} Z$.

Proof.
Consider for simplicity the case $X \in \mathbb{R}$, then we are going to show that for any choice of r.v.'s X, X' and $Z \in \mathbb{E}[X|\mathcal{G}]$ and $Z' \in \mathbb{E}[X'|\mathcal{G}]$, we have

$$X \overset{\text{a.s.}}{\leq} X' \implies Z \overset{\text{a.s.}}{\leq} Z'.$$

Then we will choose $X' = X$ and since the reverse holds because of symmetry we obtain a double inequality which implies strict equality.

By contradiction, assume instead that $\mathbb{P}(Z > Z') > 0$, which is the same of saying that $Z \not\overset{\text{a.s.}}{\leq} Z'$. Then, we would have that

$$\begin{aligned} 0 \overset{!!}{<} \mathbb{E}[(Z - Z') \cdot \mathbb{1}_{Z>Z'}] &= \mathbb{E}[Z \cdot \mathbb{1}_{Z>Z'}] - \mathbb{E}[Z' \cdot \mathbb{1}_{Z>Z'}] && (\{Z > Z'\} \in \mathcal{G}) \\ &= \mathbb{E}[X \cdot \mathbb{1}_{Z>Z'}] - \mathbb{E}[X' \cdot \mathbb{1}_{Z>Z'}] && (\text{def}) \\ &= \mathbb{E}[(X - X') \cdot \mathbb{1}_{Z>Z'}] \leq 0. && (X \overset{\text{a.s.}}{\leq} X' \text{ by Hp.}) \end{aligned}$$

Since we had assumed that $X \overset{\text{a.s.}}{\leq} X'$ we find a contradiction, and therefore we conclude that $\mathbb{P}(Z > Z') = 0$.

□

Corollary 2 (Conditional r.v.'s are equivalence classes)
The set of $\mathbb{E}[X|\mathcal{G}]$ is an equivalence class on the $\mathcal{G}$-measurable random variables w.r. to the "$\overset{a.s.}{=}$" operator.

Remark

- This is important to remember, since in all textbooks and articles the equivalence class $\mathbb{E}[X|\mathcal{G}]$ is treated as a single random variable. All equalities and inequalities are interpreted as valid for a chosen single representative of the equivalence class.
- Any other Z that is $\mathcal{G}$-measurable has to be constant over the events E_i of the countable partition (otherwise Z^{-1} would not be a union of E_i's). If Z was a version of $\mathbb{E}[X|\mathcal{G}]$, then it would be constantly equal to $\mathbb{E}[X|E_i]$ on E_i, which is exactly the expected value defined by cases $\implies \mathbb{E}[X|\mathscr{P}]$ is unique.

Example (Trivial conditioning)
Consider $\mathcal{G} = \sigma(X)$, then clearly (i) and (ii) are trivially satisfied by taking $Z = X$. With an *abuse* of notation, we are going to write $X = \mathbb{E}[X|\mathcal{G}]$.

The following theorem guarantees existence of $\mathbb{E}[X|\mathcal{G}]$ for a particular subset of random variables X.

Thm. 17
Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, then for any sub-$\sigma$-algebra $\mathcal{G}$ it holds that $\mathbb{E}[X|\mathcal{G}]$ is not empty.

Proof.
The proof is based on the Radon-Nikodym theorem between dominated probability measures.

□

Thm. 18 (Properties of $\mathbb{E}(X|\mathcal{G})$)
Let $X, Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G} \subset \mathcal{F}$. Then, the following properties hold:

1. *(**Linearity**) For any $\alpha, \beta \in \mathbb{R}$, $\mathbb{E}[\alpha X + \beta Y|\mathcal{G}] = \alpha\mathbb{E}[X|\mathcal{G}] + \beta\mathbb{E}[Y|\mathcal{G}]$.*
2. *(**Monotonicity**) If $X \leq Y$ a.s. then $\mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}]$.*
3. *If X is $\mathcal{G}$-measurable, then $\mathbb{E}[X|\mathcal{G}] = X$.*
4. *If $\sigma(X) \perp \mathcal{G}$ then $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.*
5. *(**Tower property**) If $\mathcal{H} \subset \mathcal{G}$, then $\mathbb{E}[X|\mathcal{H}] = \mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big]$.*
6. *If Y is $\mathcal{G}$-measurable and bounded, then $\mathbb{E}[Y \cdot X|\mathcal{G}] = Y \cdot \mathbb{E}[X|\mathcal{G}]$.*
7. *If Y is independent of X and $\mathcal{G}$, then $\mathbb{E}[Y \cdot X|\mathcal{G}] = \mathbb{E}[Y] \cdot \mathbb{E}[X|\mathcal{G}]$.*

Proof.

4 : Let $Z = \mathbb{E}[X]$, then clearly (i) is satisfied since a constant random variable is measurable w.r. to any σ-algebra. As for (ii), for any event $A \in \mathcal{G}$ such that $\mathbb{P}(A) > 0$ we can write

$$\begin{aligned}
\mathbb{E}[Z|A] &= \mathbb{E}\big[\mathbb{E}[X]|A\big] \\
&= \frac{1}{\mathbb{P}(A)} \cdot \mathbb{E}\big[\mathbb{E}[X] \cdot \mathbb{1}_A\big] \\
&= \frac{1}{\mathbb{P}(A)} \cdot \mathbb{E}[X] \cdot \mathbb{E}[\mathbb{1}_A] && (\mathbb{E}[X] \text{ is a constant}) \\
&= \frac{1}{\mathbb{P}(A)} \mathbb{E}[X \cdot \mathbb{1}_A] && (\text{indep.}) \\
&= \mathbb{E}[X|A]
\end{aligned}$$

5 : Let $Z \in \mathbb{E}[X|\mathcal{H}]$ and $Y \in \mathbb{E}[X|\mathcal{G}]$ with $\mathcal{H} \subset \mathcal{G}$, we want to prove that $Z \in \mathbb{E}[Y|\mathcal{H}]$. (i) is satisfied since $Z \in \mathbb{E}[X|\mathcal{H}]$ is $\mathcal{H}$-measurable by definition of version of $\mathbb{E}[X|\mathcal{H}]$. As for (ii), for any $A \in \mathcal{H}$ such that $\mathbb{P}(A) > 0$, we have that since $\mathcal{H} \subset \mathcal{G}$,

$$\mathbb{E}[Z|A] \overset{\text{def.}}{=} \mathbb{E}[X|A] \overset{A \in \mathcal{G}}{=} \mathbb{E}[Y|A].$$

6 : $(ii) \iff \mathbb{E}[X \cdot W] = \mathbb{E}[Z \cdot W]$ for any r.v. W that is $\mathcal{G}$-measurable and bounded.

□

Remarks

- $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$ families of events are said to be independent if $A \perp B$ for any $A \in \mathcal{A}$ and $B \in \mathcal{B}$.
- We say that a r.v. Y is independent of a r.v. X and a σ-algebra $\mathcal{G}$ if $\sigma(Y) \perp \sigma\big(\sigma(X) \cup \mathcal{G}\big)$.
- Property (5) means that reducing the information from $\mathcal{F} \to \mathcal{H}$ can be done by reducing multiple times, $\mathcal{F} \to \mathcal{G} \to \mathcal{H}$.

Example (Conditioning w.r. to a random variable)
In the particular case of $\mathcal{G} = \sigma(Y)$, then we can define the following random variable,

$$\mathbb{E}[X|Y] := \mathbb{E}[X|\sigma(Y)]$$

Lemma 4 (Doob's theorem)
Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and Y r.v. that takes values in another measurable space, in this case assume $(\mathbb{R}^n, \mathcal{B})$. Then X is $\sigma(Y)$-measurable $\iff$ there exists a measurable function $\varphi : \mathbb{R}^n \to \mathbb{R}^d$ such that

$$X = \varphi(Y),$$

which however could be not unique.

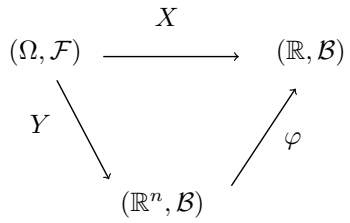

Figure 8: Schematization of Doob's theorem.

Corollary 3 (Existence of the regression function)
There exists a measurable function $\varphi : \mathbb{R}^n \to \mathbb{R}^d$ such that

$$\varphi(Y) = \mathbb{E}[X|Y],$$

*and this function is called the **regression function**.*

Proof.
We have that if the event $\{Y = y\}$ has non-negligible probability, then

$$\mathbb{E}[X|(Y=y)] \overset{\text{def}}{=} \mathbb{E}\big[\underbrace{\mathbb{E}[X|Y]}_{\text{Doob}\to\varphi(Y)} \big|(Y=y)\big] = \mathbb{E}[\varphi(Y)|(Y=y)] = \mathbb{E}[\varphi(y)] = \varphi(y).$$

□

Remark

- φ might not be unique because (i) Doob's theorem does not guarantee unicity and (ii) it belongs to the equivalence class of conditional expectations.
- φ is however almost-surely unique w.r. to the law μ_Y of Y.

Notation: The function φ is denoted as $\varphi(Y) = \mathbb{E}[X|Y = y]$, which is <u>not</u> the conditional expectation given the event $\{Y = y\}$ unless this set has positive probability.

Example

Consider the particular case in which $X = \mathbb{1}_{X\in H}$ for a Borel set $H \in \mathcal{B}$. Then, we can consider the following expected value

$$\mathbb{E}[\mathbb{1}_{X\in H}|\mathcal{G}] =: \mu_{X|\mathcal{G}}(H),$$

which we call the *conditional law of X given $\mathcal{G}$* calculated in the set H.

We can check that it is a probability measure

- If $H = \mathbb{R}^d$, we obtain $\mathbb{E}[\mathbb{1}_{X\in\mathbb{R}^d}|\mathcal{G}] = 1$.
- If $H = \emptyset$, then $\mathbb{E}[\mathbb{1}_{X\in\emptyset}|\mathcal{G}] = 0$.
- Given a sequence $(H_n)_{n\in\mathbb{N}}$ of disjoint sets, we have

$$\mathbb{E}[\mathbb{1}_{X\in\bigcup_{n\in\mathbb{N}} H_n}|\mathcal{G}] = \mathbb{E}\Big[\sum_{n\in\mathbb{N}} \mathbb{1}_{X\in H_n}|\mathcal{G}\Big] = \sum_{n\in\mathbb{N}} \mathbb{E}[\mathbb{1}_{X\in H_n}|\mathcal{G}].$$

The problem is that this holds in the almost-sure sense, i.e. everything is defined in terms of a representative of the equivalence class.

Thm. 19 (Regular conditional law)

Given $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $G \subset \mathcal{F}$ there always exist a family $\big(\mu_{X|\mathcal{G}}(\omega)\big)_{\omega\in\Omega}$ of probability measures on $\mathbb{R}^d$ such that for any Borel set H,

$$\mu_{X|\mathcal{G}}(H) = \mathbb{E}[\mathbb{1}_{X\in H}|\mathcal{G}].$$

Such family is called the ***regular version of conditional law of X given $\mathcal{G}$.***

Proof.
No.

□

Using this definition of regular conditional law, there are many results that we can compute that will give the usual known results.

Thm. 20 (Conditional expectation)

If $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ and $f(X) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, then we can write

$$\mathbb{E}[f(X)|\mathcal{G}] = \int_{\mathbb{R}^d} f(x)\mu_{X|\mathcal{G}}(dx).$$

Example (Conditional expected value)
If $f = \mathrm{id}$, then this becomes the *conditional expected value of X given $\mathcal{G}$*,

$$\mathbb{E}[X|\mathcal{G}] = \int_{\mathbb{R}^d} x\mu_{X|\mathcal{G}}(dx)$$

Thm. 21 (Expectation of the conditional measure)
Following from the tower property of the conditional expectation, we have that

$$\mu_X(H) = \mathbb{E}[\mu_{X|\mathcal{G}}(H)]$$

Example
Choosing $\mathcal{G} = \sigma(Y)$ we obtain the conditional law of X given Y as $\mu_{X|Y} := \mu_{X|\sigma(Y)}$.

Thm. 22 (Joint distribution of two random variables)
Let X, Y be r.v. with values on $\mathbb{R}^d$ and $\mathbb{R}^n$, respectively. Then, we have that for each $H \in \mathcal{B}(R^d)$ and $K \in \mathcal{B}(R^n)$,

$$\mu_{(X,Y)}(H \times K) = \mathbb{E}[\mu_{X|Y}(H) \cdot \mathbb{1}_{Y \in K}].$$

Thm. 23 (Conditional law as a function of Y)
If X, Y are r.v.'s on $\mathbb{R}^d$ and $\mathbb{R}^n$ respectively, then there exists a family $(\mu_{X|Y=y})_{y \in \mathbb{R}^n}$ of probability measures on $\mathbb{R}^d$ such that

i. For each $H \in \mathcal{B}$, the function $y \mapsto \mu_{X|Y=y}(H)$ is measurable.

ii. $(\mu_{X|Y=y})\big|_{y=Y} = \mu_{X|Y}$

Notation: Sometimes we can find this written as $\mathbb{E}[\mathbb{1}_{X \in H}|Y = y]$ and the conditional probability function coincides with this quantity if $\mathbb{P}(Y = y) > 0$.

Thm. 24 (Conditional expected value of a function)
Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that $f(X) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, then the regression function

$$\mathbb{E}[f(X)|Y = y] = \int_{\mathbb{R}^d} f(x)\mu_{X|Y=y}(dx).$$

The last thing to study is what happens when X and Y have an absolutely continuous joint probability distribution, i.e. they admit a joint density.

If (X, Y) are jointly absolutely continuous, then X and Y are also absolutely continuous and

$$\gamma_Y(y) = \int_{\mathbb{R}^d} \gamma_{(X,Y)}(x, y)\, \mathrm{d}x.$$

Thm. 25 (Conditional density of two jointly a.c. random variables)
If (X, Y) are jointly absolutely continuous, then for any $y \in \mathbb{R}^n$ such that $\gamma_Y(y) > 0$ the conditional law function $\mu_{X|Y=y}$ is absolutely continuous with density given by

$$\gamma_{X|Y=y}(x) = \frac{\gamma_{(X,Y)}(x, y)}{\gamma_Y(y)}.$$

Remark For any Borel set H, $\mu_{X|Y=y}(H) = \int_H \gamma_{X|Y=y}(\mathrm{d}x)$

Corollary 4
For two jointly absolutely continuous random variables, we have that

$$\mathbb{E}[f(X)|Y = y] = \int_{\mathbb{R}^d} f(x)\gamma_{X|Y=y}(x)\, \mathrm{d}x,$$

whereas the unconditional expected value is

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}^n} \int_{\mathbb{R}^d} f(x)\gamma_{(X|Y=y)}(x)\gamma_Y(y)\, \mathrm{d}y\, \mathrm{d}x.$$

Exercises: Given in the notes.

References

Çinlar, E. (2011). *Probability and Stochastics*. Vol. 261. Graduate Texts in Mathematics. New York.

Gut, A. (2009). *An Intermediate Course in Probability*. 2° edizione. Springer Nature.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Singapore: World Scientific Pub.

Paolella, M. S. (2007). *Intermediate Probability: A Computational Approach*. Wiley-Interscience.