3 Random elements

3.1 Measurable maps

We now study maps $X : \Omega \to \Omega'$ (often $\Omega' = \mathbb{R}^d$, $d \geq 1$, so X is a function).

Definition 3.1 (Measurability, random elements, variables, vectors, etc.)
If $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ are measurable spaces, then $X : \Omega \to \Omega'$ is *$(\mathcal{F}, \mathcal{F}')$-measurable* (or simply *measurable*) if

$$X^{-1}(\mathcal{F}') \subseteq \mathcal{F},$$

i.e. $X^{-1}(A') \in \mathcal{F}\ \forall\, A' \in \mathcal{F}'$. In this case, X is called *random element*. If $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, X is called *random variable (rv)*. If $(\Omega', \mathcal{F}') = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $d \geq 2$, X is called *random vector*, and if $d = \infty$, X is called *random sequence*.

- A rv X (term supported by William Feller) is neither random, nor a variable, it is a function (considered random in the sense that we don't know its evaluation point, the sample point $\omega \in \Omega$).
- Carefully distinguish X (the map; potential outcome) from $x = X(\omega)$ (the value, actual outcome, or realization based on the "state of nature" $\omega \in \Omega$ that happened in an experiment).

- $(\mathcal{F}, \mathcal{B}(\mathbb{R}^d))$-measurable functions $X : \Omega \to \mathbb{R}^d$, $d \geq 1$, are often simply referred to as *($\mathcal{F}$-)measurable*; notation: $X \in \mathcal{F}$. $(\mathcal{B}(\Omega), \mathcal{B}(\mathbb{R}^d))$-measurable X are called *Borel measurable* and $(\bar{\mathcal{B}}(\Omega), \mathcal{B}(\mathbb{R}^d))$-measurable X *Lebesgue measurable*.
- To allow as many maps as possible to be measurable, one typically assumes $\mathcal{F}$ to be complete ($\mathcal{F} = \bar{\mathcal{F}}$) and the definition of rvs uses the Borel σ-algebra $\mathcal{F}' = \mathcal{B}(\mathbb{R}^d)$ (not: Lebesgue σ-algebra $\bar{\mathcal{B}}(\mathbb{R}^d)$). We can then hope (see later) that $h(X)$ for a Borel measurable h is again Borel and so a rv (which it may not be if h was Lebesgue measurable).

Example 3.2 (Non-measurable $X : \Omega \to \mathbb{R}$)

1) **Trivial example:** Let $\Omega \neq \emptyset$ and $\mathcal{F} = \{\emptyset, \Omega\}$ be the trivial σ-algebra. Then any non-constant $X : \Omega \to \mathbb{R}$ is not measurable.
   Proof. Consider any $x = X(\omega)$ for some $\omega \in \Omega$. Then $\{x\} = \bigcap_{n \in \mathbb{N}} (x - \frac{1}{n}, x] \in \mathcal{B}(\mathbb{R})$, but $X^{-1}(\{x\}) \neq \emptyset$ (since $\omega \in X^{-1}(\{x\})$) and $X^{-1}(\{x\}) \neq \Omega$ (since X takes on at least two different values), so $X^{-1}(\{x\}) \notin \mathcal{F}$.
2) **Example for any $\mathcal{F}$:** If $(\Omega, \mathcal{F})$ is a measurable space and $V \subseteq \Omega : V \notin \mathcal{F}$ (e.g. $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with Vitali set V), then $X = \mathbb{1}_V$ is non-measurable.
   Proof. $X^{-1}((\frac{1}{2}, 1]) = \{\omega \in \Omega : X(\omega) = 1\} = V \notin \mathcal{F}$.

Example 3.3 (Determining measurability and $\sigma(X)$ by definition)

1) If $X = c \in \mathbb{R}$, then, $\forall\, B \in \mathcal{B}(\mathbb{R})$, we have $X^{-1}(B) = \begin{cases} \Omega, & c \in B, \\ \emptyset, & c \notin B, \end{cases}$ $\underset{\text{def.}}{\Rightarrow}$ X is a rv. Also, $\sigma(X) \underset{\text{def.}}{=} X^{-1}(\mathcal{B}(\mathbb{R})) = \{\emptyset, \Omega\}$.

2) If $X = \mathbb{1}_A$ for $A \in \mathcal{F}$, then, $\forall\, B \in \mathcal{B}(\mathbb{R})$, $X^{-1}(B) = \begin{cases} \Omega, & 0 \in B, 1 \in B, \\ A, & 0 \notin B, 1 \in B, \\ A^c, & 0 \in B, 1 \notin B, \\ \emptyset, & 0 \notin B, 1 \notin B, \end{cases}$ $\underset{\text{def.}}{\Rightarrow}$ X is a rv. Also, $\sigma(X) \underset{\text{def.}}{=} X^{-1}(\mathcal{B}(\mathbb{R})) = \{\emptyset, A, A^c, \Omega\}$. In particular, if $\Omega = \{1, \dots, 6\}$ and $\mathcal{F} = \{\emptyset, \{1,2\}, \{3,4,5,6\}, \Omega\}$ then $X = \mathbb{1}_{\{1,2\}}$ is a rv on $(\Omega, \mathcal{F})$, but $X = \mathbb{1}_{\{1,2,3\}}$ is not (e.g. $X^{-1}((1/2, 1]) = \{1,2,3\} \notin \mathcal{F}$).

3) A rv X is *simple* if, for some $n \in \mathbb{N}$ and a partition $\{A_i\}_{i=1}^n \subseteq \mathcal{F}$ of Ω,
$$X = \sum_{i=1}^n x_i \mathbb{1}_{A_i}.$$
Similarly as in 1), 2), $\sigma(X) = \{\biguplus_{i \in I} A_i : I \subseteq \{1, \dots, n\}\}$ (it contains all 2^n distinct preimages X^{-1} of Borel sets that contain any subset of $\{x_1, \dots, x_n\}$). Simple rvs are important for constructing the Lebesgue integral (see later).

Lemma 3.4 (σ-algebra generated by preimages)
Let $(\Omega, \mathcal{F})$, $(\Omega', \mathcal{F}')$ be measurable spaces and $X : \Omega \to \Omega'$. If $\mathcal{A}' \subseteq \mathcal{P}(\Omega')$, then $\sigma(X^{-1}(\mathcal{A}')) = X^{-1}(\sigma(\mathcal{A}'))$. If $\mathcal{F}' = \sigma(\mathcal{A}')$, then $\sigma(X^{-1}(\mathcal{A}')) = \sigma(X)$.

Proof.
"$\subseteq$": $\mathcal{A}' \subseteq \sigma(\mathcal{A}') \underset{\text{mon.}}{\Rightarrow} X^{-1}(\mathcal{A}') \subseteq X^{-1}(\sigma(\mathcal{A}'))$. By E. 2.9 6), $X^{-1}(\sigma(\mathcal{A}'))$ is a σ-algebra, and it contains $X^{-1}(\mathcal{A}') \underset{\text{smallest}}{\Rightarrow} \sigma(X^{-1}(\mathcal{A}')) \subseteq X^{-1}(\sigma(\mathcal{A}'))$.

"$\supseteq$": We follow the principle of good sets and first show that $\mathcal{G} := \{A' \subseteq \Omega' : X^{-1}(A') \in \sigma(X^{-1}(\mathcal{A}'))\}$ is a σ-algebra on Ω':

i) $\Omega' \in \mathcal{G}$ since $\Omega' \subseteq \Omega'$ and $X^{-1}(\Omega') = \Omega \in \sigma(X^{-1}(\mathcal{A}'))$.

ii) $A' \in \mathcal{G} \underset{\text{def. }\mathcal{G}}{\Rightarrow} A' \subseteq \Omega' : X^{-1}(A') \in \sigma(X^{-1}(\mathcal{A}')) \Rightarrow A'^c \subseteq \Omega' : X^{-1}(A'^c) \underset{\text{S. 1.2}}{=} X^{-1}(A')^c \underset{\sigma(\cdot)\ \sigma\text{-alg.}}{\in} \sigma(X^{-1}(\mathcal{A}')) \underset{\text{def. }\mathcal{G}}{\Rightarrow} A'^c \in \mathcal{G}$.

iii) $\{A'_i\}_{i \in \mathbb{N}} \subseteq \mathcal{G} \underset{\text{def. }\mathcal{G}}{\Rightarrow} A'_i \subseteq \Omega' : X^{-1}(A'_i) \in \sigma(X^{-1}(\mathcal{A}')) \Rightarrow \bigcup_{i=1}^{\infty} A'_i \subseteq \Omega' : X^{-1}(\bigcup_{i=1}^{\infty} A'_i) \underset{\text{S. 1.2}}{=} \bigcup_{i=1}^{\infty} X^{-1}(A'_i) \underset{\sigma(\cdot)\ \sigma\text{-alg.}}{\in} \sigma(X^{-1}(\mathcal{A}')) \underset{\text{def. }\mathcal{G}}{\Rightarrow} \bigcup_{i=1}^{\infty} A'_i \in \mathcal{G}$.

Now $\forall\, A' \in \mathcal{A}'$, we have $X^{-1}(A') \in \sigma(X^{-1}(\mathcal{A}'))$, so $\mathcal{A}' \subseteq \mathcal{G} \Rightarrow \sigma(\mathcal{A}') \underset{\text{smallest}}{\subseteq} \mathcal{G}$ $\underset{\text{mon.}}{\Rightarrow} X^{-1}(\sigma(\mathcal{A}')) \subseteq X^{-1}(\mathcal{G}) \underset{\text{def. }\mathcal{G}}{\subseteq} \sigma(X^{-1}(\mathcal{A}'))$.

If $\mathcal{F}' = \sigma(\mathcal{A}')$, then $\sigma(X) \underset{\text{def.}}{=} X^{-1}(\mathcal{F}') \underset{\text{ass.}}{=} X^{-1}(\sigma(\mathcal{A}')) \overset{\text{just}}{\underset{\text{shown}}{=}} \sigma(X^{-1}(\mathcal{A}'))$. □

Proposition 3.5 (Checking measurability via a generator)
Let $(\Omega, \mathcal{F})$, $(\Omega', \mathcal{F}')$ be measurable spaces, $\mathcal{A}' \subseteq \mathcal{F}' : \sigma(\mathcal{A}') = \mathcal{F}'$. Then $X : \Omega \to \Omega'$ is measurable iff $X^{-1}(\mathcal{A}') \subseteq \mathcal{F}$.

Proof.

"$\Rightarrow$": Since $\mathcal{A}' \subseteq \sigma(\mathcal{A}') \underset{\text{ass.}}{=} \mathcal{F}' \Rightarrow X^{-1}(\mathcal{A}') \underset{\text{mon.}}{\subseteq} X^{-1}(\mathcal{F}') \underset{X \text{ measurable}}{\subseteq} \mathcal{F}$

"$\Leftarrow$": $X^{-1}(\mathcal{F}') \underset{\text{ass.}}{=} X^{-1}(\sigma(\mathcal{A}')) \underset{\text{L. 3.4}}{=} \sigma(X^{-1}(\mathcal{A}')) \overset{\text{ass. } X^{-1}(\mathcal{A}') \subseteq \mathcal{F},\, \mathcal{F}\ \sigma\text{-alg.}}{\underset{\text{smallest}}{\subseteq}} \mathcal{F}$ □

Proposition 3.6 (Compositions)
Let $(\Omega, \mathcal{F})$, $(\Omega', \mathcal{F}')$, $(\Omega'', \mathcal{F}'')$ be measurable spaces and $X : \Omega \to \Omega'$, $Y : \Omega' \to \Omega''$ be measurable. Then $Y \circ X$ is $((\mathcal{F}, \mathcal{F}'')$-$)$measurable.

Proof. $\forall\, A'' \in \mathcal{F}''$,

$$(Y \circ X)^{-1}(A'') = \{\omega \in \Omega : Y(X(\omega)) \in A''\} = \{\omega \in \Omega : X(\omega) \in Y^{-1}(A'')\}$$
$$= \{\omega \in \Omega : \omega \in X^{-1}(Y^{-1}(A''))\} = X^{-1}(\underbrace{Y^{-1}(A'')}_{\underset{Y \text{ meas.}}{\in} \mathcal{F}'}) \underset{X \text{ meas.}}{\in} \mathcal{F} \quad \square$$

Proposition 3.7 (Continuous maps are measurable)
Let $(\Omega, \mathcal{T})$, $(\Omega', \mathcal{T}')$ be topological spaces and $X : \Omega \to \Omega'$ continuous. Then X is $((\mathcal{B}(\Omega), \mathcal{B}(\Omega'))$-)measurable.

Proof. $X^{-1}(\mathcal{T}') \underset{\text{cont.}}{\overset{\text{def.}}{\subseteq}} \mathcal{T} \underset{\text{def.}}{\subseteq} \sigma(\mathcal{T}) \underset{\text{def.}}{=} \mathcal{B}(\Omega) \Rightarrow \sigma(X^{-1}(\mathcal{T}')) \underset{\text{smallest}}{\subseteq} \mathcal{B}(\Omega) \underset{\text{L. 3.4}}{\Rightarrow} X^{-1}(\sigma(\mathcal{T}')) \subseteq \mathcal{B}(\Omega) \underset{\sigma(\mathcal{T}') = \mathcal{B}(\Omega')}{\Rightarrow} X^{-1}(\mathcal{B}(\Omega')) \subseteq \mathcal{B}(\Omega)$ ✓ □

Proposition 3.8 (Monotone functions are measurable)
Monotone $h : \mathbb{R} \to \mathbb{R}$ are $((\mathcal{B}(\mathbb{R}), \mathcal{B}(\mathbb{R}))$-)measurable.

Proof. Wlog, consider h to be increasing; the proof for decreasing h works similarly.

- By R. 2.24 1), $\{[t, \infty) : t \in \mathbb{R}\}$ generates $\mathcal{B}(\mathbb{R}) \underset{\text{P. 3.5}}{\Rightarrow}$ it suffices to check that $h^{-1}([t, \infty)) \in \mathcal{B}(\mathbb{R})\ \forall\, t \in \mathbb{R}$.
- For $t \in \mathbb{R}$, consider $A_t := h^{-1}([t, \infty)) = \{x \in \mathbb{R} : h(x) \in [t, \infty)\}$.
- For every $x_2 > x_1 \in A_t$, we have $h(x_2) \underset{\text{mon.}}{\geq} h(x_1) \underset{x_1 \in A_t}{\geq} t$, so $x_2 \in A_t$. Therefore, A_t must be an interval of the form $[\inf A_t, \infty)$ if $t \in \operatorname{ran}(h)$ or $(\inf A_t, \infty)$ if $h(\inf A_t) < t$, so a set in $\mathcal{B}(\mathbb{R})$. ✓ □

Example 3.9 (Lebesgue set that is not a Borel set)

- Consider $\mathcal{V} := F^{-1}(V)$, where F^{-1} is the quantile function of the Cantor df F from E. 2.51 and V is the Vitali set of T. 2.1.
- As a countable intersection of closed sets, the Cantor set $\mathcal{C}$ is a Borel null set.
- $\mathcal{V} \underset{\text{def.}}{=} F^{-1}(V) \subseteq F^{-1}([0,1]) \underset{F\text{ cont.}}{=} \mathcal{C} \Rightarrow$ As a subset of the Borel null set $\mathcal{C}$, $\mathcal{V}$ is Lebesgue measurable.
- F continuous $\Rightarrow F^{-1} \uparrow \underset{\text{P. 3.8}}{\Rightarrow} F^{-1}$ is Borel measurable.
- Suppose $\mathcal{V}$ was Borel measurable, then its preimage $(F^{-1})^{-1}(\mathcal{V})$ under the Borel measurable function F^{-1} is a Borel measurable set. However,

$$(F^{-1})^{-1}(\mathcal{V}) \underset{\text{def.}}{=} \{y : F^{-1}(y) \in \mathcal{V}\} \underset{\text{def.}}{=} \{y : F^{-1}(y) \in F^{-1}(V)\} = V \notin \mathcal{B}(\mathbb{R}) \ ⚡.$$

- So $\mathcal{V}$ is Lebesgue measurable, but cannot be Borel measurable.

If f is Borel measurable, then $f \circ g$ is Lebesgue or Borel measurable whenever g is. But if $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are both Lebesgue measurable, then $f \circ g$ may not be. If $A \in \mathcal{B}(\mathbb{R})$, then $f^{-1}(A) \in \bar{\mathcal{B}}(\mathbb{R})$, but unless also $f^{-1}(A) \in \mathcal{B}(\mathbb{R})$, there is no guarantee that $g^{-1}(f^{-1}(A)) \in \bar{\mathcal{B}}(\mathbb{R})$; see E. 3.9 above.

Proposition 3.10 (Random vectors are vectors of rvs)
Let $(\Omega, \mathcal{F})$ be a measurable space and $X : \Omega \to \mathbb{R}^d$. Then X is a random vector iff $X = (X_1, \dots, X_d)$ for rvs $X_1, \dots, X_d$.

Proof.

"$\Rightarrow$": Let $X : \Omega \to \mathbb{R}^d$ be measurable. The projection $\pi_j : \mathbb{R}^d \to \mathbb{R}$ onto the jth coordinate with $\pi_j(x_1, \dots, x_d) = x_j$ is continuous since $\forall\, \varepsilon > 0$ $\exists\, \delta > 0 : |\pi_j(\boldsymbol{x}) - \pi_j(\boldsymbol{y})| = |x_j - y_j| < \varepsilon\ \forall\, \|\boldsymbol{x} - \boldsymbol{y}\| < \delta$ (e.g. take $\delta = \varepsilon$). $\underset{\text{P. 3.7}}{\Rightarrow}$ π_j is measurable $\underset{\text{P. 3.6}}{\Rightarrow}$ $X_j := \pi_j \circ X : \Omega \to \mathbb{R}$ is measurable, so a rv.

"$\Leftarrow$": For all $-\infty < \boldsymbol{a} < \boldsymbol{b} < \infty$, we have $X^{-1}((\boldsymbol{a}, \boldsymbol{b}]) = \{\omega \in \Omega : X(\omega) \in (\boldsymbol{a}, \boldsymbol{b}]\} = \{\omega \in \Omega : X_j(\omega) \in (a_j, b_j]\ \forall\, j\} = \bigcap_{j=1}^d X_j^{-1}((a_j, b_j]) \underset{X_j \in \mathcal{F}}{\in} \mathcal{F}$. As $\sigma(\{(\boldsymbol{a}, \boldsymbol{b}] : \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d, \boldsymbol{a} < \boldsymbol{b}\}) \underset{\text{R. 2.24 1)}}{=} \mathcal{B}(\mathbb{R}^d)$, P. 3.5 implies that $X \in \mathcal{F}$. $\square$

- Because of P. 3.10, we typically use bold notation for (random) vectors, so $\boldsymbol{X} := (X_1, \dots, X_d)$. If necessary (e.g. multiplications with matrices), we interpret $\boldsymbol{X}$ as a $(d, 1)$-dimensional (random) matrix.
- Standard notation for the *sample mean* is $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$ and for the *sample maxima* $M_n := \max\{X_1, \dots, X_n\}$.

Corollary 3.11 (Sums, products, minima, maxima, etc.)
Let $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^k$ be measurable.

1) If $\boldsymbol{X}$ is a random vector, then $\boldsymbol{Y} = \boldsymbol{h}(\boldsymbol{X})$ is a random vector.
2) Sums, products, minima and maxima of rvs are rvs.

Proof.

1) P. 3.6 (composition).
2) Follows from Part 1) with $k = 1$ using the fact that sums, products, minima and maxima are continuous functions and thus measurable by P. 3.7. □

Lemma 3.12 (Measurability of quantities involving sequences of rvs)
Let $(X_i)_{i \in \mathbb{N}}$ be a sequence of rvs on a measurable space $(\Omega, \mathcal{F})$.

1) $\inf_{k \geq n} X_k$, $\sup_{k \geq n} X_k$, $\liminf_{n \to \infty} X_n$, $\limsup_{n \to \infty} X_n$ are rvs.
2) If $\lim_{n \to \infty} X_n(\omega)$ exists $\forall\, \omega \in \Omega$ (surely), then $\lim_{n \to \infty} X_n$ is a rv.
3) $\{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega)$ exists$\} \in \mathcal{F}$, so is a measurable set.

Proof. All expressions are functions $\Omega \to \mathbb{R}$. We have left to show measurability:

1) i) $(\inf_{k\ge n} X_k)^{-1}((-\infty,x]) = \{\omega\in\Omega : \inf_{k\ge n} X_k(\omega)\le x\} \overset{\text{for at least}}{\underset{\text{one } k}{=}} \bigcup_{k\ge n}\{\omega\in\Omega : X_k(\omega)\le x\} \underset{\text{def.}}{=} \bigcup_{k\ge n} X_k^{-1}((-\infty,x]) \underset{X_k^{-1}((-\infty,x])\,\in\,\mathcal{F}}{\in} \mathcal{F}\ \ \forall\, x\in\mathbb{R} \underset{\text{P. 3.5}}{\Rightarrow}$ ✓

ii) $(\sup_{k\ge n} X_k)^{-1}([x,\infty)) = \{\omega\in\Omega : \sup_{k\ge n} X_k(\omega)\ge x\} \overset{\text{for at least}}{\underset{\text{one } k}{=}} \bigcup_{k\ge n}\{\omega\in\Omega : X_k(\omega)\ge x\} \underset{\text{def.}}{=} \bigcup_{k\ge n} X_k^{-1}([x,\infty)) \underset{X_k^{-1}([x,\infty))\,\in\,\mathcal{F}}{\in} \mathcal{F}\ \ \forall\, x\in\mathbb{R} \underset{\text{P. 3.5}}{\Rightarrow}$ ✓

iii) $\liminf_{n\to\infty} X_n \underset{\text{def.}}{=} \sup_{n\ge 1}(\inf_{k\ge n} X_k) \overset{\text{i), ii)}}{\underset{\text{P. 3.6}}{\Rightarrow}}$ ✓ and similarly $\limsup_{n\to\infty} X_n \underset{\text{def.}}{=} \inf_{n\ge 1}(\sup_{k\ge n} X_k) \overset{\text{i), ii)}}{\underset{\text{P. 3.6}}{\Rightarrow}}$ ✓

2) If $\lim_{n\to\infty} X_n$ exists surely, then $\lim_{n\to\infty} X_n(\omega) = \limsup_{n\to\infty} X_n(\omega)\ \forall\,\omega \underset{\text{1) iii)}}{\Rightarrow}$ ✓

3) We have

$$\{\omega\in\Omega : \lim_{n\to\infty} X_n(\omega) \text{ exists}\}^c = \{\omega\in\Omega : \liminf_{n\to\infty} X_n(\omega) < \limsup_{n\to\infty} X_n(\omega)\}$$
$$= \bigcup_{q\in\mathbb{Q}} \{\omega\in\Omega : \liminf_{n\to\infty} X_n(\omega) \le q < \limsup_{n\to\infty} X_n(\omega)\}$$
$$= \bigcup_{q\in\mathbb{Q}} (\{\omega\in\Omega : \liminf_{n\to\infty} X_n(\omega)\le q\} \cap \{\omega\in\Omega : \limsup_{n\to\infty} X_n(\omega)\le q\}^c) \underset{1)}{\in} \mathcal{F} \quad \square$$

Lemma 3.13 (Measurability of a.e. limits under completeness)
Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $X, Y, (X_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}$. Then μ is complete iff any of the following implications hold:

1) If X is measurable and $X \overset{\text{a.e.}}{=} Y$, then Y is measurable.

2) If X_n is measurable $\forall\, n \in \mathbb{N}$ and $X_n \underset{n\to\infty}{\overset{\text{a.e.}}{\to}} X$, then X is measurable.

Proof.

1) "$\Rightarrow$": Let $N \in \mathcal{F} : \mu(N) = 0$ and $X = Y$ on N^c. For any measurable set A, we thus have $Y^{-1}(A) = (Y^{-1}(A) \cap N) \uplus (Y^{-1}(A) \cap N^c)$. Then $Y^{-1}(A) \cap N$ is measurable as a subset of a null set and $\mathcal{F}$ is complete. And on N^c, we have $Y^{-1}(A) \cap N^c = X^{-1}(A) \cap N^c$, which is measurable (as, by assumption, X is measurable, so $X^{-1}(A) \in \mathcal{F}$ and thus $X^{-1}(A) \cap N^c \in \mathcal{F}$). Therefore, $Y^{-1}(A) \cap N$ and $Y^{-1}(A) \cap N^c$ are measurable, and so is $Y^{-1}(A) = (Y^{-1}(A) \cap N) \uplus (Y^{-1}(A) \cap N^c)$. Thus Y is measurable.

"$\Leftarrow$": Let $N' \subseteq N$, for $N \in \mathcal{F} : \mu(N) = 0$. Let $X = \mathbb{1}_N$ and $Y = \mathbb{1}_{N'}$. Then $X = Y$ a.e., so that, by assumption, Y is measurable. Therefore, $N' = Y^{-1}(\{1\}) \in \mathcal{F}$, so N' is measurable. Hence μ is complete.

2) "$\Rightarrow$": Let $N \in \mathcal{F} : \mu(N) = 0$ and $X_n \underset{n \to \infty}{\to} X$ pointwise on N^c. For all $n \in \mathbb{N}$, $X_n \mathbb{1}_{N^c}$ is measurable by C. 3.11 2) and converges everywhere to $X \mathbb{1}_{N^c}$ (the convergence holds on N^c by assumption, and otherwise both are 0). Then $X \mathbb{1}_{N^c}$ is measurable by L. 3.12 2). Since $X \mathbb{1}_{N^c} = X$ a.e., 1) implies that X is measurable.

"$\Leftarrow$": Let $N' \subseteq N$, for $N \in \mathcal{F} : \mu(N) = 0$. Let $X_n = \mathbb{1}_N$, $n \in \mathbb{N}$, and $X = \mathbb{1}_{N'}$. Then $X_n \underset{n \to \infty}{\overset{\text{a.e.}}{\to}} X$, so X is measurable. Therefore, $N' = X^{-1}(\{1\}) \in \mathcal{F}$, so N' is measurable. Hence μ is complete.

□

3.2 Distributions

Proposition 3.14 (X induces a measure on $(\Omega', \mathcal{F}')$)
If $(\Omega, \mathcal{F}, \mu)$ is a measure space, $(\Omega', \mathcal{F}')$ a measurable space and $X : \Omega \to \Omega'$ measurable, then $\mu_X := \mu \circ X^{-1}$ is a measure on $(\Omega', \mathcal{F}')$, the *distribution of X* (also *push-forward measure* or *image measure of μ wrt X*).

Proof.

i) $\mu_X : \mathcal{F}' \to [0,1]$ ✓

ii) $\mu_X(\emptyset) \underset{\text{def.}}{=} \mu(X^{-1}(\emptyset)) \underset{\text{def.}}{=} \mu(\emptyset) = 0$ ✓

iii) $\{A_i'\}_{i\in\mathbb{N}} \subseteq \mathcal{F}'$, $A_i' \cap A_j' = \emptyset \; \forall\, i \neq j \Rightarrow \mu_X(\biguplus_{i\in\mathbb{N}} A_i') \underset{\text{def.}}{=} \mu(X^{-1}(\biguplus_{i\in\mathbb{N}} A_i')) \underset{\text{S. 1.2}}{=}$
$\mu(\biguplus_{i\in\mathbb{N}} X^{-1}(A_i')) \overset{X^{-1}(A_i') \in \mathcal{F}}{\underset{\sigma\text{-add.}}{=}} \sum_{i=1}^\infty \mu(X^{-1}(A_i')) \underset{\text{def.}}{=} \sum_{i=1}^\infty \mu_X(A_i')$ ✓ □

- $\forall\, A' \in \mathcal{F}'$, $\mu(X \in A') := \mu(\{\omega \in \Omega : X(\omega) \in A'\}) \underset{\text{def. } X^{-1}}{=} \mu(X^{-1}(A')) \underset{\text{def. } \mu_X}{=} \mu_X(A')$
- We typically consider the case where $\mu = \mathbb{P}$ is a probability measure and $(\Omega', \mathcal{F}') = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $d \geq 1$. Then

$$\mathbb{P}(\boldsymbol{a} < \boldsymbol{X} \leq \boldsymbol{b}) := \mathbb{P}(\boldsymbol{X} \in (\boldsymbol{a}, \boldsymbol{b}]) = \mathbb{P}(\{\omega \in \Omega : \boldsymbol{X}(\omega) \in (\boldsymbol{a}, \boldsymbol{b}]\})$$
$$\underset{\text{above}}{=} \mathbb{P}_{\boldsymbol{X}}((\boldsymbol{a}, \boldsymbol{b}]), \quad -\boldsymbol{\infty} < \boldsymbol{a} < \boldsymbol{b} < \boldsymbol{\infty}. \tag{1}$$

- We can now make a connection between the distribution $\mathbb{P}_{\boldsymbol{X}}$ of $\boldsymbol{X}$ (a Borel probability measure) and its df F.

Proposition 3.15 (Characterization of $\boldsymbol{X}$ by F via $\mathbb{P}_{\boldsymbol{X}}$)

1) $\boldsymbol{X}$ induces F: If $\boldsymbol{X}$ is a random vector, then its distribution $\mathbb{P}_{\boldsymbol{X}}$ is a probability measure on $\mathbb{R}^d$ with df $F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$.
2) F induces $\boldsymbol{X}$: If F is a df on $\mathbb{R}^d$, then there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random vector $\boldsymbol{X} : \Omega \to \mathbb{R}^d$ such that $\mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}) = F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$.

Proof.

1) By P. 3.14 with $(\Omega', \mathcal{F}') = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we know that $\mathbb{P}_{\boldsymbol{X}}$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$. Its df is $F(\boldsymbol{x}) \underset{\text{R. 2.49 1)}}{=} \mathbb{P}_{\boldsymbol{X}}((-\boldsymbol{\infty}, \boldsymbol{x}]) \underset{(1)}{=} \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$.
2) We only consider the case $d = 1$; the general case can be shown iteratively based on "conditional distributions functions" we have not introduced yet. If $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1], \bar{\mathcal{B}}((0,1]), \lambda)$ and $X(\omega) := F^{-1}(\omega)$, $\omega \in \Omega$, then

$$\mathbb{P}(X \le x) \underset{(1)}{=} \mathbb{P}(\{\omega \in \Omega : X(\omega) \le x\}) \underset{\text{def. } X}{=} \mathbb{P}(\{\omega \in \Omega : F^{-1}(\omega) \le x\})$$

$$\underset{\text{L. 2.53 3)}}{=} \mathbb{P}(\{\omega \in \Omega : \omega \le F(x)\}) = \mathbb{P}((0, F(x)]) \underset{\text{def. } \mathbb{P}}{=} \lambda((0, F(x)]) \underset{\text{def. } \lambda}{=} F(x). \quad \square$$

Remark 3.16

1) Because of the correspondence between $\boldsymbol{X}$, $\mathbb{P}_{\boldsymbol{X}}$ and F, one writes $\boldsymbol{X} \sim F$, also $\boldsymbol{X} \sim F_{\boldsymbol{X}}$ to distinguish different dfs. If, for $j = 1, 2$, $\boldsymbol{X}_j \sim F$ is a random vector on $(\Omega_j, \mathcal{F}_j, \mathbb{P}_j)$, we also write $\boldsymbol{X}_1 \overset{\text{d}}{=} \boldsymbol{X}_2$, where $\overset{\text{d}}{=}$ denotes the equality *in distribution*. Also, for $\boldsymbol{X} \sim F$, we let $\operatorname{supp}(\boldsymbol{X}) := \operatorname{supp}(F)$.

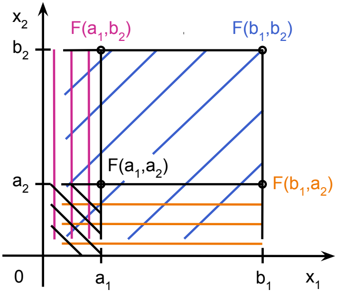
2) By P. 3.14, the distribution $\mathbb{P}_{\boldsymbol{X}} = \mathbb{P} \circ \boldsymbol{X}^{-1}$ of $\boldsymbol{X} \sim F$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$ with df F. We thus must have

$$\mathbb{P}(\boldsymbol{X} \in (\boldsymbol{a}, \boldsymbol{b}]) \underset{(1)}{=} \mathbb{P}_{\boldsymbol{X}}((\boldsymbol{a}, \boldsymbol{b}]) \underset{\text{R. 2.49 2)}}{=} \Delta_{(\boldsymbol{a},\boldsymbol{b}]} F, \quad -\boldsymbol{\infty} < \boldsymbol{a} \le \boldsymbol{b} < \boldsymbol{\infty},$$

so the F-volume is the probability of $\boldsymbol{X} \sim F$ to take values in $(\boldsymbol{a}, \boldsymbol{b}]$.

Visual confirmation ($d = 2$):

$$\begin{aligned}\Delta_{(\boldsymbol{a},\boldsymbol{b}]} F &\underset{\text{def.}}{=} F(b_1, b_2) - F(b_1, a_2) \\ &\quad - F(a_1, b_2) + F(a_1, a_2) \\ &\underset{\text{P. 3.15}}{=} \mathbb{P}\left(\boldsymbol{X} \le \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}\right) - \mathbb{P}\left(\boldsymbol{X} \le \begin{pmatrix} b_1 \\ a_2 \end{pmatrix}\right) \\ &\quad - \mathbb{P}\left(\boldsymbol{X} \le \begin{pmatrix} a_1 \\ b_2 \end{pmatrix}\right) + \mathbb{P}\left(\boldsymbol{X} \le \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}\right) \\ &\underset{\text{indeed}}{\overset{\text{visually}}{=}} \mathbb{P}(\boldsymbol{X} \in (\boldsymbol{a}, \boldsymbol{b}]).\end{aligned}$$



3) Random vectors are *discrete*, *continuous (singular)*, *absolutely continuous* or *mixed-type* if their dfs F are. As densities (pmfs) f also uniquely determine F, one also occasionally writes $\boldsymbol{X} \sim f$ meaning $\boldsymbol{X} \sim F$ for F having density (pmf) f; this also applies to other quantities that uniquely characterize F.
4) One also overloads the notation of F and uses $F(B) := \mathbb{P}_{\boldsymbol{X}}(B),\ B \in \bar{\mathcal{B}}(\mathbb{R}^d)$; e.g. $F((-\infty, \boldsymbol{x}]) = \mathbb{P}_{\boldsymbol{X}}((-\infty, \boldsymbol{x}])$. One therefore often uses the terms "distribution" and "distribution function" interchangeably.

Corollary 3.17 (Equality in distribution of functional transforms)
If, for $j = 1, 2$, $\boldsymbol{X}_j \sim F$ is a random vector on $(\Omega_j, \mathcal{F}_j, \mathbb{P}_j)$, then $\boldsymbol{h}(\boldsymbol{X}_1) \stackrel{\text{d}}{=} \boldsymbol{h}(\boldsymbol{X}_2)$.

Proof. $\boldsymbol{X}_1 \stackrel{\text{d}}{=} \boldsymbol{X}_2 \Rightarrow \forall$ measurable $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^k$, one has

$$F_{\boldsymbol{h}(\boldsymbol{X}_1)}(\boldsymbol{x}) = \mathbb{P}_1(\boldsymbol{h}(\boldsymbol{X}_1) \le \boldsymbol{x}) = \mathbb{P}_1\big(\boldsymbol{X}_1 \in \boldsymbol{h}^{-1}((-\infty, \boldsymbol{x}])\big) = F\big(\boldsymbol{h}^{-1}((-\infty, \boldsymbol{x}])\big) \underset{\text{backwards}}{\overset{\text{same}}{=}} F_{\boldsymbol{h}(\boldsymbol{X}_2)}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^k,$$

so $\boldsymbol{h}(\boldsymbol{X}_1) \stackrel{\text{d}}{=} \boldsymbol{h}(\boldsymbol{X}_2)$. □

3.3 Margins

- Let $\boldsymbol{X} = (X_1, \dots, X_d) \sim F$, $J \subseteq \{1, \dots, d\}$. Then the J-margin F_J of F is

  $$F_J(\boldsymbol{x}_J) \underset{\text{R. 2.49 8)}}{=} \lim_{\boldsymbol{x}_{J^c} \to \boldsymbol{\infty}} F(\boldsymbol{x}) \underset{\text{P. 3.15}}{=} \lim_{\boldsymbol{x}_{J^c} \to \boldsymbol{\infty}} \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}) \overset{\text{cont.}}{\underset{\text{below}}{=}} \mathbb{P}(\boldsymbol{X}_J \le \boldsymbol{x}_J), \ \boldsymbol{x}_J \in \mathbb{R}^{|J|},$$

  which, again by P. 3.15, is the distribution function of the random vector $\boldsymbol{X}_J = (X_j)_{j \in J}$. We therefore also call $\boldsymbol{X}_J$ the *J-margin of $\boldsymbol{X}$*.
- For $J = \{j\}$, $j = 1, \dots, d$, the *jth margin of F* or $\boldsymbol{X}$ is $F_j(x_j) = \mathbb{P}(X_j \le x_j)$, $x_j \in \mathbb{R}$.
- If F is absolutely continuous, then

  $$F_J(\boldsymbol{x}_J) = F(\boldsymbol{\infty}_{J \leftarrow \boldsymbol{x}_J}) = \int_{(-\boldsymbol{\infty}, \boldsymbol{\infty}_{J \leftarrow \boldsymbol{x}_J}]} f(\boldsymbol{z}) \,\mathrm{d}\boldsymbol{z} \overset{\text{Tonelli}}{\underset{\text{see later}}{=}} \int_{-\boldsymbol{\infty}}^{\boldsymbol{x}_J} \int_{-\boldsymbol{\infty}}^{\boldsymbol{\infty}} f(\boldsymbol{z}) \,\mathrm{d}\boldsymbol{z}_{J^c} \mathrm{d}\boldsymbol{z}_J,$$

  so that F_J is absolutely continuous with *J-marginal density of F* or $\boldsymbol{X}$ given by

  $$f_J(\boldsymbol{x}_J) = \int_{-\boldsymbol{\infty}}^{\boldsymbol{\infty}} f(\boldsymbol{z}_{J \leftarrow \boldsymbol{x}_J}) \,\mathrm{d}\boldsymbol{z}_{J^c}.$$
- So we see that F being absolutely continuous implies that all lower-dimensional margins are absolutely continuous, and we obtain their marginal densities by integrating out joint densities over the remaining variables.

- The converse does not hold in general, so absolutely continuous margins do not necessarily imply an absolutely continuous F, only if the underlying dependence structure (see later) has a density; see E. 3.22 later.
- Similarly, if F is discrete one has *joint/marginal probability mass functions*; in this case replace integrals by sums.

3.4 Survival functions

- Several quantities related to a df F of $\boldsymbol{X}$ are frequently of interest, one being the *survival function (sf)* of $\boldsymbol{X} \sim F$ defined by $\bar{F}(\boldsymbol{x}) := \mathbb{P}(\boldsymbol{X} > \boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, and often considered for $\boldsymbol{X} \geq \mathbf{0}$ a.s., so with support $[\mathbf{0}, \infty)$ and $\bar{F}(\mathbf{0}) = 1$.
- As for the margins of F, the *J-margin* of $\bar{F}$ is $\bar{F}_J(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X}_J > \boldsymbol{x}_J)$.
- The *jth marginal sf* is $\bar{F}_j(x_j) := \mathbb{P}(X_j > x_j) = 1 - F_j(x_j)$, $j = 1, \dots, d$.
- For $d = 1$, expressing $\bar{F}$ in terms of F is straightforward via $\bar{F}(x) = \mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x) = 1 - F(x)$, $x \in \mathbb{R}$, but for $d \geq 2$, $\bar{F}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} > \boldsymbol{x}) \neq 1 - \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) = 1 - F(\boldsymbol{x})$. For example, for $d = 2$,

$$\bar{F}(x_1, x_2) \underset{\text{def.}}{=} \mathbb{P}(X_1 > x_1, X_2 > x_2) \overset{\text{tot.}}{\underset{\text{prob.}}{=}} \mathbb{P}(X_1 > x_1) - \mathbb{P}(X_1 > x_1, X_2 \leq x_2)$$

$$\overset{\text{tot.}}{\underset{\text{prob.}}{=}} \mathbb{P}(X_1 > x_1) - (\mathbb{P}(X_2 \le x_2) - \mathbb{P}(X_1 \le x_1, X_2 \le x_2))$$

$$= 1 - F_1(x_1) - F_2(x_2) + F(x_1, x_2), \quad x_1, x_2 \in \mathbb{R}.$$

In general, we have the following result.

Lemma 3.18 ($\bar{F}$ in terms of F)

If $\boldsymbol{X} \sim F$, then

$$\bar{F}(\boldsymbol{x}) = \sum_{J \subseteq \{1,\dots,d\}} (-1)^{|J|} F(\boldsymbol{\infty}_{J \leftarrow \boldsymbol{x}_J}), \quad \boldsymbol{x} \in \mathbb{R}^d.$$

Proof.

$$\bar{F}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} > \boldsymbol{x}) \underset{()^c}{=} 1 - \mathbb{P}\Biggl(\bigcup_{j=1}^{d} \{X_j \le x_j\}\Biggr)$$

$$\overset{\text{incl.}}{\underset{\text{excl.}}{=}} 1 - \sum_{j=1}^{d} (-1)^{j-1} \sum_{J \subseteq \{1,\dots,d\}: |J| = j} \mathbb{P}\Biggl(\bigcap_{k \in J} \{X_k \le x_k\}\Biggr)$$

$$= 1 + \sum_{j=1}^{d} (-1)^{j} \sum_{J \subseteq \{1,\dots,d\}: |J| = j} \mathbb{P}\Biggl(\bigcap_{k \in J} \{X_k \le x_k\}\Biggr)$$

$$
\underset{(*)}{=} \sum_{j=0}^{d} (-1)^j \sum_{J\subseteq\{1,\dots,d\}:|J|=j} \mathbb{P}\Bigl(\bigcap_{k\in J} \{X_k \le x_k\}\Bigr)
$$

$$
= \sum_{j=0}^{d} (-1)^j \sum_{J\subseteq\{1,\dots,d\}:|J|=j} F(\boldsymbol{\infty}_{J\leftarrow \boldsymbol{x}_J}) = \sum_{J\subseteq\{1,\dots,d\}} (-1)^{|J|} F(\boldsymbol{\infty}_{J\leftarrow \boldsymbol{x}_J}),
$$

where we in $(*)$ assume that $\bigcap_{k\in\emptyset}\{X_k \le x_k\} = \Omega$. □

3.5 Extensions

- Similarly as in P. 3.10, one can show that $\boldsymbol{X}$ is a random sequence iff $\boldsymbol{X} = (X_1, X_2, \dots)$ for rvs $X_1, X_2, \dots$. Such an $\boldsymbol{X}$ is also known as a *discrete-time stochastic process*.
- The step to a *continuous-time stochastic process*, i.e. $(\boldsymbol{X}_t)_{t\in I}$ for $I \subseteq \mathbb{R}$ can be made via Kolmogorov's extension theorem.
- To this end, $\forall\, k \in \mathbb{N}$, $t_1 < \dots < t_k \in I$, the probability measures $\mathbb{P}_{t_1,\dots,t_k}(\prod_{i=1}^{k} B_i) := \mathbb{P}(\boldsymbol{X}_{t_1} \in B_1, \dots, \boldsymbol{X}_{t_k} \in B_k)$, $B_i \in \mathcal{B}(\mathbb{R}^d)$ $\forall\, i = 1, \dots, k$, are the *finite-dimensional distributions* of $\boldsymbol{X} = (\boldsymbol{X}_t)_{t\in I}$.

Theorem 3.19 (Kolmogorov's extension theorem)
$\forall\, k \in \mathbb{N}$, $t_1 < \dots < t_k \in I$, let $\mathbb{P}_{t_1,\dots,t_k}$ be probability measures on $(\mathbb{R}^d)^k$ (Cartesian products of k Borel sets, each in $\mathbb{R}^d$) satisfying

i) $\mathbb{P}_{t_{\pi(1)},\dots,t_{\pi(k)}}(\prod_{i=1}^k B_{\pi(i)}) = \mathbb{P}_{t_1,\dots,t_k}(\prod_{i=1}^k B_i)$, $\forall\, B_1,\dots,B_k \in \mathcal{B}(\mathbb{R}^d)$ and all permutations π of $\{1,\dots,k\}$ (permutation invariance); and
ii) $\mathbb{P}_{t_1,\dots,t_{k+1}}(\prod_{i=1}^k B_i \times \mathbb{R}^d) = \mathbb{P}_{t_1,\dots,t_k}(\prod_{i=1}^k B_i)$ $\forall\, B_1,\dots,B_k \in \mathcal{B}(\mathbb{R}^d)$ (compatibility).

Then there exists a probability measure $\mathbb{P}$ on $\big((\mathbb{R}^d)^I, \mathcal{B}((\mathbb{R}^d)^I)\big)$ (note: $(\mathbb{R}^d)^I$ denotes all functions from I to $\mathbb{R}^d$) such that the stochastic process (coordinate process) $\boldsymbol{X} = (\boldsymbol{X}_t)_{t\in I} : (\mathbb{R}^d)^I \to \mathbb{R}^d$ evaluating its argument from $(\mathbb{R}^d)^I$ at t has (the given compatible) $(\mathbb{P}_{t_1,\dots,t_k})_{t_1<\dots<t_k\in I,\ k\in\mathbb{N}}$ as its finite-dimensional distributions (so there exists a probability space, namely $\big((\mathbb{R}^d)^I, \mathcal{B}((\mathbb{R}^d)^I), \mathbb{P}\big)$, and a stochastic process, namely $\boldsymbol{X}$, with these finite-dimensional distributions).

Proof. Application of Carathéodory's extension theorem; see Billingsley (1995, T. 36.1) based on product-σ-algebra and finite intersections of preimages of projections (cylinder sets). □

If the finite-dimensional dfs are jointly normal, one obtains Brownian motion.

3.6 Examples of distributions

3.6.1 Examples of univariate distributions

Example 3.20 (Bernoulli distribution)

- The df of the *Bernoulli distribution* $\mathrm{B}(1,p)$ is $F(x) = (1-p)\mathbb{1}_{[0,\infty)}(x) + p\mathbb{1}_{[1,\infty)}(x)$, $x \in \mathbb{R}$, for some parameter $p \in [0,1]$.
- A rv $X \sim \mathrm{B}(1,p)$ satisfies $\mathbb{P}(X = 0) = 1-p$ and $\mathbb{P}(X = 1) = p$. Thus $\mathbb{1}_A \sim \mathrm{B}(1,p)$ for any event $A \in \mathcal{F}$ with $\mathbb{P}(A) = p$.

Example 3.21 (Uniform distribution)

- The df of the *uniform distribution* $\mathrm{U}(a,b)$ is $F(x) = \frac{\min\{\max\{x,a\},b\}-a}{b-a}$, $x \in \mathbb{R}$, or $F(x) = \frac{x-a}{b-a}$, $x \in [a,b]$, for some parameters $-\infty < a < b < \infty$.
- F is differentiable a.e. with $F'(x) = \frac{1}{b-a}\mathbb{1}_{(a,b)}(x)$, $x \notin \{a,b\}$, which integrates to F, so F is abs. cont. with density $f(x) = F'(x) = \frac{1}{b-a}\mathbb{1}_{(a,b)}(x)$, $x \in \mathbb{R}$.
  Interpretation: A constant density $\frac{1}{b-a}$ over (a,b) means for $X \sim \mathrm{U}(a,b)$ that

$$\mathbb{P}(X \in (x, x+h]) = \int_x^{x+h} 1/(b-a)\,\mathrm{d}z = h/(b-a), \quad x \in [a, b-h],$$

i.e. there is equal probability of $X \sim \mathrm{U}(a,b)$ to fall in any interval of length h.

- For rvs following $\mathrm{U}(0,1)$, one typically writes the letter U. There are various algorithms to generate realizations of $U \sim \mathrm{U}(0,1)$ with software.
- $\mathbb{P}(X=x) \underset{\text{R. 2.49 3)}}{=} F(x)-F(x-) \underset{\text{cont.}}{=} F(x)-F(x) = 0$, $x \in [a,b]$. In particular, $\mathbb{P}(X=x) \neq \frac{1}{b-a} = f(x)$, $x \in (a,b)$. Densities should be interpreted via
$$f(x) \underset{h>0 \text{ small}}{\approx} \frac{\mathbb{P}(X \in (x,x+h])}{h}$$
since $\mathbb{P}(X \in (x,x+h]) = F(x+h)-F(x) = \int_x^{x+h} f(z)\,\mathrm{d}z \underset{h>0 \text{ small}}{\approx} f(x)(x+h-x) = f(x)h$.
- Also note that "$X \sim \mathrm{U}(a,b)$ has density $f(x) = \frac{1}{b-a}$" is wrong. For densities or pmfs, always provide a domain (here: $x \in (a,b)$, not $x \in \mathbb{R}$).

Question: Is there a uniform distribution on an unbounded interval $I=(a,b)$ with $a=-\infty$ or $b=\infty$? No.

Proof.

Case 1): $f(x)=0,\ x \in I \Rightarrow f(x)=0,\ x \in \mathbb{R} \Rightarrow \int_{-\infty}^{\infty} 0\,\mathrm{d}z = 0 \neq 1$ ↯.

Case 2): Suppose $f(x)=c>0,\ x \in I$. Then $\int_{-\infty}^{\infty} f(z)\,\mathrm{d}z \geq \int_I f(z)\,\mathrm{d}z = \int_I c\,\mathrm{d}z = \infty \neq 1$ ↯ □

We can now provide a counterexample for a statement in S. 3.3.

Example 3.22 (Abs. cont. margins $\not\Rightarrow$ abs. cont. F)

$\boldsymbol{X} = (U, U)$ for $U \sim \mathrm{U}(0,1)$ has $\mathrm{U}(0,1)$ margins with $f_{X_j}(x_j) = \mathbb{1}_{(0,1)}(x_j)$, $j = 1, 2$, but $F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}(U \le x_1, U \le x_2) = \mathbb{P}(U \le \min\{x_1, x_2\}) = \min\{x_1, x_2\}$ has no density since

$$\frac{\partial^2}{\partial x_2 \partial x_1} F_{\boldsymbol{X}}(\boldsymbol{x}) = \begin{cases} \frac{\partial^2}{\partial x_2 \partial x_1} x_1 = \frac{\partial}{\partial x_2} 1 = 0, & x_1 < x_2, \\ \frac{\partial^2}{\partial x_2 \partial x_1} x_2 = \frac{\partial}{\partial x_2} 0 = 0, & x_1 > x_2, \end{cases}$$

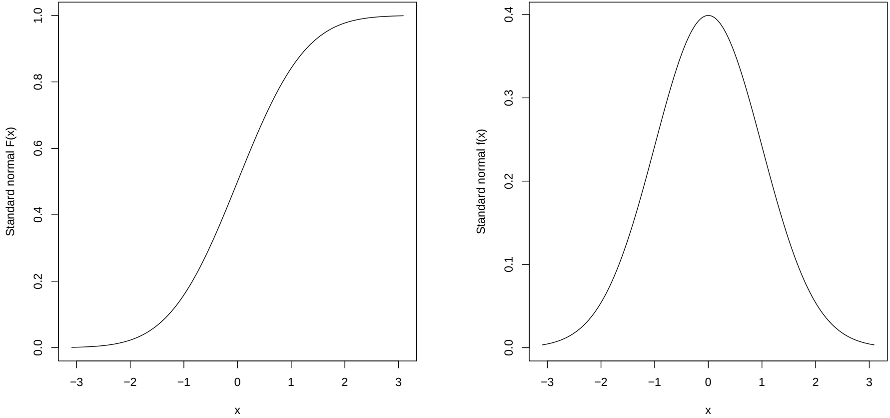
which integrates to $0 \neq 1$.

Example 3.23 (Exponential distribution)

- The *exponential df* is $F(x) = (1 - e^{-\lambda x})\mathbb{1}_{[0,\infty)}(x)$ for some $\lambda > 0$.
- F is differentiable a.e. with $F'(x) = \lambda e^{-\lambda x}\mathbb{1}_{(0,\infty)}(x)$, $x \neq 0$, which integrates to F, so F is absolutely continuous with density $f(x) = F'(x) = \lambda e^{-\lambda x}\mathbb{1}_{(0,\infty)}(x)$.
- The quantile function of F (obtained by solving $F(x) = y$ wrt x) is $F^{-1}(y) = -\frac{1}{\lambda}\log(1 - y)$, $y \in (0, 1]$.

Example 3.24 (Normal distribution)

The normal df is $F(x) = \Phi(\frac{x-\mu}{\sigma}) = \int_{-\infty}^{x} f(z)\,\mathrm{d}z$ (left) with density $f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$, $x \in \mathbb{R}$ (right) for $\mu \in \mathbb{R}$ (here: $\mu = 0$) and $\sigma > 0$ (here: $\sigma = 1$). The case $\mu = 0$, $\sigma = 1$ is known as standard normal df.



This rather non-tractable F is of interest due to its properties (role in the "central limit theorem", sums of "jointly normal random variables", etc.).

Verifying that the integrable, non-negative $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ is indeed a proper density on $\mathbb{R}$ can be done as follows. With

$$y = \frac{x-\mu}{\sigma}$$

we obtain that

$$\int_{-\infty}^{\infty} f(z)\,\mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}y \underset{\text{sym.}}{=} \sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-y^2/2}\,\mathrm{d}y =: \sqrt{\frac{2}{\pi}} I.$$

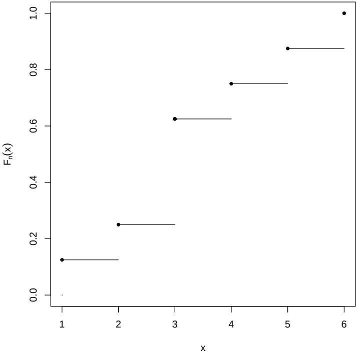
Then

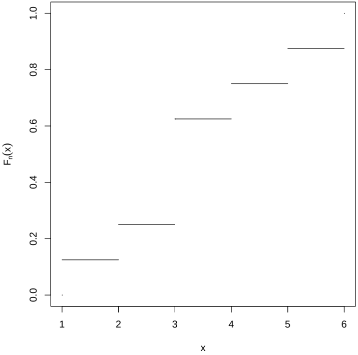
$$\begin{aligned} I^2 &= I \cdot I = \int_0^\infty e^{-\frac{x^2}{2}}\,\mathrm{d}x \int_0^\infty e^{-\frac{y^2}{2}}\,\mathrm{d}y = \int_0^\infty \Bigl(\int_0^\infty e^{-\frac{y^2}{2}}\,\mathrm{d}y\Bigr) e^{-\frac{x^2}{2}}\,\mathrm{d}x \\ &= \int_0^\infty \Bigl(\underbrace{\int_0^\infty e^{-\frac{x^2+y^2}{2}}\,\mathrm{d}y}_{\underset{\text{y = xt}}{=} \int_0^\infty x e^{-\frac{x^2(1+t^2)}{2}}\,\mathrm{d}t} \Bigr)\,\mathrm{d}x \underset{\text{Tonelli}}{=} \int_0^\infty \Bigl(\int_0^\infty x e^{-\frac{x^2(1+t^2)}{2}}\,\mathrm{d}x\Bigr)\,\mathrm{d}t \\ &= \int_0^\infty \Bigl[-\frac{1}{1+t^2} e^{-\frac{x^2(1+t^2)}{2}}\Bigr]_0^\infty \mathrm{d}t = \int_0^\infty \frac{1}{1+t^2}\,\mathrm{d}t = [\arctan(t)]_0^\infty = \frac{\pi}{2}, \end{aligned}$$

so $I = \sqrt{\frac{\pi}{2}}$ and thus $\int_{-\infty}^{\infty} f(z)\,\mathrm{d}z = \sqrt{\frac{2}{\pi}}\sqrt{\frac{\pi}{2}} = 1$.

Example 3.25 (Empirical distribution)

The discrete empirical df (edf) of the n (here: $n = 8$) data points $\boldsymbol{X} = (X_1, \dots, X_n) = (1, 2, 3, 3, 3, 4, 5, 6)$ is $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \le x\}}$ (relative frequency of points $\le x$; left).





- F_n jumps at each unique data point $x = X_i$ by its relative frequency (e.g. at $x = 1$ by $1/8$, at $x = 3$ by $3/8$); for larger n, one can use vertical bars (right).
- F_n approximates to the (true underlying, but unknown df) F that produced $\boldsymbol{X}$. Later: "$F_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} F$" pointwise and uniformly.

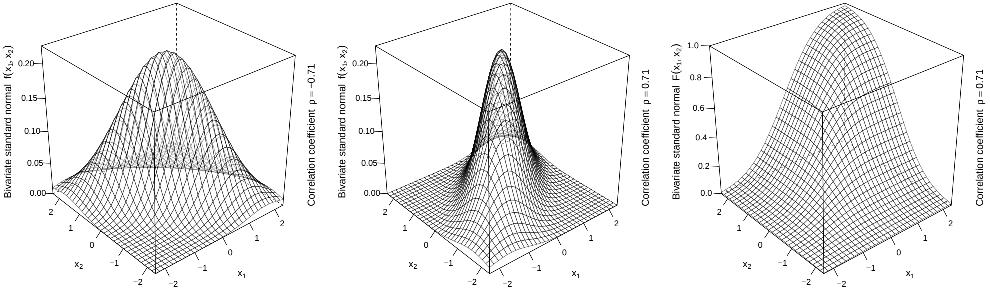
3.6.2 Examples of multivariate distributions

Example 3.26 (Absolutely continuous distribution)

The density of the normal df $F(\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$ can be shown to be

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\Bigl(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Bigr), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

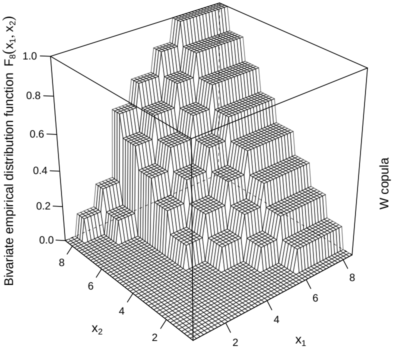
for $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ being symmetric, positive definite (a *covariance matrix*, more later; here: $\boldsymbol{\mu} = \boldsymbol{0}$ (all figures), $\Sigma = \left(\begin{smallmatrix} 1 & -0.71 \\ -0.71 & 1 \end{smallmatrix}\right)$ (left), $\Sigma = \left(\begin{smallmatrix} 1 & 0.71 \\ 0.71 & 1 \end{smallmatrix}\right)$ (middle, right)).



Example 3.27 (Discrete distribution)

The discrete empirical df (edf) of

$$X = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} \\ & \vdots \\ X_{n,1} & X_{n,2} \end{pmatrix} = \begin{pmatrix} 1 & 8 \\ 2 & 7 \\ 3 & 6 \\ 3 & 5 \\ 3 & 4 \\ 4 & 3 \\ 5 & 2 \\ 6 & 1 \end{pmatrix}$$

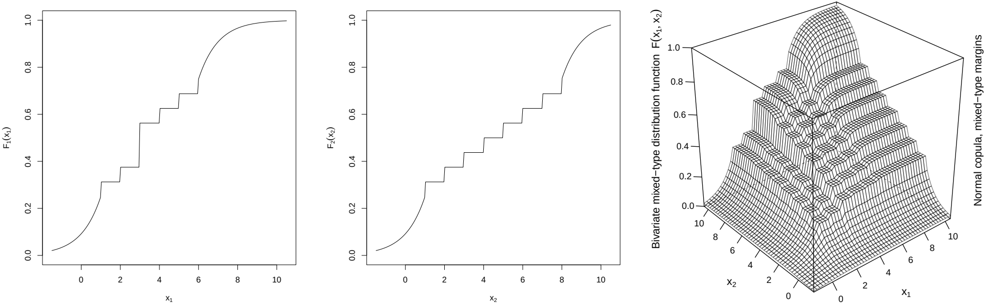


is $F_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{X}_i \le \boldsymbol{x}\}}$ (the first column or *component sample* is as in E. 3.25).

- Due to *ties* (equal values in at least one component sample) $F_n(x_1, x_2) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_{i,1}\le x_1,\ X_{i,2}\le x_2\}}$ jumps 6 (instead of 8) times in the first dimension.
- As for $d = 1$, F_n approximates the (true underlying, but unknown df) F that produced X, and "$F_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} F$" pointwise and uniformly.

Example 3.28 (Mixed-type distribution)

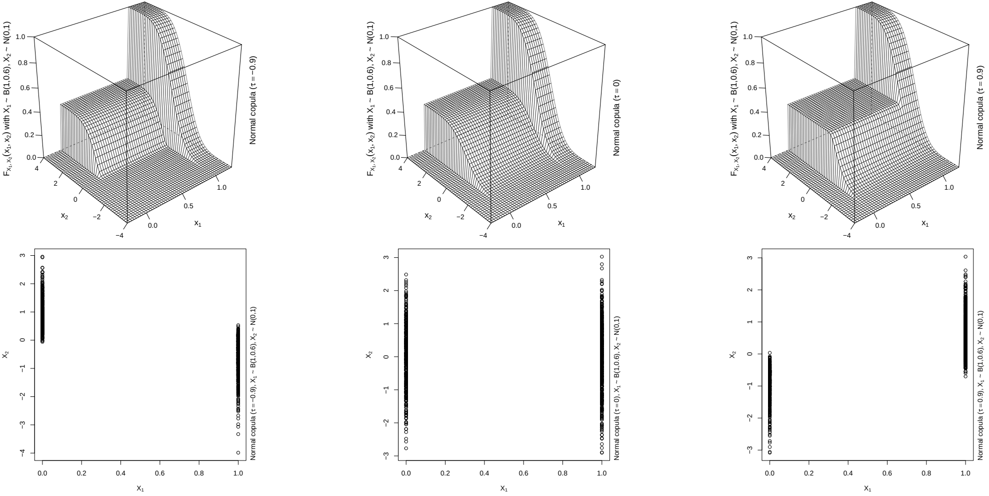
Based on the same data as in E. 3.27 and with absolutely continuous parts on $(-\infty, 1)$ (covering the y-axis range $(0, 1/4)$) and $[6, \infty)$ for F_1 and $[8, \infty)$ for F_2 (covering the y-axis range $[3/4, 1)$), we obtain a bivariate mixed-type df:



- The left shows F_1, the middle F_2, and the right F.
- Such distributions (with much larger number of samples in the body) frequently appear in practice, combining actual observations in the body with (absolutely) continuous and thus more tractable distributions in the tails.
- **Question:** Why is F_1 (F_2) best visible for large x_2 (x_1)?

Example 3.29 (Distribution with discrete and continuous margins)

F (first row) with discrete F_1 ($X_1 \sim \mathrm{B}(1, 0.6)$), continuous F_2 ($X_2 \sim \mathrm{N}(0, 1)$), and "dependence" between X_1 and X_2 given by a normal copula (more later) mimicking negative dependence (left), independence (centre) and positive dependence (right). The dependence is also visible from realizations of $(X_1, X_2) \sim F$ (second row).



3.7 Univariate transforms

We now discuss two important transformations of random variables.

Proposition 3.30 (Quantile transform)
Let F be a df and $U \sim \mathrm{U}(0,1)$. Then $F^{-1}(U) \sim F$.

Proof. $\mathbb{P}(F^{-1}(U) \le x) \underset{\text{L. 2.53 3)}}{=} \mathbb{P}(U \le F(x)) = F(x)$, $x \in \mathbb{R}$. □

- See also the proof of P. 3.15 2) (F induces X).
- P. 3.30 provides a representation of $X \sim F$ via $X \overset{\mathrm{d}}{=} F^{-1}(U)$. A representation in distribution of a random element in terms of others, typically more elementary ones, is called a *stochastic representation (sr)*. Srs are at the core of understanding stochastic models.

Example 3.31 (Stochastic representations)

1) By P. 3.30 and E. 3.20, $X \sim \mathrm{B}(1,p)$ has sr $X \overset{\mathrm{d}}{=} \mathbb{1}_{\{U \le p\}}$ for $U \sim \mathrm{U}(0,1)$.
2) $X \sim \mathrm{U}(a,b)$ has sr $X \overset{\mathrm{d}}{=} a + (b-a)U$ for $U \sim \mathrm{U}(0,1)$.

Proof.

- Either directly via $F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}(a + (b-a)U \le x) = \mathbb{P}(U \le \frac{x-a}{b-a}) = \frac{x-a}{b-a}$, $x \in [a, b]$.
- Or, by P. 3.30, $X \overset{\text{d}}{=} F^{-1}(U)$ with $F^{-1}(y) = a + (b-a)y$. □

3) $X \sim \mathrm{Exp}(\lambda)$, $\lambda > 0$, has sr $X \overset{\text{d}}{=} -\frac{1}{\lambda}\log(1-U) \overset{\text{d}}{=} -\frac{1}{\lambda}\log(U)$ for $U \sim \mathrm{U}(0,1)$.

Under continuity of F a converse is given as follows; an extension to arbitrary dfs F can also be given.

Proposition 3.32 (Probability transform)
If $X \sim F$, F continuous, then $F(X) \sim \mathrm{U}(0,1)$.

Proof. F continuous $\underset{\text{L. 2.53 4)}}{\overset{\text{no jumps}}{\Rightarrow}}$ $F^{-1} \uparrow$ on $[0,1]$, so

$$\begin{aligned}\mathbb{P}(F(X) \le u) &= \mathbb{P}(F^{-1}(F(X)) \le F^{-1}(u)) \underset{\text{L. 2.53 1)}}{\overset{F \uparrow \text{ on } \mathrm{supp}(F)}{=}} \mathbb{P}(X \le F^{-1}(u)) \\ &= F(F^{-1}(u)) \underset{\text{L. 2.53 2)}}{=} u, \quad u \in \mathrm{ran}(F) \cup \{0,1\} \underset{F \text{ cont.}}{=} [0,1]. \end{aligned}$$ □

If F is not continuous $\Rightarrow \exists\, x \in \mathbb{R} : F(x) > F(x-) \Rightarrow F(X) \notin (F(x-), F(x)) \subseteq [0,1] \Rightarrow F(X)$ cannot be $\mathrm{U}(0,1)$ distributed.

References

Billingsley, P. (1995), Probability and Measure, 3rd ed., Wiley.

4 Ordinary conditional probability, independence and dependence

4.1 Ordinary conditional probability

Proposition 4.1 (Ordinary conditional probability)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $B \in \mathcal{F} : \mathbb{P}(B) > 0$. Then $\mathbb{P}(A \,|\, B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, $A \in \mathcal{F}$ is a probability measure on $(\Omega, \mathcal{F})$, the *(ordinary) conditional probability of A given B*.

Proof.

1) $0 \le \mathbb{P}(A \cap B) \underset{\text{mon.}}{\le} \mathbb{P}(B)\ \forall\, A \in \mathcal{F} \Rightarrow \mathbb{P}(\cdot \,|\, B) : \mathcal{F} \to [0, 1]$.

2) $\mathbb{P}(\Omega \,|\, B) = \mathbb{P}(\Omega \cap B)/\mathbb{P}(B) = \mathbb{P}(B)/\mathbb{P}(B) = 1$.

3) $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F},\ A_i \cap A_j = \emptyset\ \forall\, i \neq j \Rightarrow \mathbb{P}\big(\biguplus_{i=1}^\infty A_i \,\big|\, B\big) \underset{\text{def.}}{=} \frac{\mathbb{P}((\biguplus_{i=1}^\infty A_i) \cap B)}{\mathbb{P}(B)}$

$\underset{\text{distr.}}{=} \frac{\mathbb{P}(\biguplus_{i=1}^\infty (A_i \cap B))}{\mathbb{P}(B)} \underset{\sigma\text{-add.}}{=} \frac{\sum_{i=1}^\infty \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^\infty \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} \underset{\text{def.}}{=} \sum_{i=1}^\infty \mathbb{P}(A_i \,|\, B)$. □

- If $\mathbb{P}(B) = 0$, we use the convention that $\mathbb{P}(A \,|\, B)\mathbb{P}(B) = 0$, motivated by $\mathbb{P}(A \,|\, B)\mathbb{P}(B) \underset{\text{def.}}{=} \mathbb{P}(A \cap B) \underset{\text{mon.}}{\le} \mathbb{P}(B) = 0$, which makes sense for any definition of $\mathbb{P}(A \,|\, B)$ if $\mathbb{P}(B) = 0$.
- Important results involving ordinary conditional probabilities are the law of total probability and Bayes' theorem.

Theorem 4.2 (Law of total probability)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{B_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$ a partition of Ω. Then

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \underbrace{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}_{=0 \text{ if } \mathbb{P}(B_i)=0}, \quad A \in \mathcal{F}.$$

Proof. $\mathbb{P}(A) \underset{\text{meas.}}{\overset{\text{tot.}}{=}} \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) \underset{\text{def.}}{=} \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$ □

Theorem 4.3 (Bayes' theorem)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A \in \mathcal{F} : \mathbb{P}(A) > 0$. Then

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}, \quad B \in \mathcal{F}.$$

If $\{B_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$ is a partition of Ω, then

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} \underset{\text{prob.}}{\overset{\text{tot.}}{=}} \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{\infty} \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}, \quad i \in \mathbb{N}.$$

Proof. We have

$$\mathbb{P}(B \mid A) \underset{\text{def.}}{=} \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \underset{\text{def.}}{=} \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$ □

4.2 Independence

4.2.1 Definition and properties

If $\mathbb{P}(A \,|\, B)$ does not depend on B, then $\mathbb{P}(A \cap B) = \mathbb{P}(A \,|\, \not{B})\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B)$. This motivates the following definition of the notion of independent events.

Definition 4.4 (Independence)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $n \geq 2$.

1) $A_1, \dots, A_n \in \mathcal{F}$ are *independent* if $\mathbb{P}(\bigcap_{i \in I} A_i) = \prod_{i \in I} \mathbb{P}(A_i)\ \forall\, I \subseteq \{1, \dots, n\}$.
2) $\mathcal{A}_1, \dots, \mathcal{A}_n \subseteq \mathcal{F}$ are *independent* if $A_1, \dots, A_n$ are $\forall\, A_i \in \mathcal{A}_i$, $i = 1, \dots, n$.
3) $\mathcal{A}_i \subseteq \mathcal{F}$, $i \in I$, are *independent* if $\mathcal{A}_{i_1}, \dots, \mathcal{A}_{i_n}$ are $\forall\, \{i_1, \dots, i_n\} \subseteq I$, $n \geq 2$.
4) Random elements X_i, $i \in I$, defined on the same probability space are *independent* if $\sigma(X_i)$, $i \in I$, are independent.

In particular, rvs are independent if $X_{i_1}^{-1}(B_{i_1}), \dots, X_{i_n}^{-1}(B_{i_n})$ are independent $\forall\, B_{i_1}, \dots, B_{i_n} \in \mathcal{B}(\mathbb{R})$, $\{i_1, \dots, i_n\} \subseteq I$, $n \geq 2$. And random vectors are independent if $\boldsymbol{X}_{i_1}^{-1}(B_{i_1}), \dots, \boldsymbol{X}_{i_n}^{-1}(B_{i_n})$ are independent $\forall\, B_{i_1} \in \mathcal{B}(\mathbb{R}^{d_1}), \dots, B_{i_n} \in \mathcal{B}(\mathbb{R}^{d_n})$, $\{i_1, \dots, i_n\} \subseteq I$, $n \geq 2$.

Example 4.5 (Factorization $\not\Rightarrow$ independence)

Consider rolling a fair die twice, so $\Omega = \{(\omega_1, \omega_2) : \omega_j \in \{1, \dots, 6\}, j = 1, 2\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$, $A \in \mathcal{F}$ (Laplace probability space). Let

$$A_1 = \text{“first roll} \le 3\text{”}, \quad A_2 = \text{“first roll is } 3, 4 \text{ or } 5\text{”}, \quad A_3 = \text{“sum is } 9\text{”}.$$

We show that $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$ but no two of the three events are independent.

- $\mathbb{P}(A_1) = \frac{1}{2} = \mathbb{P}(A_2)$. $\mathbb{P}(A_3) = \mathbb{P}(\{(3,6), (4,5), (5,4), (6,3)\}) = \frac{1}{9}$.
- $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(\{(3,6)\}) = \frac{1}{36} = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$.
- We have

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(\text{“first roll is 3”}) = \frac{6}{36} = \frac{1}{6} \neq \frac{1}{4} = \mathbb{P}(A_1)\mathbb{P}(A_2),$$
$$\mathbb{P}(A_1 \cap A_3) = \mathbb{P}(\{(3,6)\}) = \frac{1}{36} \neq \frac{1}{18} = \mathbb{P}(A_1)\mathbb{P}(A_3),$$
$$\mathbb{P}(A_2 \cap A_3) = \mathbb{P}(\{(3,6), (4,5), (5,4)\}) = \frac{3}{36} = \frac{1}{12} \neq \frac{1}{18} = \mathbb{P}(A_2)\mathbb{P}(A_3).$$

Example 4.6 (Pairwise independence $\not\Rightarrow$ independence)

Again consider rolling a fair die twice, so $\Omega = \{(\omega_1, \omega_2) : \omega_j \in \{1, \dots, 6\}, j = 1, 2\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$, $A \in \mathcal{F}$. Let

$$A_1 = \text{“first roll is even”},$$
$$A_2 = \text{“second roll is even”},$$
$$A_3 = \text{“both rolls are even or both rolls are odd”}.$$

We show that A_1, A_2, A_3 are pairwise independent but not independent.

- First, $\mathbb{P}(A_1) = \frac{1}{2} = \mathbb{P}(A_2)$, and $\mathbb{P}(A_3) = \mathbb{P}(\text{“both even”}) + \mathbb{P}(\text{“both odd”}) = \frac{9}{36} + \frac{9}{36} = \frac{18}{36} = \frac{1}{2}$.
- Then

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(\text{“both are even”}) = \frac{9}{36} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A_1)\mathbb{P}(A_2),$$
$$\mathbb{P}(A_1 \cap A_3) = \mathbb{P}(\text{“both are even”}) = \mathbb{P}(A_1)\mathbb{P}(A_3),$$
$$\mathbb{P}(A_2 \cap A_3) = \mathbb{P}(\text{“both are even”}) = \mathbb{P}(A_2)\mathbb{P}(A_3).$$

- $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(\text{“both are even”}) = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$.

Lemma 4.7 (Independence via generators)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $\mathcal{A}_i \subseteq \mathcal{F}$, $i \in I$, are independent π-systems, then $\sigma(\mathcal{A}_i)$, $i \in I$, are independent. If $\{A_i\}_{i\in I} \subseteq \mathcal{F}$ are ind., so are $\{A_i^c\}_{i\in I}$.

Proof. For $n \geq 2$, let $\{i_1, \dots, i_n\} \subseteq I$ and fix $A_{i_k} \in \mathcal{A}_{i_k}$, $k = 2, \dots, n$. Let $\mathcal{D} = \{A \in \mathcal{F} : \mathbb{P}(A \cap A_{i_2} \cap \ldots \cap A_{i_n}) = \mathbb{P}(A) \prod_{k=2}^{n} \mathbb{P}(A_{i_k})\}$ (good sets). Then

i) $\Omega \in \mathcal{D}$ (by def. of independence of $\mathcal{A}_{i_k}$, $k = 2, \dots, n$) ✓

ii) $A \in \mathcal{D}$ implies that $A^c \in \mathcal{D}$, since

$$\mathbb{P}\Big(A^c \cap \bigcap_{k=2}^{n} A_{i_k}\Big) = \mathbb{P}\Big(\bigcap_{k=2}^{n} A_{i_k} \Big\backslash A\Big) = \mathbb{P}\Big(\bigcap_{k=2}^{n} A_{i_k} \Big\backslash \Big(A \cap \bigcap_{k=2}^{n} A_{i_k}\Big)\Big)$$

$$\underset{\text{subtr.}}{=} \mathbb{P}\Big(\bigcap_{k=2}^{n} A_{i_k}\Big) - \mathbb{P}\Big(A \cap \bigcap_{k=2}^{n} A_{i_k}\Big)$$

$$\underset{A \in \mathcal{D}}{\overset{\text{ass.}}{=}} \prod_{k=2}^{n} \mathbb{P}(A_{i_k}) - \mathbb{P}(A)\Big(\prod_{k=2}^{n} \mathbb{P}(A_{i_k})\Big) = \mathbb{P}(A^c) \prod_{k=2}^{n} \mathbb{P}(A_{i_k}).$$

iii) If $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{D}$, $A_i \cap A_j = \emptyset$ $\forall i \neq j$, then $\biguplus_{i=1}^{\infty} A_i \in \mathcal{D}$ since

$$\mathbb{P}\Big(\Big(\biguplus_{i=1}^{\infty} A_i\Big) \cap \bigcap_{k=2}^{n} A_{i_k}\Big) \underset{\text{distr.}}{=} \mathbb{P}\Big(\biguplus_{i=1}^{\infty} \Big(A_i \cap \bigcap_{k=2}^{n} A_{i_k}\Big)\Big) \underset{\sigma\text{-add.}}{=} \sum_{i=1}^{\infty} \mathbb{P}\Big(A_i \cap \bigcap_{k=2}^{n} A_{i_k}\Big)$$

$$\underset{A_i \in \mathcal{D}}{=} \sum_{i=1}^{\infty}\Bigl(\mathbb{P}(A_i)\prod_{k=2}^{n}\mathbb{P}(A_{i_k})\Bigr) = \prod_{k=2}^{n}\mathbb{P}(A_{i_k})\cdot\sum_{i=1}^{\infty}\mathbb{P}(A_i)$$

$$\underset{\sigma\text{-add.}}{=} \prod_{k=2}^{n}\mathbb{P}(A_{i_k})\cdot\mathbb{P}\Bigl(\biguplus_{i=1}^{\infty}A_i\Bigr).$$

- Hence $\mathcal{D}$ is a Dynkin system. Since $\mathcal{A}_{i_1},\dots,\mathcal{A}_{i_n}$ are independent by ass., $\mathcal{D}$ also contains $\mathcal{A}_{i_1}$. By ass., $\mathcal{A}_{i_1}$ is a π-system $\underset{\pi\text{-}\lambda}{\Rightarrow} \sigma(\mathcal{A}_{i_1}) \subseteq \mathcal{D} \Rightarrow \sigma(\mathcal{A}_{i_1}), \mathcal{A}_{i_2},\dots,\mathcal{A}_{i_n}$ are independent $\underset{\text{iteratively}}{\Rightarrow}$ $\sigma(\mathcal{A}_{i_1}), \sigma(\mathcal{A}_{i_2}),\dots,\sigma(\mathcal{A}_{i_n})$ are independent.
- $\{A_i\}_{i\in I} \subseteq \mathcal{F}$ ind. $\Rightarrow \{\mathcal{A}_i\}_{i\in I}$ with $\mathcal{A}_i = \{A_i\}$ are ind. π-systems $\Rightarrow \{\sigma(\mathcal{A}_i)\}_{i\in I}$ are ind. Since $A_i^c \in \sigma(\mathcal{A}_i)$, $i \in I$, we have that $\{A_i^c\}_{i\in I}$ are ind. □

Lemma 4.8 (Grouping lemma)

Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space. If $\mathcal{A}_i \subseteq \mathcal{F}$, $i \in I$, are independent π-systems and $I = \biguplus_{j\in J} I_j$, then $\sigma(\bigcup_{i\in I_j}\mathcal{A}_i)$, $j \in J$, are independent.

Proof.

- For $j \in J$, $\mathcal{A}_{I_j} := \{\bigcap_{i\in\tilde{I}_j} A_i : A_i \in \mathcal{A}_i,\ \tilde{I}_j \subseteq I_j,\ |\tilde{I}_j| < \infty\}$ are π-systems.
- $\{\mathcal{A}_i\}_{i\in I}$ are ind. π-sys. $\underset{\text{def.}}{\Rightarrow} \{\mathcal{A}_{I_j}\}_{j\in J}$ are ind. π-sys. $\underset{\text{L. 4.7}}{\Rightarrow} \{\sigma(\mathcal{A}_{I_j})\}_{j\in J}$ are ind.

- We now show that $\sigma(\bigcup_{i\in I_j}\mathcal{A}_i)=\sigma(\mathcal{A}_{I_j})$, which implies the statement.
  "$\subseteq$": $\forall\, i\in I_j$ we have $\mathcal{A}_i \underset{\text{take } \tilde{I}_j=\{i\}}{\subseteq} \mathcal{A}_{I_j}\subseteq\sigma(\mathcal{A}_{I_j}) \Rightarrow \bigcup_{i\in I_j}\mathcal{A}_i\subseteq\sigma(\mathcal{A}_{I_j}) \underset{\text{smallest}}{\Rightarrow}$ $\sigma(\bigcup_{i\in I_j}\mathcal{A}_i)\subseteq\sigma(\mathcal{A}_{I_j})$.
  "$\supseteq$": If $A_i\in\mathcal{A}_i$, $i\in\tilde{I}_j\subseteq I_j$, $|\tilde{I}_j|<\infty$, then $A_i\in\bigcup_{i\in I_j}\mathcal{A}_i\subseteq\sigma(\bigcup_{i\in I_j}\mathcal{A}_i)$ $\forall\, i\in\tilde{I}_j \underset{\sigma\text{-algebra}}{\Rightarrow} \bigcap_{i\in\tilde{I}_j}A_i\in\sigma(\bigcup_{i\in I_j}\mathcal{A}_i) \underset{\text{def. } \mathcal{A}_{I_j}}{\Rightarrow} \mathcal{A}_{I_j}\subseteq\sigma(\bigcup_{i\in I_j}\mathcal{A}_i) \underset{\text{smallest}}{\Rightarrow}$ $\sigma(\mathcal{A}_{I_j})\subseteq\sigma(\bigcup_{i\in I_j}\mathcal{A}_i)$. □

4.2.2 Zero-one laws

A *zero-one law* is a result that states that an event must have probability in $\{0,1\}$.

Theorem 4.9 (Borel–Cantelli (BC) lemmas)

1) If $\{A_i\}_{i\in\mathbb{N}}\subseteq\mathcal{F}$, $\sum_{i=1}^{\infty}\mathbb{P}(A_i)<\infty \Rightarrow \mathbb{P}(A_n \text{ io})=0$ (BC1).
2) If $\{A_i\}_{i\in\mathbb{N}}\subseteq\mathcal{F}$ are independent, $\sum_{i=1}^{\infty}\mathbb{P}(A_i)=\infty \Rightarrow \mathbb{P}(A_n \text{ io})=1$ (BC2).

In particular, for independent $\{A_i\}_{i\in\mathbb{N}}\subseteq\mathcal{F}$, $\mathbb{P}(A_n \text{ io})\in\{0,1\}$.

Proof.

1) $\mathbb{P}(A_n \text{ io}) \underset{\text{L. 1.12)}}{=} \mathbb{P}(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty}A_k)\le\mathbb{P}(\bigcup_{k=n}^{\infty}A_k) \underset{\sigma\text{-subadd.}}{\le} \sum_{k=n}^{\infty}\mathbb{P}(A_k) \underset{n\to\infty}{\overset{\text{ass.}}{\to}} 0.$

2) By assumption, we must have $\sum_{k=n}^{\infty} \mathbb{P}(A_k) = \infty \;\forall\, n \in \mathbb{N}$ and thus

$$\mathbb{P}(A_n \text{ io}) \underset{\text{L. 1.1 2)}}{=} \mathbb{P}(\limsup_{n\to\infty} A_n) = 1 - \mathbb{P}((\limsup_{n\to\infty} A_n)^c) \underset{\text{L. 1.2 2)}}{=} 1 - \mathbb{P}(\liminf_{n\to\infty} A_n^c)$$

$$\underset{\text{L. 1.3 2)}}{=} 1 - \mathbb{P}(\lim_{n\to\infty} \inf_{k\ge n} A_k^c) \underset{\text{def.}}{=} 1 - \mathbb{P}\Big(\lim_{n\to\infty} \bigcap_{k=n}^{\infty} A_k^c\Big)$$

$$\underset{\text{below}}{\overset{\text{cont.}}{=}} 1 - \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{k=n}^{\infty} A_k^c\Big) \underset{\text{L. 4.7}}{=} 1 - \lim_{n\to\infty} \prod_{k=n}^{\infty} \underbrace{\mathbb{P}(A_k^c)}_{=1-\mathbb{P}(A_k) \underset{1-x\le e^{-x}}{\le} e^{-\mathbb{P}(A_k)}}$$

$$\ge 1 - \lim_{n\to\infty} \underbrace{e^{-\sum_{k=n}^{\infty} \mathbb{P}(A_k)}}_{=0\;\forall\, n} = 1. \qquad \square$$

Definition 4.10 (Tail σ-algebra, tail events)
Let $(\Omega, \mathcal{F})$ be a measurable space and $\mathcal{A}_n \subseteq \mathcal{F}$, $n \in \mathbb{N}$, π-systems. Then $\mathcal{T} := \bigcap_{n=1}^{\infty} \sigma(\bigcup_{k=n}^{\infty} \mathcal{A}_k) = \bigcap_{n=1}^{\infty} \sigma(\mathcal{A}_n, \mathcal{A}_{n+1}, \dots)$ is the *tail σ-algebra* and its elements are *tail events*.

Theorem 4.11 (Kolmogorov's zero-one law)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{A}_n \subseteq \mathcal{F}$, $n \in \mathbb{N}$, be independent π-systems. Then $\mathbb{P}(A) \in \{0, 1\} \;\forall\, A \in \mathcal{T}$.

Proof. Let $\mathcal{T}_n := \sigma(\bigcup_{k=1}^{n-1} \mathcal{A}_k)$, $n \in \mathbb{N} \cup \{\infty\}$.

i) By L. 4.8 and ass., $\mathcal{T}_n = \sigma(\bigcup_{k=1}^{n-1} \mathcal{A}_k)$ and $\sigma(\bigcup_{k=n}^{\infty} \mathcal{A}_k)$ are ind. σ-algebras $\forall\, n$.

ii) $\mathcal{T} \underset{\text{def.}}{=} \bigcap_{n=1}^{\infty} \sigma(\bigcup_{k=n}^{\infty} \mathcal{A}_k) \subseteq \sigma(\bigcup_{k=n}^{\infty} \mathcal{A}_k)\ \forall\, n \in \mathbb{N}$. $\overset{\text{i)}}{\underset{\text{L. 4.7}}{\Rightarrow}}$ $\mathcal{T}_n, \mathcal{T}$ are independent $\forall\, n \in \mathbb{N} \underset{\text{def.}}{\Rightarrow} \bigcup_{n=1}^{\infty} \mathcal{T}_n, \mathcal{T}$ are independent.

iii) $\bigcup_{n=1}^{\infty} \mathcal{T}_n$ is a π-system since $A_1, A_2 \in \bigcup_{n=1}^{\infty} \mathcal{T}_n \Rightarrow A_i \in \mathcal{T}_{n_i}$, $i = 1, 2 \underset{\mathcal{T}_n \nearrow}{\Rightarrow}$ $A_1, A_2 \in \mathcal{T}_{\max\{n_1,n_2\}} \underset{\mathcal{T}_{\max\{n_1,n_2\}}\ \sigma\text{-alg.}}{\Rightarrow} A_1 \cap A_2 \in \mathcal{T}_{\max\{n_1,n_2\}} \subseteq \bigcup_{n=1}^{\infty} \mathcal{T}_n$.

iv) $\overset{\text{ii)}}{\underset{\text{L. 4.7}}{\Rightarrow}}$ $\sigma(\bigcup_{n=1}^{\infty} \mathcal{T}_n), \mathcal{T}$ are independent.

v) $\sigma(\bigcup_{n=1}^{\infty} \mathcal{T}_n) \underset{\mathcal{A}_n \subseteq \mathcal{T}_{n+1}}{\supseteq} \sigma(\bigcup_{n=1}^{\infty} \mathcal{A}_n) \underset{\text{def.}}{=} \mathcal{T}_\infty$. And $\mathcal{T}_n \subseteq \mathcal{T}_\infty \Rightarrow \bigcup_{n=1}^{\infty} \mathcal{T}_n \subseteq \mathcal{T}_\infty \Rightarrow$ $\sigma(\bigcup_{n=1}^{\infty} \mathcal{T}_n) \subseteq \mathcal{T}_\infty$. Hence $\sigma(\bigcup_{n=1}^{\infty} \mathcal{T}_n) = \mathcal{T}_\infty \underset{\text{iv)}}{\Rightarrow} \mathcal{T}_\infty, \mathcal{T}$ are independent.

vi) Since $\mathcal{T} \underset{\text{def.}}{=} \bigcap_{n=1}^{\infty} \sigma(\bigcup_{k=n}^{\infty} \mathcal{A}_k) \underset{n=1}{\subseteq} \sigma(\bigcup_{k=1}^{\infty} \mathcal{A}_k) = \mathcal{T}_\infty$, every $A \in \mathcal{T}$ is independent of itself $\Rightarrow \mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A)^2 \Rightarrow \mathbb{P}(A) \in \{0, 1\}$. □

- Kolmogorov's zero-one law is often applied to independent events ($\mathcal{A}_n = \{A_n\}$, $n \in \mathbb{N}$) or rvs ($\mathcal{A}_n = \sigma(X_n)$, $n \in \mathbb{N}$).

- In case of rvs, note that $\sigma(\bigcup_{k=n}^{\infty}\sigma(X_k)) \underset{\text{def.}}{=} \sigma(X_k,\ k\ge n)$. So

$$\mathcal{T} \underset{\text{def.}}{=} \bigcap_{n=1}^{\infty}\sigma\Big(\bigcup_{k=n}^{\infty}\sigma(X_k)\Big) = \bigcap_{n=1}^{\infty}\sigma(X_n, X_{n+1}, \dots). \tag{2}$$

Example 4.12

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent rvs.

1) $A := \{\omega\in\Omega : \sum_{k=1}^{\infty} X_k(\omega) \text{ converges}\} = \{\omega\in\Omega : \sum_{k=n}^{\infty} X_k(\omega) \text{ converges}\} \underset{\text{can show}}{\in} \sigma(X_n, X_{n+1}, \dots)\ \forall\, n\in\mathbb{N} \underset{\text{def.}}{\Rightarrow} A \in \bigcap_{n=1}^{\infty}\sigma(X_n, X_{n+1}, \dots) \underset{(2)}{=} \mathcal{T}$
$\underset{\text{ind.}}{\Rightarrow} \mathbb{P}(\sum_{n=1}^{\infty} X_n \text{ converges}) \underset{\text{K. 0-1}}{\in} \{0,1\}$.

2) $A := \{\omega\in\Omega : \lim_{n\to\infty}\bar{X}_n(\omega) = c\in\mathbb{R}\} = \Big\{\omega\in\Omega : \lim_{n\to\infty}\frac{\sum_{k=m}^{n} X_k(\omega)}{n} = c\in\mathbb{R}\Big\} \underset{\text{can show}}{\in} \sigma(X_m, X_{m+1}, \dots)\ \forall\, m\in\mathbb{N} \underset{\text{def.}}{\Rightarrow} A\in\bigcap_{m=1}^{\infty}\sigma(X_m, X_{m+1}, \dots) \underset{(2)}{=} \mathcal{T}$
$\underset{\text{ind.}}{\Rightarrow} \mathbb{P}(\lim_{n\to\infty}\bar{X}_n = c\in\mathbb{R}) \underset{\text{K. 0-1}}{\in} \{0,1\}$.

4.2.3 Independence of random variables

Question: Is there an easier way to check independence of rvs than by definition?

Theorem 4.13 (Characterization of independence of rvs)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X_i : \Omega \to \mathbb{R}$, $i \in I$, be rvs. Then X_i, $i \in I$, are independent iff $\forall\, n \in \mathbb{N}$, $\{i_1, \dots, i_n\} \subseteq I$, $\mathbb{P}(X_{i_1} \le x_{i_1}, \dots, X_{i_n} \le x_{i_n}) = \prod_{k=1}^n \mathbb{P}(X_{i_k} \le x_{i_k})$, $\forall\, x_{i_1}, \dots, x_{i_n} \in \mathbb{R}$.

Proof.

"$\Rightarrow$": X_i, $i \in I$, ind. $\underset{\text{def.}}{\Rightarrow}$ $\sigma(X_i)$, $i \in I$, ind. $\underset{\text{def. }\sigma()}{\Rightarrow}$ $X_{i_1}^{-1}((-\infty, x_{i_1}]), \dots, X_{i_n}^{-1}((-\infty, x_{i_n}])$ are ind. $\forall x_{i_1}, \dots, x_{i_n} \in \mathbb{R}$ $\underset{\text{def.}}{\Rightarrow}$ $\mathbb{P}(X_{i_1} \le x_{i_1}, \dots, X_{i_n} \le x_{i_n}) \underset{\text{def.}}{=} \mathbb{P}(\bigcap_{k=1}^n \{X_{i_k} \le x_{i_k}\}) \underset{\text{def.}}{=} \mathbb{P}(\bigcap_{k=1}^n X_{i_k}^{-1}((-\infty, x_{i_k}])) \underset{\text{ind.}}{=} \prod_{k=1}^n \mathbb{P}(X_{i_k}^{-1}((-\infty, x_{i_k}])) \underset{\text{def.}}{=} \prod_{k=1}^n \mathbb{P}(X_{i_k} \le x_{i_k})$ $\forall x_{i_1}, \dots, x_{i_n} \in \mathbb{R}$.

"$\Leftarrow$": As preimages are closed under intersection, $\mathcal{A}_{i_k} := \{X_{i_k}^{-1}((-\infty, x])\}_{x \in \mathbb{R}}$, $k = 1, \dots, n$, are π-systems $\underset{\text{ass.}}{\Rightarrow}$ $\mathcal{A}_{i_1}, \dots, \mathcal{A}_{i_n}$ are ind. $\underset{\text{L. 4.7}}{\Rightarrow}$ $\sigma(\mathcal{A}_{i_1}), \dots, \sigma(\mathcal{A}_{i_n})$ are ind. Now $\sigma(\mathcal{A}_{i_k}) = \sigma(X_{i_k}^{-1}(\{(-\infty, x]\}_{x \in \mathbb{R}})) \underset{\text{L. 3.4}}{=} X_{i_k}^{-1}(\sigma(\{(-\infty, x]\}_{x \in \mathbb{R}})) \underset{\text{R. 2.24 1)}}{=} X_{i_k}^{-1}(\mathcal{B}(\mathbb{R})) \underset{\text{def.}}{=} \sigma(X_{i_k})$, $k = 1, \dots, n$ $\underset{\text{def.}}{\Rightarrow}$ $X_{i_1}, \dots, X_{i_n}$ are ind. $\square$

Corollary 4.14 (Equivalent ways to check independence)

1) Rvs $X_1, \dots, X_d$ with $\boldsymbol{X} = (X_1, \dots, X_d) \sim F$ and margins $F_1, \dots, F_d$ are independent iff

$$F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}) \underset{\text{T. 4.13}}{=} \prod_{j=1}^{d} \mathbb{P}(X_j \le x_j) = \prod_{j=1}^{d} F_j(x_j), \quad \boldsymbol{x} \in \mathbb{R}^d.$$

2) If F admits continuous partial derivatives with respect to each component once in $\operatorname{supp}(F)$, then F is absolutely continuous with density $f(\boldsymbol{x}) = \frac{\partial^d}{\partial x_d \dots \partial x_1} F(\boldsymbol{x})$ and $X_1, \dots, X_d$ are independent iff

$$f(\boldsymbol{x}) = \frac{\partial^d}{\partial x_d \dots \partial x_1} F(\boldsymbol{x}) \underset{1)}{=} \prod_{j=1}^{d} \frac{\partial}{\partial x_j} F_j(x_j) = \prod_{j=1}^{d} f_j(x_j), \quad \boldsymbol{x} \in \operatorname{supp}(F).$$

3) Similarly, for discrete $\boldsymbol{X} = (X_1, \dots, X_d)$ with $\operatorname{supp}(F_j) = \{x_{j,1}, x_{j,2}, \dots\}$, $X_1, \dots, X_d$ are independent iff $f(\boldsymbol{x}) = \prod_{j=1}^{d} f_j(x_j)$, $\boldsymbol{x} \in \prod_{j=1}^{d} \operatorname{supp}(F_j)$.

4) When using $\mathcal{A}'_{i_k} = \{X_{i_k}^{-1}((x, \infty))\}_{x \in \mathbb{R}}$ in the proof of T. 4.13, we see that $X_1, \dots, X_d$ are independent iff $\bar{F}(\boldsymbol{x}) = \prod_{j=1}^{d} \bar{F}_j(x_j)$, $\boldsymbol{x} \in \mathbb{R}^d$; one can also see "$\Rightarrow$" from L. 3.18 by induction.

Question: Are functions of independent rvs independent?

Proposition 4.15 (Functions of independent rvs are independent)
If $X_{i,j}$, $i \in \mathbb{N}$, $j = 1, \dots, d_i$, are ind. rvs and $h_i : \mathbb{R}^{d_i} \to \mathbb{R}$ are $(\mathcal{B}(\mathbb{R}^{d_i}), \mathcal{B}(\mathbb{R}))$-measurable, then $Y_i = h_i(X_{i,1}, \dots, X_{i,d_i})$, $i \in \mathbb{N}$, are independent rvs.

Proof.

1) $X_{i,j}$, $i \in \mathbb{N}$, $j = 1, \dots, d_i$, independent $\underset{\text{def.}}{\Rightarrow} \sigma(X_{i,j})$, $i \in \mathbb{N}$, $j = 1, \dots, d_i$, are independent σ-algebras $\underset{\text{L. 4.8}}{\Rightarrow} \sigma(\bigcup_{j=1}^{d_i} \sigma(X_{i,j}))$, $i \in \mathbb{N}$, are independent.
2) For $\boldsymbol{X}_i = (X_{i,1}, \dots, X_{i,d_i})$, it is straightforward to check that $\mathcal{F}_i := \{B \in \mathcal{B}(\mathbb{R}^{d_i}) : \boldsymbol{X}_i^{-1}(B) \in \sigma(\bigcup_{j=1}^{d_i} \sigma(X_{i,j}))\}$ is a σ-algebra. Let $B = \prod_{j=1}^{d_i} B_j \in \mathcal{B}(\mathbb{R}^{d_i})$. Then $\boldsymbol{X}_i^{-1}(B) = \bigcap_{j=1}^{d_i} X_{i,j}^{-1}(B_j) \underset{X_{i,j}^{-1}(B_j) \in \sigma(X_{i,j})}{\in} \sigma(\bigcup_{j=1}^{d_i} \sigma(X_{i,j})) \underset{\text{def.}}{\Rightarrow} B \in \mathcal{F}_i$ $\overset{B\text{'s generate}}{\underset{\mathcal{B}(\mathbb{R}^{d_i})}{\Rightarrow}} \mathcal{B}(\mathbb{R}^{d_i}) \subseteq \mathcal{F}_i \underset{\mathcal{F}_i \subseteq \mathcal{B}(\mathbb{R}^{d_i})}{\Rightarrow} \mathcal{F}_i = \mathcal{B}(\mathbb{R}^{d_i})$.
3) Therefore, $\forall\, B \in \mathcal{B}(\mathbb{R}^{d_i})$, we have $\boldsymbol{X}_i^{-1}(B) \in \sigma(\bigcup_{j=1}^{d_i} \sigma(X_{i,j}))$, so $\boldsymbol{X}_i$ is $\big(\sigma(\bigcup_{j=1}^{d_i} \sigma(X_{i,j})), \mathcal{B}(\mathbb{R}^{d_i})\big)$-measurable, $i \in \mathbb{N}$. As a composition, $Y_i = h_i \circ \boldsymbol{X}_i$ is $\big(\sigma(\bigcup_{j=1}^{d_i} \sigma(X_{i,j})), \mathcal{B}(\mathbb{R})\big)$-measurable $\forall\, i \in \mathbb{N}$. Therefore Y_i is a rv.
4) $\sigma(Y_i) \underset{\text{def.}}{=} Y_i^{-1}(\mathcal{B}(\mathbb{R})) \underset{\text{def. } Y_i}{=} \boldsymbol{X}_i^{-1}\big(h_i^{-1}(\mathcal{B}(\mathbb{R}))\big) \underset{\text{mon.}}{\subseteq} \boldsymbol{X}_i^{-1}(\mathcal{B}(\mathbb{R}^{d_i})) \underset{3)}{\subseteq} \sigma(\bigcup_{j=1}^{d_i} \sigma(X_{i,j}))$.
5) By 1), $\sigma(\bigcup_{j=1}^{d_i} \sigma(X_{i,j}))$, $i \in \mathbb{N}$, are independent and $\sigma(Y_i) \underset{4)}{\subseteq} \sigma(\bigcup_{j=1}^{d_i} \sigma(X_{i,j}))$, $i \in \mathbb{N}$, so $\sigma(Y_i)$, $i \in \mathbb{N}$, are independent $\underset{\text{def.}}{\Rightarrow} Y_i$, $i \in \mathbb{N}$, are independent. □

Question: How can we construct independent rvs?

Proposition 4.16 (Construction of independent rvs)
For dfs $F_1, \dots, F_d$, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and independent rvs $X_j \sim F_j$, $j = 1, \dots, d$.

Proof.

- Since $F_1, \dots, F_d$ are dfs, T. 2.48 2) and E. 2.29 imply that $(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \prod_{j=1}^d \lambda_{F_j})$ is a probability space.
- Projections are continuous $\underset{\text{P. 3.7}}{\Rightarrow}$ they are measurable $\Rightarrow X_j(\boldsymbol{\omega}) = \pi_j(\boldsymbol{\omega}) = \omega_j$ are rvs and, by P. 3.10, $\boldsymbol{X}(\boldsymbol{\omega}) = \boldsymbol{\omega}$ is a random vector, $\boldsymbol{\omega} = (\omega_1, \dots, \omega_d) \in \mathbb{R}^d$. Since

$$F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}) = \mathbb{P}(\{\boldsymbol{\omega} \in \Omega : \boldsymbol{\omega} \le \boldsymbol{x}\}) = \mathbb{P}\biggl(\bigcap_{j=1}^d \{\omega_j \in \mathbb{R} : \omega_j \le x_j\}\biggr)$$
$$= \mathbb{P}\biggl(\prod_{j=1}^d (-\infty, x_j]\biggr) \underset{\text{def.}}{=} \prod_{j=1}^d \lambda_{F_j}((-\infty, x_j]) \underset{\text{R. 2.49 1)}}{=} \prod_{j=1}^d F_j(x_j), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

  $X_1, \dots, X_d$ are independent by C. 4.14 1).
- $\forall\, k \ne j$, letting $x_k \to \infty$ implies $\mathbb{P}(X_j \le x_j) = F_j(x_j)$, so $X_j \sim F_j \;\forall\, j$. □

- If $F_1 = \dots = F_d =: F$, one says $X_1, \dots, X_d$ are *independent and identically distributed (iid)* from F or *independent copies of $X \sim F$* (notation: $X_1, \dots, X_d \overset{\text{ind.}}{\sim} F$ or $X_1, \dots, X_d \overset{\text{iid}}{\sim} F$).
- Infinite sequences of independent rvs can be constructed with Kolmogorov's extension theorem.

Example 4.17 (Countable mixtures through conditioning)

- Conditioning can be used to construct flexible distributions from given ones.
- If $I \sim F_I$ with pmf $(p_i)_{i \in \mathbb{N}}$ on $\mathbb{N}$ and $\boldsymbol{X}_i \sim F_i \; \forall\, i \in \mathbb{N}$ are independent, then

$$\boldsymbol{X} = \boldsymbol{X}_I \quad \text{(interpreted as } \boldsymbol{X}_i \text{ if } I = i\text{)} \tag{3}$$

  follows the *mixture (distribution)* of $F_1, F_2, \dots$ wrt F_I.
- The df of $\boldsymbol{X}$ is

$$F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}) = \mathbb{P}(\boldsymbol{X}_I \le \boldsymbol{x}) \underset{\text{prob.}}{\overset{\text{tot.}}{=}} \sum_{i=1}^{\infty} \mathbb{P}(\boldsymbol{X}_I \le \boldsymbol{x} \,|\, I = i)\mathbb{P}(I = i)$$

$$= \sum_{i=1}^{\infty} p_i \mathbb{P}(\boldsymbol{X}_i \le \boldsymbol{x} \,|\, I = i) \underset{\text{ind.}}{=} \sum_{i=1}^{\infty} p_i \mathbb{P}(\boldsymbol{X}_i \le \boldsymbol{x}) = \sum_{i=1}^{\infty} p_i F_i(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d.$$

- If F_i has density $f_i \; \forall\, i \in \mathbb{N}$, then $\boldsymbol{X}$ has density $f(\boldsymbol{x}) = \sum_{i=1}^{\infty} p_i f_i(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$.
- Using $\int$ instead of $\sum$, an extension to uncountable mixtures is possible but requires more work since if, e.g., $I \sim \mathrm{U}(0,1)$, then $\mathbb{P}(I = i) = 0 \; \forall\, i \in I$, so $\mathbb{P}(\boldsymbol{X}_I \le \boldsymbol{x} \,|\, I = i)$ is not defined anymore. Nevertheless, sr (3) is still valid.

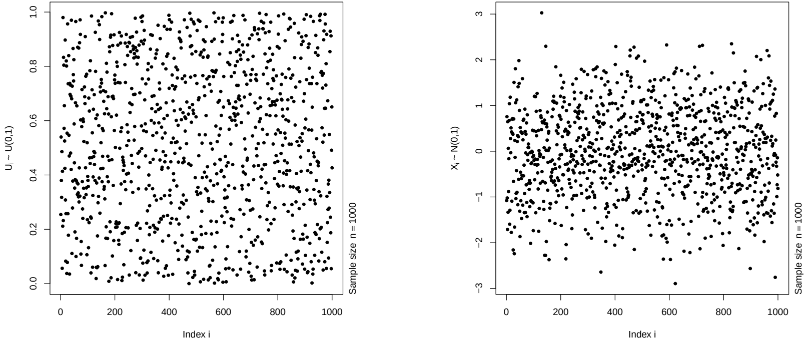
4.2.4 Sampling

- If $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \overset{\text{ind.}}{\sim} F$, then $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ is a *random sample* from F.
- Applications often require to generate realizations of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \overset{\text{ind.}}{\sim} F$, so to *sample* from F or $\boldsymbol{X} \sim F$.
- There are many algorithms known for generating realizations of $U \sim \mathrm{U}(0,1)$ with a computer, so-called *pseudo-random number generators*. They are reproducible once the *seed* is fixed, the first number in the produced sequence.
- By P. 3.30, it suffices to know how to construct realizations from $U \sim \mathrm{U}(0,1)$ and how to evaluate the qf F^{-1} of the df F in order to sample from X. This method is known as *inversion method* for sampling $X \sim F$. Extensions to random vectors $\boldsymbol{X} = (X_1, \dots, X_d)$ are typically based on srs ($\Rightarrow$ realizations $x = (\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_n^\top)^\top \in \mathbb{R}^{n \times d}$).

- There are also other *sampling algorithms* known for how to transform $\mathrm{U}(0,1)$ realizations to realizations of a given distribution.

Example 4.18 ($\mathrm{U}(0,1)$ and $\mathrm{N}(0,1)$ samples)

Realizations $x_1, \dots, x_n$ of $X_1, \dots, X_n$, $n = 1000$, for $\mathrm{U}(0,1)$ (left) and $\mathrm{N}(0,1)$ (right):



There is no visible structure (= uniformity) for the $\mathrm{U}(0,1)$ realizations. The $\mathrm{N}(0,1)$ realizations are dispersed around the location $\mu = 0$. The larger σ (here: $\sigma = 1$), the larger the dispersion. The probability transform can be used for assessing whether a sample has a continuous F as df.

Example 4.19 (Sampling of edfs)

- An edf F_n can be written as

$$F_n(\boldsymbol{x}) = \sum_{i=1}^n \frac{1}{n}\mathbb{1}_{[\boldsymbol{X}_i,\infty)}(\boldsymbol{x}) = \sum_{i=1}^n p_i F_i(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

  for $p_i = \frac{1}{n}$, $i = 1, \dots, n$, (the pmf of $\mathrm{U}(\{1, \dots, n\})$) and $F_i(\boldsymbol{x}) = \mathbb{1}_{[\boldsymbol{X}_i,\infty)}(\boldsymbol{x})$ (a *degenerate* df corresponding to the point mass at $\boldsymbol{X}_i$). Therefore, F_n is a mixture which puts mass $1/n$ on each $\boldsymbol{X}_i$.

- With sr $I = \lceil nU \rceil \sim \mathrm{U}(\{1, \dots, n\})$ for $U \sim \mathrm{U}(0, 1)$, we have the sr

$$\boldsymbol{X}_{\lceil nU \rceil} \underset{(3)}{\sim} F_n.$$

  We can thus sample F_n by first drawing $I \sim \mathrm{U}(\{1, \dots, n\})$ and then returning $\boldsymbol{X}_I$, so by randomly drawing from $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$.

- This idea is applied in the "bootstrap".

- It remains valid even if there are *ties* (i.e. equal realizations) among $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$.

4.3 Dependence

Question: How can we quantify dependence between $X_1, \dots, X_d$, $d \ge 2$?

Definition 4.20 (Copula)
A d-dimensional *copula* C is a d-dimensional df with $\mathrm{U}(0,1)$ margins.

Proposition 4.21 (Characterization of copulas)
A function $C : [0,1]^d \to [0,1]$ is a d-dimensional copula iff

1) $C(\boldsymbol{u}) = 0$ if $\exists\, j \in \{1, \dots, d\}$ such that $u_j = 0$ (*groundedness*);
2) $C(1, \dots, 1, u_j, 1, \dots, 1) = u_j$, $u_j \in [0,1]$, $\forall\, j$ ($\mathrm{U}(0,1)$ margins); and
3) $\Delta_{(\boldsymbol{a},\boldsymbol{b}]} C \ge 0 \;\forall\, \boldsymbol{0} \le \boldsymbol{a} \le \boldsymbol{b} \le \boldsymbol{1}$ (*d-increasingness*).

- As dfs with support in $[0,1]^d$ (instead of $\mathbb{R}^d$), 0 has now the role of $-\infty$ (lower endpoint of the support) and 1 that of ∞ (upper endpoint of the support); see R. 2.49 5) for how to extend C to $\mathbb{R}^d$.
- In practice we never find data with perfect $\mathrm{U}(0,1)$ margins. However, we will see later why copulas are useful nonetheless.

Example 4.22 (Fundamental copulas and srs)

1) $C(\boldsymbol{u}) = \prod_{j=1}^d u_j$, $\boldsymbol{u} \in [0,1]^d$, is a copula (the *independence copula*) since $\boldsymbol{U} \sim C$ for $\boldsymbol{U} = (U_1, \dots, U_d)$ with $U_1, \dots, U_d \overset{\text{ind.}}{\sim} \mathrm{U}(0,1)$.

   Proof. $\mathbb{P}(\boldsymbol{U} \le \boldsymbol{u}) \underset{\text{T. 4.13}}{=} \prod_{j=1}^d \mathbb{P}(U_j \le u_j) = \prod_{j=1}^d u_j$, $\boldsymbol{u} \in [0,1]^d$.

2) $C(\boldsymbol{u}) = M(\boldsymbol{u}) := \min\{u_1, \dots, u_d\}$, $\boldsymbol{u} \in [0,1]^d$, is a copula (the *comonotone copula*) since $\boldsymbol{U} \sim C$ for $\boldsymbol{U} = (U, \dots, U)$ for $U \sim \mathrm{U}(0,1)$.

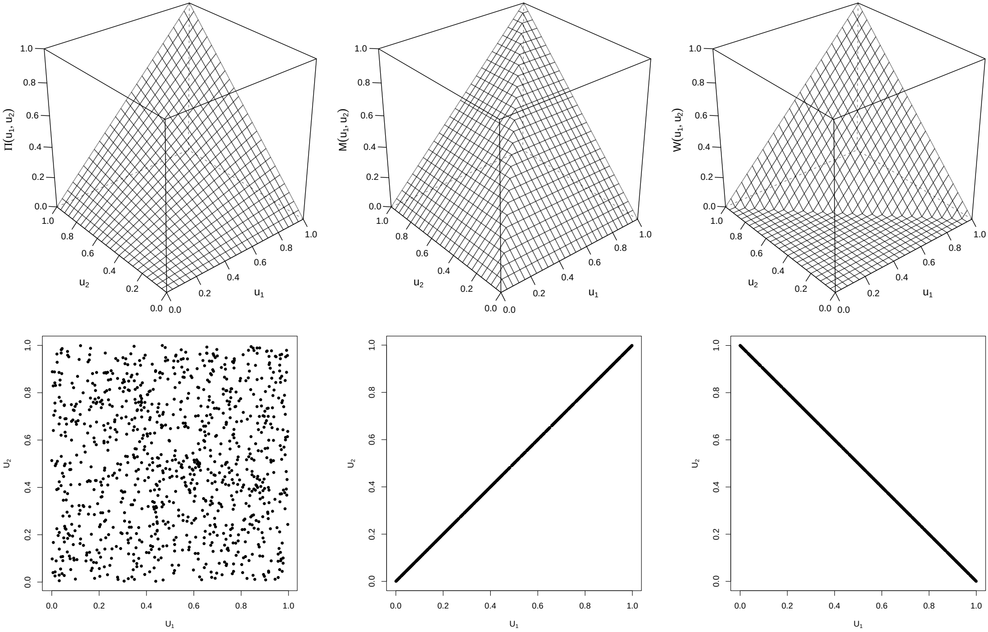
   Proof. $\mathbb{P}(\boldsymbol{U} \le \boldsymbol{u}) = \mathbb{P}(U \le u_1, \dots, U \le u_d) = \mathbb{P}(U \le \min\{u_1, \dots, u_d\}) = \min\{u_1, \dots, u_d\} = M(\boldsymbol{u})$, $\boldsymbol{u} \in [0,1]^d$.

3) $C(\boldsymbol{u}) = W(\boldsymbol{u}) := \max\{(\sum_{j=1}^d u_j) - d + 1, 0\}$, $\boldsymbol{u} \in [0,1]^d$, is a copula (the *countermonotone copula*) for $d = 2$ but not $d \ge 3$.

   Proof. For $d = 2$, $(U, 1-U)$ satisfies $\mathbb{P}(U \le u_1, 1-U \le u_2) = \mathbb{P}(1-u_2 \le U \le u_1) = \max\{u_1 - (1-u_2), 0\} = W(u_1, u_2)$. For $d \ge 3$, see E. 2.41.

   Caution: Never blindly move from $\mathbb{P}$-computations to F-computations, e.g. blindly applying $\mathbb{P}(1-u_2 \le U \le u_1) = u_1 - (1-u_2)$ is only correct if $u_1 \ge 1 - u_2$, otherwise you get $u_1 - (1-u_2) < 0$ ↯.

Copulas Π, M, W (top) and corresponding samples (mass distributions; bottom):



The importance of copulas is due to the following result.

Theorem 4.23 (Sklar (1959))

1) For any df F with margins $F_1, \dots, F_d$, $\exists$ a copula C such that

$$F(\boldsymbol{x}) = C(F_1(x_1), \dots, F_d(x_d)), \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{$*$}$$

   C is uniquely defined on $\prod_{j=1}^d \operatorname{ran}(F_j)$ and there given by $C(\boldsymbol{u}) = F(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))$, $\boldsymbol{u} \in \prod_{j=1}^d \operatorname{ran}(F_j)$.

2) Given any copula C and univariate dfs $F_1, \dots, F_d$, F defined by $(*)$ is a df with margins $F_1, \dots, F_d$.

Proof. We only provide a proof in the case where $F_1, \dots, F_d$ are continuous.

1) Let $\boldsymbol{X} \sim F$ and $\boldsymbol{U} := (F_1(X_1), \dots, F_d(X_d))$. By the probability transform, $\boldsymbol{U}$ has $\mathrm{U}(0,1)$ margins, so $\boldsymbol{U}$ has a copula, say C, as df. Since $X_j \overset{F_j \uparrow \text{ on } \operatorname{supp}(F_j)}{\underset{\text{L. 2.53 1)}}{=}} F_j^{-1}(F_j(X_j)) \underset{\text{def.}}{=} F_j^{-1}(U_j)$, $j = 1, \dots, d$, we have

$$F(\boldsymbol{x}) = \mathbb{P}(X_j \le x_j \ \forall\, j) = \mathbb{P}(F_j^{-1}(U_j) \le x_j \ \forall\, j) \underset{\text{L. 2.53 3)}}{=} \mathbb{P}(U_j \le F_j(x_j) \ \forall\, j)$$
$$= C(F_1(x_1), \dots, F_d(x_d)), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

so C satisfies $(*)$. It is uniquely given by $C(\boldsymbol{u}) \underset{\text{L. 2.53 2)}}{=} C\big(F_1(F_1^{-1}(u_1)), \dots, F_d(F_d^{-1}(u_d))\big) \underset{(*)}{=} F(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))$, $\boldsymbol{u} \in [0,1]^d$.

2) Let $\boldsymbol{U} \sim C$ and $\boldsymbol{X} := (F_1^{-1}(U_1), \dots, F_d^{-1}(U_d))$. Since

$$\begin{aligned}\mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}) &\underset{\text{L. 2.53 3)}}{=} \mathbb{P}(U_1 \le F_1(x_1), \dots, U_d \le F_d(x_d)) \\ &= C(F_1(x_1), \dots, F_d(x_d)), \quad \boldsymbol{x} \in \mathbb{R}^d,\end{aligned}$$

F defined by $(*)$ is a df (namely that of $\boldsymbol{X}$) with margins $F_1, \dots, F_d$ (by the quantile transform). □

Remark 4.24

1) The proofs of both parts are constructive. Part 1) utilizes the probability transform (d-times) and Part 2) the quantile transform.
2) Part 1) allows us to decompose any continuous F into its copula C and the margins $F_1, \dots, F_d$ of F. C is thus the function which contains all the information about the dependence between $X_1, \dots, X_d$ (as it tells us how to combine the margins $F_1, \dots, F_d$ to get the joint df F). This is used in statistical applications (e.g. estimation).

3) Part 2) allows us to construct flexible new multivariate dfs F. This is used in probabilistic applications (e.g. model building, stress testing), e.g. in finance, insurance or risk management (especially under *information asymmetry*).

4) By L. 2.50, $|C(\boldsymbol{b}) - C(\boldsymbol{a})| \le \sum_{j=1}^{d} |b_j - a_j|$, $\boldsymbol{a}, \boldsymbol{b} \in [0,1]^d$, so all copulas are (uniformly equi)continuous, so F is continuous iff $F_1, \dots, F_d$ are. However, if C does not have a density, then F is not absolutely continuous even if all $F_1, \dots, F_d$ are; see E. 3.22, where we chose $C = M$.

5) If F admits continuous partial derivatives with respect to each component once, then C admits a density, given by

$$c(\boldsymbol{u}) = \frac{f(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))}{\prod_{j=1}^{d} f_j(F_j^{-1}(u_j))}, \quad \boldsymbol{u} \in (0,1)^d, \tag{4}$$

where f_j denotes the density of F_j, $j = 1, \dots, d$.

We say that $\boldsymbol{X} \sim F$ *has copula C* if $(*)$ holds, i.e. if $F(\boldsymbol{x}) = C(F_1(x_1), \dots, F_d(x_d))$, $\boldsymbol{x} \in \mathbb{R}^d$, for some margins $F_1, \dots, F_d$.

Theorem 4.25 (Invariance principle)
Let $\boldsymbol{X} = (X_1, \dots, X_d) \sim F$ with continuous margins $F_1, \dots, F_d$ and copula C. If $T_j \uparrow$ on $\operatorname{supp}(F_j)$, $j = 1, \dots, d$, then $(T_1(X_1), \dots, T_d(X_d))$ has unique copula C (so the same copula as that of $\boldsymbol{X}$).

Proof.

- **Idea:** Find the df of $(T_1(X_1), \dots, T_d(X_d))$ and its margins (here: the latter first, to verify continuity), then apply Sklar's theorem.
- As $T_j \uparrow$ on $\operatorname{supp}(F_j)$, T_j has at most countably-many discontinuities. Assume wlog that T_j is right-continuous at its discontinuities (since X_j is continuously distributed, $\mathbb{P}(X_j = x) = 0\ \forall\, x$, so we only change $T_j(X_j)$ on a $\mathbb{P}$-null set).
- By right-continuity of T_j, $T_j(T_j^{-1}(x_j)) \underset{\text{L. 2.53 2)}}{=} x_j\ \forall\, x_j \in \operatorname{ran}(T_j)$. The df G_j of $T_j(X_j)$ is thus

$$\begin{aligned} G_j(x_j) &:= \mathbb{P}(T_j(X_j) \le x_j) \underset{\text{L. 2.53 2)}}{=} \mathbb{P}\big(T_j(X_j) \le T_j(T_j^{-1}(x_j))\big) \\ &\underset{T_j \uparrow \text{ on } \operatorname{supp}(F_j)}{=} \mathbb{P}(X_j \le T_j^{-1}(x_j)) = F_j(T_j^{-1}(x_j)), \quad x_j \in \operatorname{ran}(T_j). \end{aligned}$$

  Since $T_j(X_j)$ puts no mass outside $\operatorname{ran}(T_j)$, $G_j(x_j) = F_j(T_j^{-1}(x_j))$, $x_j \in \mathbb{R}$.

- By L. 2.53 4), T_j^{-1} is continuous on $\operatorname{ran}(T_j)$. And F_j is continuous by assumption. So $G_j = F_j \circ T_j^{-1}$ is continuous on $\operatorname{ran}(T_j)$. By definition, G_j puts no mass outside $\operatorname{supp}(T_j(X_j)) \subseteq \operatorname{ran}(T_j)$, so G_j is continuous on $\mathbb{R}$.
- Then the joint df of $(T_1(X_1), \dots, T_d(X_d))$ is

$$\begin{aligned}
\mathbb{P}(T_j(X_j) \le x_j \ \forall\, j) &\underset{\text{of } T_j(X_j)}{\overset{\text{cont.}}{=}} \mathbb{P}(T_j(X_j) < x_j \ \forall\, j) \underset{\text{L. 2.53 3)}}{\overset{T_j \text{ right-cont.}}{=}} \mathbb{P}(X_j < T_j^{-1}(x_j) \ \forall\, j) \\
&\underset{\text{cont.}}{=} \mathbb{P}(X_j \le T_j^{-1}(x_j) \ \forall\, j) = F(T_1^{-1}(x_1), \dots, T_d^{-1}(x_d)) \\
&\underset{\text{Sklar to } F}{\overset{\text{apply}}{=}} C\big(F_1(T_1^{-1}(x_1)), \dots, F_d(T_d^{-1}(x_d))\big) \\
&\underset{\text{def.}}{=} C(G_1(x_1), \dots, G_d(x_d)), \quad \boldsymbol{x} \in \mathbb{R}^d.
\end{aligned}$$

Since $T_j(X_j) \sim G_j$, $j = 1, \dots, d$, Sklar's theorem implies that $(T_1(X_1), \dots, T_d(X_d))$ has copula C and C is unique since $G_1, \dots, G_d$ are continuous. □

The invariance principle is fundamental to copula modelling, it identifies transformations we are allowed to apply (componentwise) to multivariate data without changing their dependence.

Corollary 4.26

Let $\boldsymbol{X} = (X_1, \dots, X_d)$ with continuous margins $F_1, \dots, F_d$. Then $\boldsymbol{X}$ has unique copula C iff $(F_1(X_1), \dots, F_d(X_d)) \sim C$.

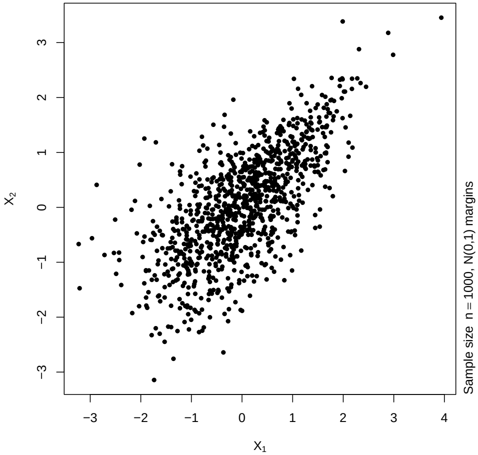
Proof. Let $\boldsymbol{U} := (F_1(X_1), \dots, F_d(X_d))$.

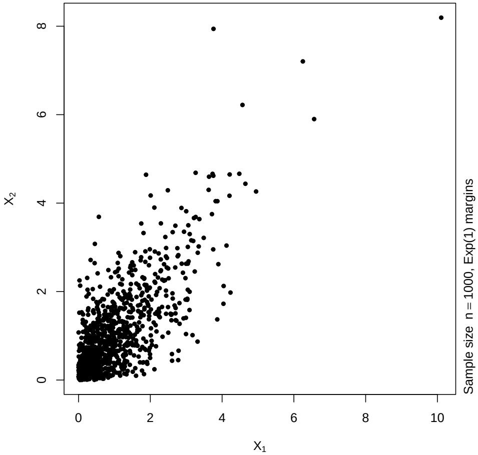
"$\Rightarrow$": Since $F_j \uparrow \mathrm{supp}(F_j)$, $j = 1, \dots, d$, the invariance principle implies that $\boldsymbol{U}$ also has unique copula C. By the probability transform, the margins of $\boldsymbol{U}$ are $\mathrm{U}(0,1)$, so, by Sklar's theorem, the df of $\boldsymbol{U}$ is C.

"$\Leftarrow$": $\boldsymbol{X} \underset{\text{L. 2.53 1)}}{=} \big(F_1^{-1}(F_1(X_1)), \dots, F_d^{-1}(F_d(X_d))\big) \underset{\text{def.}}{=} (F_1^{-1}(U_1), \dots, F_d^{-1}(U_d))$. By assumption, $\boldsymbol{U} \sim C$. And by L. 2.53 4), $F_j^{-1} \uparrow$ on $[0,1]$ (which contains the support of $\mathrm{U}(0,1)$), $j = 1, \dots, d$. By the invariance principle, $\boldsymbol{X}$ thus has unique copula C. □

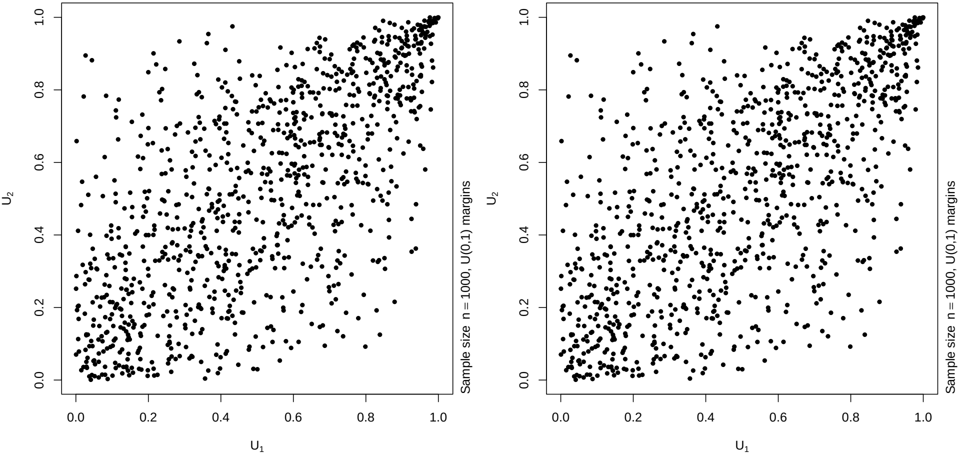
Say we have realizations of $\boldsymbol{X}_i = (X_{i,1}, X_{i,2}) \overset{\text{ind.}}{\sim} F$, $i = 1, \dots, n = 1000$, from two bivariate dfs F with continuous margins F_1, F_2.

Question: For which of the two samples is the dependence between $X_1 \sim F_1$ and $X_2 \sim F_2$ larger?



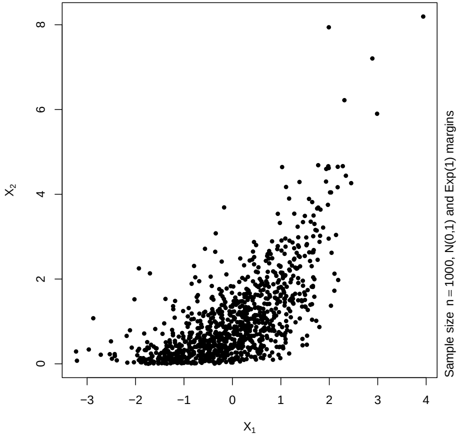


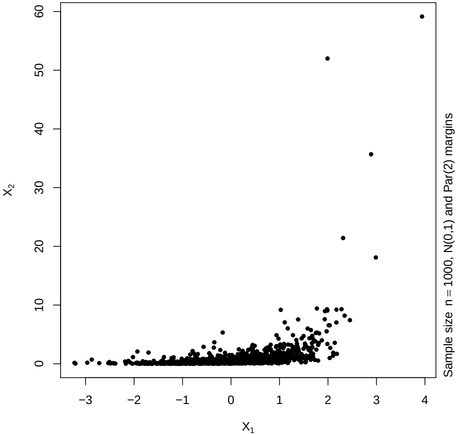
This is not easy to answer as the marginal dfs in the two plots differ. If we apply the respective F_1, F_2 to each of the samples (in applications, use the edfs $F_{n,1}, F_{n,2}$ $\Rightarrow$ componentwise scaled ranks known as *pseudo-observations*), we obtain:



- So the realizations of the two $\boldsymbol{X} \sim F$ only differed in their margins F_1, F_2, their copulas are identical.
- By C. 4.26 (or the proof of Sklar's theorem Part 1)), we indeed see realizations of $\boldsymbol{U} \sim C$, where C is the copula of the two $\boldsymbol{X}$.
- This allows us to study dependence questions in terms of the realizations of $\boldsymbol{U}$.

If we apply $\mathrm{N}(0,1)$ and $\mathrm{Exp}(1)$ (left) and $\mathrm{N}(0,1)$ and $\mathrm{Par}(2)$ (right) qfs ($\mathrm{Exp}(1)$: df $F(x) = 1 - e^{-x}$; $\mathrm{Par}(2)$: df $F(x) = 1 - 1/(1+x)^2$):





- Sklar's theorem Part 2) $\Rightarrow$ We see realizations of $\boldsymbol{X}$ with copula C and the chosen margins.
- Invariance principle $\Rightarrow$ the underlying copula C is the same in the last 6 pictures.

For $(X_1, \dots, X_d) \sim F$ with continuous margins $F_1, \dots, F_d$ and copula C, the invariance principle implies the following meanings of Π, M, W for F:

- $X_1, \dots, X_d$ are independent iff $C = \Pi$.
- $X_j = T_j(X_1)$, $j = 2, \dots, d$, a.s. for the strictly increasing $T_j(x) = F_j^{-1}(F_1(x))$ iff $C = M$.
- $X_2 = T(X_1)$ a.s. for the strictly decreasing $T(x) = F_2^{-1}(1 - F_1(x))$ iff $C = W$.

These are intuitive (but extreme) forms of dependence.

W and M are also extremal bounds analytically as the following result shows. We use it later to derive bounds on dependence-related summary statistics.

Theorem 4.27 (Fréchet–Hoeffding bounds)

1) For any d-dimensional copula C, one has $W(\boldsymbol{u}) \le C(\boldsymbol{u}) \le M(\boldsymbol{u})$, $\boldsymbol{u} \in [0,1]^d$.
2) W is a copula for $d = 2$ but not for $d \ge 3$.
3) M is a copula $\forall\, d \ge 2$.

Proof.

1) $W \leq C$: By L. 2.50, $1 - C(\boldsymbol{u}) = C(\mathbf{1}) - C(\boldsymbol{u}) \underset{\boldsymbol{a}=\boldsymbol{u}}{\overset{\boldsymbol{b}=\mathbf{1}}{\leq}} \sum_{j=1}^{d}(1-u_j) = d - \sum_{j=1}^{d} u_j$, so $C(\boldsymbol{u}) \geq (\sum_{j=1}^{d} u_j) - d + 1$. Clearly, also $C(\boldsymbol{u}) \geq 0$, so $C(\boldsymbol{u}) \geq \max\{(\sum_{j=1}^{d} u_j) - d + 1, 0\}$, which is $W(\boldsymbol{u})$.

$C \leq M$: By L. 2.37 5), any d-increasing F is componentwise increasing. Therefore, $C(\boldsymbol{u}) \leq C(1, u_2, \dots, u_d) \underset{d-2 \text{ times}}{\leq} C(1, \dots, 1, u_j, 1, \dots, 1) \underset{\text{margins}}{\overset{\mathrm{U}(0,1)}{=}} u_j \;\forall\, j = 1, \dots, d$, and thus $C(\boldsymbol{u}) \leq \min\{u_1, \dots, u_d\} = M(\boldsymbol{u})$.

2) By E. 4.22 3), W is a copula for $d = 2$. By E. 2.41, W is not a copula $\forall\, d \geq 3$.

3) By E. 4.22 2), M is a copula $\forall\, d \geq 2$. □

5 Integration and expectation

5.1 Construction

Question: What value do we obtain on average when rolling a fair die?

- $\Omega = \{1, \dots, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$, $A \in \mathcal{F}$ (Laplace probability space).
- Then

$$\sum_{\text{“all outcomes”}} \text{“outcome”} \cdot \text{“probability to obtain this outcome”}$$

$$= \sum_{\omega \in \Omega} \omega \cdot \mathbb{P}(\{\omega\}) = \sum_{i=1}^{6} i \cdot \frac{1}{6} = 3.5 \underset{X(\omega)\,=\,\omega}{=} \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}).$$

- If Ω is uncountable, we could consider (if this exists)

$$\mathbb{E}(X) \underset{\text{notation}}{:=} \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\{\omega\}) \underset{\text{notation}}{=:} \int_\Omega X\,\mathrm{d}\mathbb{P}.$$

Question: Is there a general construction principle for $\mathbb{E}(X)$ for general measures μ (instead of $\mathbb{P}$) that also answers the existence question?

- If $(\Omega, \mathcal{F}, \mu)$ be a measure space and $X : \Omega \to \mathbb{R}$ measurable, one can construct a notion of $\int_\Omega X\,\mathrm{d}\mu$ in a three-step process known as *standard argument* (also *algebraic induction*). Let $\bar{\mathbb{R}}_+ := [0, \infty]$ and $\bar{\mathbb{R}} := [-\infty, \infty]$.

1) One starts to define $\int_\Omega X \,\mathrm{d}\mu$ for simple X.
2) One then extends the notion of $\int_\Omega X \,\mathrm{d}\mu$ to non-negative $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}_+))$-measurable $X : \Omega \to \bar{\mathbb{R}}_+$.
3) Finally, one extends $\int_\Omega X \,\mathrm{d}\mu$ to $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}))$-measurable $X : \Omega \to \bar{\mathbb{R}}$.

Complex-valued X can be handled by their real/imaginary parts.

5.1.1 Expectation of simple rvs

Let $(\Omega, \mathcal{F})$ be a measurable space. Recall that a rv $X : \Omega \to \mathbb{R}$ is *simple* if it has the form $X = \sum_{i=1}^n x_i \mathbb{1}_{A_i}$ for some $n \in \mathbb{N}$, $x_i \in \mathbb{R}$, $i = 1, \dots, n$, and a partition $\{A_i\}_{i=1}^n \subseteq \mathcal{F}$ of Ω; $\{A_i\}_{i=1}^n \subseteq \mathcal{F}$ is required for measurability of X.

Example 5.1 (Examples of simple rvs)
Let $X = \sum_{i=1}^n x_i \mathbb{1}_{A_i}$ and $Y = \sum_{k=1}^m y_k \mathbb{1}_{B_k}$ be simple. Then one can easily verify that the following are simple rvs:

1) $aX + bY = \sum_{i,k} (ax_i + by_k) \mathbb{1}_{A_i \cap B_k}$
2) $XY = \sum_{i,k} (x_i y_k) \mathbb{1}_{A_i \cap B_k}$
3) $\min\{X, Y\} = \sum_{i,k} \min\{x_i, y_k\} \mathbb{1}_{A_i \cap B_k}$, $\max\{X, Y\} = \sum_{i,k} \max\{x_i, y_k\} \mathbb{1}_{A_i \cap B_k}$

Definition 5.2 (Expectation of simple rvs)

Let $X = \sum_{i=1}^n x_i \mathbb{1}_{A_i}$ be a simple rv. Then

$$\mathbb{E}(X) := \int_\Omega X \,\mathrm{d}\mu := \sum_{i=1}^n x_i \mu(A_i)$$

is the *expectation* (or *mean*, *integral*) of X (or its df F) wrt μ with the convention $0 \cdot \infty = 0$ if $x_i = 0$ and $\mu(A_i) = \infty$ for some i. Furthermore, let $\int_A X \,\mathrm{d}\mu := \mathbb{E}(X \mathbb{1}_A)$ $\forall\, A \in \mathcal{F}$.

Lemma 5.3 (Properties of simple rvs)

Let $X, Y : \Omega \to \mathbb{R}$ be simple. Then

1) $X \geq 0$ (pointwise) $\Rightarrow \mathbb{E}(X) \geq 0$ (*non-negativity*);
2) $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ $\forall\, a, b \in \mathbb{R}$ (*linearity*);
3) $X \leq Y \Rightarrow \mathbb{E}(X) \leq \mathbb{E}(Y)$ (*monotonicity*);
4) $X \geq 0 \Rightarrow \nu(A) := \int_A X \,\mathrm{d}\mu$ is a measure on $(\Omega, \mathcal{F})$; and
5) $\mathbb{E}(\mathbb{1}_A) = \mu(A)$, $A \in \mathcal{F}$.

Proof.

1) $X \geq 0$ (pointwise) $\Rightarrow x_i \geq 0,\ i = 1, \dots, n \Rightarrow \mathbb{E}(X) = \sum_{i=1}^n x_i \mu(A_i) \geq 0$.

2) $aX + bY \underset{\text{E. 5.1}}{=} \sum_{i,k} (ax_i + by_k) \mathbb{1}_{A_i \cap B_k}$ is again simple, so

$$\mathbb{E}(aX + bY) \underset{\text{def.}}{=} \sum_{i,k} (ax_i + by_k) \mu(A_i \cap B_k)$$

$$\underset{\text{out}}{\overset{\text{multiply}}{=}} a \sum_{i=1}^n x_i \sum_{k=1}^m \mu(A_i \cap B_k) + b \sum_{k=1}^m y_k \sum_{i=1}^n \mu(A_i \cap B_k)$$

$$\underset{\text{meas.}}{\overset{\text{tot.}}{=}} a \sum_{i=1}^n x_i \mu(A_i) + b \sum_{k=1}^m y_k \mu(B_k) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

3) $X \leq Y \Rightarrow Y - X \geq 0$ is simple $\Rightarrow \mathbb{E}(Y) = \mathbb{E}((Y - X) + X) \underset{2)}{=} \mathbb{E}(Y - X) + \mathbb{E}(X) \underset{1)}{\geq} 0 + \mathbb{E}(X) = \mathbb{E}(X)$.

4) $X \geq 0 \Rightarrow \nu(A) := \int_A X \,\mathrm{d}\mu$ is a measure on $\mathcal{F}$:

i) $X \geq 0 \Rightarrow \nu(A) \underset{1)}{\geq} 0\ \forall A \in \mathcal{F}$, so $\nu : \mathcal{F} \to [0, \infty]$.

ii) $\nu(\emptyset) = \int_\emptyset X \,\mathrm{d}\mu \underset{\text{def.}}{=} \int_\Omega X \mathbb{1}_\emptyset \,\mathrm{d}\mu = \int_\Omega 0 \,\mathrm{d}\mu \underset{\text{simple}}{\overset{\text{def.}}{=}} 0\mu(\Omega) \underset{\text{convention}}{=} 0$.

iii) Let $X = \sum_{j=1}^n x_j \mathbb{1}_{B_j}$ be simple. If $\{A_i\}_{i\in\mathbb{N}} \in \mathcal{F} : A_i \cap A_j = \emptyset \ \forall\, i \neq j$,

$$\nu\Bigl(\biguplus_{i=1}^{\infty} A_i\Bigr) \underset{\text{def.}}{=} \int_{\biguplus_{i=1}^{\infty} A_i} X \,\mathrm{d}\mu \underset{\text{def.}}{=} \mathbb{E}\bigl(X\mathbb{1}_{\biguplus_{i=1}^{\infty} A_i}\bigr) \underset{\text{def. } X}{=} \mathbb{E}\Bigl(\sum_{j=1}^n x_j \mathbb{1}_{B_j\cap\biguplus_{i=1}^{\infty} A_i}\Bigr)$$

$$\underset{\text{def.}}{=} \sum_{j=1}^n x_j \mu\Bigl(B_j \cap \biguplus_{i=1}^{\infty} A_i\Bigr) \underset{\sigma\text{-add.}}{=} \sum_{j=1}^n x_j \sum_{i=1}^{\infty} \mu(B_j \cap A_i)$$

$$\underset{(*)}{=} \sum_{i=1}^{\infty}\sum_{j=1}^n x_j \mu(B_j\cap A_i) \underset{\text{def.}}{=} \sum_{i=1}^{\infty} \mathbb{E}\Bigl(\sum_{j=1}^n x_j \mathbb{1}_{B_j\cap A_i}\Bigr) \underset{\text{line 1}}{\overset{\text{as in}}{=}} \sum_{i=1}^{\infty} \nu(A_i),$$

where $(*)$ holds since the series is absolutely convergent (possibly ∞ in which case $\exists\, i : \nu(A_i) = \infty$, so the claim also holds) and thus, by the Riemann series theorem, any reordering converges to the same value.

5) $\mathbb{E}(\mathbb{1}_A) = \mathbb{E}(1\cdot\mathbb{1}_A + 0\cdot\mathbb{1}_{A^c}) \underset{\text{def.}}{=} 1\cdot\mu(A) + 0\cdot\mu(A^c) = \mu(A) \ \forall\, A \in \mathcal{F}$ with the convention $0\cdot\infty = 0$. □

5.1.2 Expectation of non-negative rvs

We now extend the definition of expectation to non-negative rvs $X : \Omega \to \bar{\mathbb{R}}_+ = [0,\infty]$ (potentially ∞; see R. 2.24 2)).

Lemma 5.4 (Approximating sequence)

Let $(\Omega, \mathcal{F})$ be a measurable space. Then $X : \Omega \to \bar{\mathbb{R}}_+$ is $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}_+))$-measurable iff $\exists$ simple $X_n : \Omega \to \mathbb{R}_+$, $X_n \nearrow X$ pointwise and uniformly on any set on which X is bounded.

Proof. We call any such $(X_n)_{n\in\mathbb{N}}$ an *approximating sequence* to X.

"$\Rightarrow$": Let

$$X_n := \min\left\{\frac{\lfloor 2^n X\rfloor}{2^n}, n\right\} = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbb{1}_{X^{-1}\left(\left(\frac{k-1}{2^n}, \frac{k}{2^n}\right]\right)} + n\mathbb{1}_{X^{-1}(n,\infty]}.$$

By measurability of X, X_n is simple. Also, since $2\lfloor x\rfloor \le \lfloor 2x\rfloor$, one can verify that $0 \le X_n \le X_{n+1}$ $\forall\, n \in \mathbb{N}$.

- If $X \le n$, then $0 \le X - X_n \le \frac{1}{2^n} \underset{n\to\infty}{\to} 0$, so the convergence is uniform in ω (and thus pointwise) on any set on which X is bounded.
- If $X(\omega) = \infty$, then $X_n(\omega) = n \underset{n\to\infty}{\to} \infty$, so pointw. convergence $\forall\, \omega \in \Omega$.

"$\Leftarrow$": Since X_n is simple $\forall\, n \in \mathbb{N}$, we know that X_n is measurable $\forall\, n \in \mathbb{N}$. Similarly as in L. 3.12, one can show that, as a limit, X is measurable. $\square$

Definition 5.5 (Expectation of non-negative rvs)
Let $X : \Omega \to \bar{\mathbb{R}}_+$ be measurable. Then

$$\mathbb{E}(X) := \int_\Omega X \,\mathrm{d}\mu := \sup_{\substack{0 \le Y \le X, \\ Y \text{ simple}}} \mathbb{E}(Y)$$

is the *expectation* (or *mean*, *integral*) of X (or its df F) wrt μ. Furthermore, $\int_A X \,\mathrm{d}\mu := \mathbb{E}(X \mathbb{1}_A) \;\forall\, A \in \mathcal{F}$. The set of all $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}_+))$-measurable $X : \Omega \to \bar{\mathbb{R}}_+$ is denoted by $L_+ := L_+(\Omega, \mathcal{F}, \mu)$.

Lemma 5.6 (Monotonicity, scaling)
Let $X, Y \in L_+$. Then $X \le Y \Rightarrow \mathbb{E}(X) \le \mathbb{E}(Y)$ and $\mathbb{E}(cX) = c\,\mathbb{E}(X) \;\forall\, c \ge 0$.

Proof. Monotonicity is clear by definition (supremum over a larger set cannot be smaller). Now consider scaling. For $c = 0$, scaling is also clear by definition (since $\mathbb{E}(0) \underset{\text{simple}}{=} 0$). And for $c > 0$,

$$\mathbb{E}(cX) = \sup_{\substack{0 \le Y \le cX, \\ Y \text{ simple}}} \mathbb{E}(Y) \overset{\text{subs.}}{\underset{Y = cZ}{=}} \sup_{\substack{0 \le Z \le X, \\ Z \text{ simple}}} \mathbb{E}(cZ) \underset{\text{L. 5.3 2)}}{=} \sup_{\substack{0 \le Z \le X, \\ Z \text{ simple}}} c\,\mathbb{E}(Z) = c\,\mathbb{E}(X). \qquad \square$$

Theorem 5.7 (Monotone convergence (MON))
If $(X_n)_{n\in\mathbb{N}} \subseteq L_+$, $X_n \nearrow X$ (pointwise), then $X \in L_+$ and $\mathbb{E}(X_n) \nearrow \mathbb{E}(X)$.

Proof. We have $X \underset{\text{ass.}}{\geq} X_n \underset{\text{ass.}}{\geq} 0$. Similarly as in L. 3.12, one can show that, as a limit, X is measurable. Therefore, $X \in L_+$. Since $X_n \nearrow \underset{\text{L. 5.6}}{\Rightarrow} \mathbb{E}(X_n) \nearrow \Rightarrow \lim_{n\to\infty} \mathbb{E}(X_n)$ exists (possibly ∞).

"$\leq$": $X_n \underset{\text{ass.}}{\leq} X \underset{\text{L. 5.6}}{\Rightarrow} \mathbb{E}(X_n) \leq \mathbb{E}(X)\ \forall\, n \in \mathbb{N} \Rightarrow \lim_{n\to\infty} \mathbb{E}(X_n) \leq \mathbb{E}(X)$.

"$\geq$": Let Y be simple, $0 \leq Y \leq X$ and $A_n := \{\omega \in \Omega : X_n(\omega) \geq \alpha Y(\omega)\}$ for $\alpha \in (0,1)$. Then A_n is measurable (since $X_n - \alpha Y$ is) and, as $X_n(\omega) \nearrow X(\omega) \underset{\text{constr.}}{\geq} Y(\omega)\ \forall\, \omega \in \Omega$, we have $A_n \nearrow \Omega$. Furthermore, $\mathbb{E}(X_n) = \mathbb{E}(X_n \mathbb{1}_\Omega) \underset{\text{L. 5.6}}{\geq} \mathbb{E}(X_n \mathbb{1}_{A_n}) \underset{\text{L. 5.6}}{\geq} \alpha \mathbb{E}(Y \mathbb{1}_{A_n})\ \forall\, n \in \mathbb{N}$. Then

$$\lim_{n\to\infty} \mathbb{E}(X_n) \overset{\text{just}}{\underset{\text{shown}}{\geq}} \alpha \lim_{n\to\infty} \mathbb{E}(Y \mathbb{1}_{A_n}) \underset{\text{L. 5.3 4)}}{=} \alpha \lim_{n\to\infty} \nu(A_n) \overset{\text{cont. below}}{\underset{A_n \nearrow \Omega}{=}} \alpha \nu(\Omega) \underset{\text{def.}}{=} \alpha \mathbb{E}(Y).$$

For $\alpha \to 1-$, we obtain $\lim_{n\to\infty} \mathbb{E}(X_n) \geq \mathbb{E}(Y) \overset{\sup_Y}{\underset{\text{def. } \mathbb{E}(X)}{\Rightarrow}} \lim_{n\to\infty} \mathbb{E}(X_n) \geq \mathbb{E}(X)$. □

Lemma 5.8 (Properties of non-negative rvs)

Let $X, Y \in L_+$, $(X_n)_{n\in\mathbb{N}} \subseteq L_+$. Then

1) $\mathbb{E}(X) = 0$ iff $X = 0$ a.e.;
2) $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)\ \forall\, a, b \ge 0$ (*linearity*);
3) $\mathbb{E}(\sum_{n=1}^{\infty} X_n) = \sum_{n=1}^{\infty} \mathbb{E}(X_n)$;
4) $X_n \nearrow X$ a.e. $\Rightarrow \lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$;
5) If $\mathbb{E}(X) < \infty$, then $\mu(X = \infty) = 0$, i.e. X is finite a.e.; and
6) $\nu(A) := \int_A X \,\mathrm{d}\mu$ is a measure on $(\Omega, \mathcal{F})$.

Proof.

1) If $X = \sum_{i=1}^{n} x_i \mathbb{1}_{A_i}$ is simple, then $\mathbb{E}(X) = 0 \underset{(*)}{\Leftrightarrow} x_i = 0$ or $\mu(A_i) = 0$ $\forall\, i = 1, \dots, n \Leftrightarrow X = 0$ a.e. Now consider $X \in L_+$.

   "$\Rightarrow$": Let $A_n := \{\omega \in \Omega : X(\omega) \ge 1/n\}$, $n \in \mathbb{N}$. Then $X \ge \frac{1}{n}\mathbb{1}_{A_n}$ and $\mu(\bigcup_{n=1}^{\infty} A_n) = \mu(\{\omega \in \Omega : X(\omega) > 0\})$. Assume $\mu(\bigcup_{n=1}^{\infty} A_n) > 0 \Rightarrow$ $\exists\, n_0 : \mu(A_{n_0}) > 0$ (otherwise $\mu(\bigcup_{n=1}^{\infty} A_n) \le \sum_{n=1}^{\infty} \mu(A_n) = 0$ ↯) $\Rightarrow$ $\mathbb{E}(X) \underset{\text{mon.}}{\ge} \mathbb{E}(\frac{1}{n_0}\mathbb{1}_{A_{n_0}}) \overset{\text{simple}}{\underset{\text{L. 5.6}}{=}} \frac{1}{n_0}\mu(A_{n_0}) > 0$ ↯.

   "$\Leftarrow$": $\forall\, Y$ simple with $0 \le Y \le X \overset{\text{a.e.}}{\underset{\text{ass.}}{=}} 0$, we have $Y = 0$ a.e., and thus by $(*)$ that $\mathbb{E}(Y) = 0$, so $\mathbb{E}(X) = \sup_{0\le Y\le X,\ Y \text{ simple}} \mathbb{E}(Y) = 0$.

2) By scaling (L. 5.6), it suffices to show $\mathbb{E}(X+Y)=\mathbb{E}(X)+\mathbb{E}(Y)$. By L. 5.4, $\exists\,(X_n)_{n\in\mathbb{N}},(Y_n)_{n\in\mathbb{N}}$ simple, non-negative: $X_n\nearrow X$, $Y_n\nearrow Y\Rightarrow(X_n+Y_n)_{n\in\mathbb{N}}$ is simple, non-negative and $X_n+Y_n\nearrow X+Y\Rightarrow\mathbb{E}(X+Y)\underset{\text{MON}}{=}\lim_{n\to\infty}\mathbb{E}(X_n+Y_n)\overset{\text{simple}}{\underset{\text{lin.}}{=}}\lim_{n\to\infty}\mathbb{E}(X_n)+\lim_{n\to\infty}\mathbb{E}(Y_n)\underset{\text{MON}}{=}\mathbb{E}(X)+\mathbb{E}(Y)$.

3) $\mathbb{E}(\sum_{n=1}^\infty X_n)=\mathbb{E}(\lim_{N\to\infty}\sum_{n=1}^N X_n)\underset{\text{MON}}{=}\lim_{N\to\infty}\mathbb{E}(\sum_{n=1}^N X_n)\underset{2)}{=}\lim_{N\to\infty}\sum_{n=1}^N\mathbb{E}(X_n)\underset{\text{def.}}{=}\sum_{n=1}^\infty\mathbb{E}(X_n)$.

4) $X_n(\omega)\nearrow X(\omega)\ \forall\,\omega\in N^c:\mu(N)=0$. Since $X_n-X_n\mathbb{1}_{N^c}=0$ a.e. and $X-X\mathbb{1}_{N^c}=0$ a.e., 1) implies $\mathbb{E}(X_n-X_n\mathbb{1}_{N^c})=0$, $\mathbb{E}(X-X\mathbb{1}_{N^c})=0$. By 2), $\mathbb{E}(X_n)=\mathbb{E}(X_n\mathbb{1}_{N^c})$, $\mathbb{E}(X)=\mathbb{E}(X\mathbb{1}_{N^c})$. Therefore, $\lim_{n\to\infty}\mathbb{E}(X_n)=\lim_{n\to\infty}\mathbb{E}(X_n\mathbb{1}_{N^c})\underset{\text{MON}}{=}\mathbb{E}(X\mathbb{1}_{N^c})=\mathbb{E}(X)$.

5) Let $A_n:=\{X\ge n\}$, $n\in\mathbb{N}\cup\{\infty\}$. Suppose $\mu(A_\infty)>0$, then $\mathbb{E}(X)\underset{\text{mon.}}{\ge}\mathbb{E}(X\mathbb{1}_{A_n})\underset{\text{mon.}}{\ge}\mathbb{E}(n\mathbb{1}_{A_n})\underset{\text{lin.}}{=}n\mathbb{E}(\mathbb{1}_{A_n})\underset{\text{def.}}{=}n\mu(A_n)\overset{A_n\supseteq A_\infty}{\underset{\text{mon.}}{\ge}}n\mu(A_\infty)\underset{n\to\infty}{\to}\infty$ ↯.

6) $\nu(A):=\int_A X\,\mathrm{d}\mu$ is a measure on $(\Omega,\mathcal{F})$: i) $\nu(A)\underset{\text{def.}}{=}\mathbb{E}(X\mathbb{1}_A)\underset{\text{mon.}}{\ge}0\ \forall\,A\in\mathcal{F}$, so $\nu:\mathcal{F}\to[0,\infty]$; ii) $\nu(\emptyset)=\mathbb{E}(X\mathbb{1}_\emptyset)=\mathbb{E}(0)\underset{\text{simple}}{=}0$; and iii) If $\{A_i\}_{i\in\mathbb{N}}\in\mathcal{F}:A_i\cap A_j=\emptyset\ \forall\,i\ne j$, then $\nu(\biguplus_{i=1}^\infty A_i)\underset{\text{def.}}{=}\mathbb{E}(X\mathbb{1}_{\biguplus_{i=1}^\infty A_i})\underset{\text{logic}}{=}\mathbb{E}(\sum_{i=1}^\infty X\mathbb{1}_{A_i})\underset{3)}{=}\sum_{i=1}^\infty\mathbb{E}(X\mathbb{1}_{A_i})\underset{\text{def.}}{=}\sum_{i=1}^\infty\nu(A_i)$. □

5.1.3 Expectation of real-valued, measurable functions

- For $X : \Omega \to \bar{\mathbb{R}}$, let $X^+ := \max\{X, 0\} \geq 0$ be the *positive part* and $X^- := \max\{-X, 0\} \geq 0$ be the *negative part* of X.
- If $X \geq 0$, then $X = X^+$ and $X^- = 0$, and if $X < 0$, then $X^+ = 0$ and $X^- = -X$. Therefore $X = X^+ - X^-$ and $|X| = X^+ + X^-$.
- X is measurable iff X^+, X^- are (both "$\Rightarrow$" and "$\Leftarrow$" follow from C. 3.11 2)).

Definition 5.9 (Quasi-integrable, integrable)
Let $X : \Omega \to \bar{\mathbb{R}}$ be measurable. If $\mathbb{E}(X^+) < \infty$ or $\mathbb{E}(X^-) < \infty$, X is *quasi-integrable* and

$$\mathbb{E}(X) := \int_\Omega X \, \mathrm{d}\mu = \mathbb{E}(X^+) - \mathbb{E}(X^-)$$

is the *expectation* (or *mean*, *integral*) of X (or its df F) wrt μ. Furthermore, $\int_A X \, \mathrm{d}\mu := \mathbb{E}(X \mathbb{1}_A) \; \forall \, A \in \mathcal{F}$. If $\mathbb{E}(X^+) < \infty$ and $\mathbb{E}(X^-) < \infty$, X is *integrable*. If $\mathbb{E}(X^+) = \infty$ and $\mathbb{E}(X^-) = \infty$, X is *non-integrable*. The set of all $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable and integrable $X : \Omega \to \mathbb{R}$ is denoted by $L^1 := L^1(\Omega, \mathcal{F}, \mu)$.

Remark 5.10

1) If the integration variable is important, one also writes $\int_A X(\omega)\,\mathrm{d}\mu(\omega)$. Furthermore, some authors write $\int_A X(\omega)\,\mu(\mathrm{d}\omega)$. If the measure is important to denote, one also writes $\mathbb{E}_\mu(X)$.
2) Examples of quasi-integrable but not integrable X are sufficiently heavy-tailed distributions on $\mathbb{R}_+$ ($\Rightarrow \mathbb{E}(X^-) = 0 < \infty$, $\mathbb{E}(X^+) = \infty$), e.g. $X \sim \mathrm{Par}(\theta)$ with $F(x) = 1 - (1+x)^{-\theta}$, $x \geq 0$, $\theta \in (0,1]$.
3) Examples of non-integrable X are two-sided heavy-tailed distributions, e.g. the Cauchy distribution with density $f(x) = \frac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$.
4) Since $|X| = X^+ + X^-$, we have that X is integrable iff $|X|$ is integrable as a non-negative rv, so iff $\mathbb{E}(|X|) < \infty$. We thus have

$$L^1 = \{X : \Omega \to \mathbb{R} : X \text{ measurable, } \mathbb{E}(|X|) < \infty\}.$$

We will see from L. 5.13 below that integrable rvs are finite a.e., hence the definition of L^1 excludes the possiblity that $\mu(X = \pm\infty) > 0$.

Lemma 5.11 (Properties of quasi-integrable rvs)

Let X, Y be quasi-integrable.

1) If $\mathbb{E}(X^-) < \infty$, $\mathbb{E}(Y^-) < \infty$ or if $\mathbb{E}(X^+) < \infty$, $\mathbb{E}(Y^+) < \infty$, then $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ (*additivity*).
2) $\mathbb{E}(cX) = c\mathbb{E}(X)$, $c \in \mathbb{R}$ (*scaling*).
3) $X \geq 0 \Rightarrow \mathbb{E}(X) \geq 0$ (*non-negativity*).
4) $X \leq Y \Rightarrow \mathbb{E}(X) \leq \mathbb{E}(Y)$ (*monotonicity*).
5) $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$ (*Δ-inequality*).
6) If $\mu(A) = 0$, then $\int_A X \,\mathrm{d}\mu = 0$.

Proof.

1) $Z := X + Y \Rightarrow Z^+ - Z^- \underset{\text{def.}}{=} Z \underset{\text{def.}}{=} X + Y \underset{\text{def.}}{=} X^+ - X^- + Y^+ - Y^- \underset{\text{add}}{\Rightarrow} Z^+ + X^- + Y^- = Z^- + X^+ + Y^+ \underset{\text{L. 5.8 2)}}{\Rightarrow} \mathbb{E}(Z^+) + \mathbb{E}(X^-) + \mathbb{E}(Y^-) \underset{(*)}{=} \mathbb{E}(Z^-) + \mathbb{E}(X^+) + \mathbb{E}(Y^+)$. Wlog assume the case $\mathbb{E}(X^-) < \infty$, $\mathbb{E}(Y^-) < \infty \overset{Z^- \leq X^- + Y^-}{\underset{\text{mon.}}{\Rightarrow}} \mathbb{E}(Z^-) \underset{\text{lin.}}{\leq} \mathbb{E}(X^-) + \mathbb{E}(Y^-) < \infty$ and we can thus subtract the expectations of the three negative parts from $(*)$ to obtain $\mathbb{E}(Z) \underset{\text{def.}}{=} \mathbb{E}(Z^+) - \mathbb{E}(Z^-) \underset{(*)}{=} \mathbb{E}(X^+) - \mathbb{E}(X^-) + \mathbb{E}(Y^+) - \mathbb{E}(Y^-) \underset{\text{def.}}{=} \mathbb{E}(X) + \mathbb{E}(Y)$.

2) If $c \geq 0$, then $\mathbb{E}(cX) \underset{\text{def.}}{=} \mathbb{E}((cX)^+) - \mathbb{E}((cX)^-) \underset{\text{logic}}{=} \mathbb{E}(cX^+) - \mathbb{E}(cX^-) \underset{\text{lin.}}{=} c\mathbb{E}(X^+) - c\mathbb{E}(X^-) \underset{\text{def.}}{=} c\mathbb{E}(X)$. And if $c < 0$, then $\mathbb{E}(cX) \underset{\text{def.}}{=} \mathbb{E}((cX)^+) - \mathbb{E}((cX)^-) \underset{\text{logic}}{=} \mathbb{E}(-cX^-) - \mathbb{E}(-cX^+) \underset{\text{lin.}}{=} -c\mathbb{E}(X^-) - (-c)\mathbb{E}(X^+) \underset{\text{def.}}{=} c\mathbb{E}(X)$.

3) $X \geq 0 \Rightarrow \mathbb{E}(X) \underset{X = X^+}{=} \mathbb{E}(X^+) \underset{\text{mon.}}{\geq} 0$.

4) i) By ass., X is quasi-integrable.

ii) $Y - X \geq 0 \Rightarrow \mathbb{E}((Y - X)^-) = 0 < \infty \Rightarrow Y - X$ quasi-integrable.

$\Rightarrow \mathbb{E}(Y) = \mathbb{E}((Y - X) + X) \overset{\text{ii),iii)}}{\underset{1)}{=}} \mathbb{E}(Y - X) + \mathbb{E}(X) \underset{3)}{\geq} \mathbb{E}(X)$.

5) $|\mathbb{E}(X)| \underset{\text{def.}}{=} |\mathbb{E}(X^+) - \mathbb{E}(X^-)| \underset{\Delta}{\leq} |\mathbb{E}(X^+)| + |\mathbb{E}(X^-)| = \mathbb{E}(X^+) + \mathbb{E}(X^-) \underset{\text{lin.}}{=} \mathbb{E}(X^+ + X^-) = \mathbb{E}(|X|)$.

6) i) First consider $X \geq 0$, so $X\mathbb{1}_A \geq 0$. Any simple $Y = \sum_{i=1}^n y_i \mathbb{1}_{B_i} \geq 0$ with $Y \leq X\mathbb{1}_A$ must satisfy $y_i \mathbb{1}_{B_i} \leq X\mathbb{1}_A$, $i = 1, \dots, n$. For $i = 1, \dots, n$, we thus have $B_i \subseteq A$ (and thus $0 \leq \mu(B_i) \underset{\text{mon.}}{\leq} \mu(A) \underset{\text{ass.}}{=} 0$) or, if not, then we must have $y_i = 0$. Hence $\mathbb{E}(Y) = \sum_{i=1}^n y_i \mu(B_i) = 0$.

ii) By definition, $\int_A X \,\mathrm{d}\mu = \mathbb{E}(X\mathbb{1}_A) = \sup_{\substack{0 \leq Y \leq X, \\ Y \text{ simple}}} \mathbb{E}(Y) \underset{\text{i)}}{=} 0$.

iii) For general X, $\mathbb{E}(X\mathbb{1}_A) \underset{\text{def.}}{=} \mathbb{E}(X^+\mathbb{1}_A) - \mathbb{E}(X^-\mathbb{1}_A) \underset{\text{ii)}}{=} 0 - 0 = 0$. □

The following is needed for conditional expectations later.

Lemma 5.12 (σ-additivity for expectations)
Let $A, \{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F} : A = \biguplus_{i=1}^{\infty} A_i$ and let Z be a rv such that $Z\mathbb{1}_A \in L^1$, then $\mathbb{E}(Z\mathbb{1}_A) = \sum_{i=1}^{\infty} \mathbb{E}(Z\mathbb{1}_{A_i})$.

Proof. $\mathbb{E}(|Z|\mathbb{1}_A) = \mathbb{E}(|Z\mathbb{1}_A|) \underset{Z\mathbb{1}_A \in L^1}{<} \infty$ and so $\mathbb{E}(Z^+\mathbb{1}_A) \le \mathbb{E}(|Z|\mathbb{1}_A) < \infty$ and $\mathbb{E}(Z^-\mathbb{1}_A) \le \mathbb{E}(|Z|\mathbb{1}_A) < \infty$. Therefore, $\mathbb{E}(Z\mathbb{1}_A) = \mathbb{E}((Z^+ - Z^-)\mathbb{1}_A) \underset{\text{lin.}}{=} \mathbb{E}(Z^+\mathbb{1}_A) - \mathbb{E}(Z^-\mathbb{1}_A)$. By L. 5.8 6), $\nu(B) = \mathbb{E}(Z^+\mathbb{1}_B)$, $B \in \mathcal{F}$, is a measure, so σ-additivity implies

$$\mathbb{E}(Z^+\mathbb{1}_A) = \nu(A) = \nu\Bigl(\biguplus_{i=1}^{\infty} A_i\Bigr) = \sum_{i=1}^{\infty} \nu(A_i) = \sum_{i=1}^{\infty} \mathbb{E}(Z^+\mathbb{1}_{A_i}).$$

Similarly for $\mathbb{E}(Z^-\mathbb{1}_A)$. We thus obtain that

$$\mathbb{E}(Z\mathbb{1}_A) \underset{\text{lin.}}{=} \mathbb{E}(Z^+\mathbb{1}_A) - \mathbb{E}(Z^-\mathbb{1}_A) \overset{\text{as}}{\underset{\text{shown}}{=}} \sum_{i=1}^{\infty} \mathbb{E}(Z^+\mathbb{1}_{A_i}) - \sum_{i=1}^{\infty} \mathbb{E}(Z^-\mathbb{1}_{A_i})$$

$$\overset{\text{both}}{\underset{\text{finite}}{=}} \sum_{i=1}^{\infty} (\mathbb{E}(Z^+\mathbb{1}_{A_i}) - \mathbb{E}(Z^-\mathbb{1}_{A_i})) \underset{\text{lin.}}{=} \sum_{i=1}^{\infty} \mathbb{E}((Z^+ - Z^-)\mathbb{1}_{A_i}) = \sum_{i=1}^{\infty} \mathbb{E}(Z\mathbb{1}_{A_i}) \quad \square$$

Lemma 5.13 (Integrable rvs are finite a.e.)
If X is integrable, then $\mu(X = \pm\infty) = 0$, i.e. X is finite a.e..

Proof. X integrable $\Rightarrow \mathbb{E}(X^+) < \infty$, $\mathbb{E}(X^-) < \infty \underset{\text{L. 5.8 5)}}{\Rightarrow} \mu(X^+ = \infty) = \mu(X^- = \infty) = 0 \Rightarrow \mu(X = \pm\infty) = \mu(\{X^+ = \infty\} \uplus \{X^- = \infty\}) \underset{\text{add.}}{=} \mu(X^+ = \infty) + \mu(X^- = \infty) = 0 + 0 = 0.$ □

Proposition 5.14 (Properties of integrable rvs)
If X, Y are integrable, the following are equivalent:

1) $\mathbb{E}(X\mathbb{1}_A) = \mathbb{E}(Y\mathbb{1}_A)\ \forall\, A \in \mathcal{F}$
2) $\mathbb{E}(|X - Y|) = 0$
3) $X = Y$ a.e.

Proof.

1) $\Rightarrow$ 3): Let $Z := X - Y$ and assume $\mu(Z \neq 0) > 0 \Rightarrow \mu(A) > 0$ for $A = \{Z^+ > 0\}$ or $A = \{Z^- > 0\}$, wlog the former. Then $0 \underset{\text{ass.}}{=} \mathbb{E}(X\mathbb{1}_A) - \mathbb{E}(Y\mathbb{1}_A) \underset{\text{lin.}}{=} \mathbb{E}(Z\mathbb{1}_A) \underset{Z\mathbb{1}_A = Z^+\mathbb{1}_A}{=} \mathbb{E}(Z^+\mathbb{1}_A) \underset{\text{L. 5.8 1)}}{\Rightarrow} Z^+\mathbb{1}_A = 0$ a.e., a contradiction to $Z^+ > 0$ on A with $\mu(A) > 0$.

3) $\Rightarrow$ 2): Apply L. 5.8 1) to $Z := |X - Y| \in L_+$.

2) $\Rightarrow$ 1): $|\mathbb{E}(X\mathbb{1}_A) - \mathbb{E}(Y\mathbb{1}_A)| \underset{\text{lin.}}{=} |\mathbb{E}((X-Y)\mathbb{1}_A)| \underset{\text{L. 5.11 5)}}{\leq} \mathbb{E}(|X-Y|\mathbb{1}_A) \leq \mathbb{E}(|X - Y|) \underset{\text{ass.}}{=} 0 \ \forall A \in \mathcal{F}$. □

Lemma 5.15 (Fatou's lemma)

If $(X_n)_{n\in\mathbb{N}} \subseteq L_+$, then

$$\mathbb{E}(\liminf_{n\to\infty} X_n) \leq \liminf_{n\to\infty} \mathbb{E}(X_n).$$

Proof. We use from analysis that, pointwise, $\liminf_{n\to\infty} X_n = \lim_{n\to\infty} \inf_{k\geq n} X_k$. We thus have

$$0 = \mathbb{E}(0) \underset{\text{mon.}}{\leq} \mathbb{E}(\liminf_{n\to\infty} X_n) \underset{\text{analysis}}{=} \mathbb{E}(\lim_{n\to\infty} \inf_{k\geq n} X_k) \overset{\text{MON}}{\underset{\inf \nearrow}{=}} \lim_{n\to\infty} \mathbb{E}(\inf_{k\geq n} X_k)$$

$$\underset{\mathbb{E}(\cdot)\nearrow}{=} \liminf_{n\to\infty} \mathbb{E}(\inf_{k\geq n} X_k) \underset{\text{mon.}}{\leq} \liminf_{n\to\infty} \mathbb{E}(X_n).$$

□

Theorem 5.16 (Dominated convergence (DOM))

If $(X_n)_{n\in\mathbb{N}} \subseteq L^1$, $X_n \underset{n\to\infty}{\overset{\text{a.e.}}{\to}} X$, $|X_n| \overset{\text{a.e.}}{\le} Y \;\forall\, n \in \mathbb{N}$ for $Y \in L^1$, then $X \in L^1$ and $\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$.

Proof.

- By ass., $\exists\, N_X \in \mathcal{F} : \mu(N_X) = 0$, $\lim_{n\to\infty} X_n(\omega) = X(\omega) \;\forall\, \omega \in N_X^c$.
- By ass., $\exists\, N_{X_n,Y} \in \mathcal{F} : \mu(N_{X_n,Y}) = 0$, $|X_n(\omega)| \le Y(\omega) \underset{\text{L. 5.13}}{<} \infty \;\forall\, \omega \in N_{X_n,Y}^c$.
- Let $N := N_X \cup N_{X_n,Y} \in \mathcal{F}$ and, wlog, redefine $X := (\lim_{n\to\infty} X_n)\mathbb{1}_{N^c}$ (a.e. equal to the original $X = \lim_{n\to\infty} X_n$). As a composition, X is measurable.
- On N^c, $|X_n| \le Y < \infty \;\forall\, n \in \mathbb{N}$, so that $|X| \le Y < \infty$ and thus $X \in L^1$, as well as $Y + X_n \ge 0$ and $Y - X_n \ge 0$.

1) $\mathbb{E}(Y)+\mathbb{E}(X) \underset{\text{lin.}}{=} \mathbb{E}(Y+X) \underset{\text{on } N^c}{\overset{\liminf_{n\to\infty} X_n = X}{=}} \mathbb{E}(\liminf_{n\to\infty}(Y+X_n)) \underset{\text{Fatou}}{\le} \liminf_{n\to\infty} \mathbb{E}(Y + X_n) \underset{\text{lin.}}{=} \mathbb{E}(Y) + \liminf_{n\to\infty} \mathbb{E}(X_n)$; and

2) $\mathbb{E}(Y)-\mathbb{E}(X) \underset{\text{lin.}}{=} \mathbb{E}(Y-X) \underset{\text{on } N^c}{\overset{\liminf_{n\to\infty} X_n = X}{=}} \mathbb{E}(\liminf_{n\to\infty}(Y-X_n)) \underset{\text{Fatou}}{\le} \liminf_{n\to\infty} \mathbb{E}(Y - X_n) \underset{\text{lin.}}{=} \mathbb{E}(Y) - \limsup_{n\to\infty} \mathbb{E}(X_n)$.

$\underset{-\mathbb{E}(Y)}{\Rightarrow} \liminf_{n\to\infty} \mathbb{E}(X_n) \underset{1)}{\ge} \mathbb{E}(X) \underset{2)}{\ge} \limsup_{n\to\infty} \mathbb{E}(X_n) \ge \liminf_{n\to\infty} \mathbb{E}(X_n)$. $\square$

Under DOM, we actually have the even stronger statement

$$\lim_{n\to\infty} \mathbb{E}(|X - X_n|) = 0.$$

To see this, apply DOM to $|X - X_n| \underset{n\to\infty}{\overset{\text{a.e.}}{\to}} 0$ and note that $|X - X_n|$ is dominated by $|X - X_n| \underset{\Delta}{\leq} |X| + |X_n| \overset{\text{a.e.}}{\leq} |X| + |Y| \in L^1$.

Example 5.17 (Why domination is required)

Consider $(\Omega, \mathcal{F}, \mu) = ([0,1], \mathcal{B}([0,1]), \lambda)$. Let $X_n := n^2 \mathbb{1}_{(0,1/n)} \geq 0$, $n \in \mathbb{N}$, (so X_n explodes on smaller and smaller set) and $X := 0$. Then $X_n \to X$ pointwise, but

$$\mathbb{E}(X_n) = n^2 \mathbb{E}(\mathbb{1}_{(0,1/n)}) = \frac{n^2}{n} = n \underset{n\to\infty}{\to} \infty \neq 0 = \mathbb{E}(X).$$

Note that Fatou's lemma still holds, since

$$\mathbb{E}(\liminf_{n\to\infty} X_n) = \mathbb{E}(0) \underset{\text{simple}}{=} 0 \leq \infty = \liminf_{n\to\infty} \mathbb{E}(X_n).$$

Corollary 5.18 (Expectation of infinite series)

Suppose $(X_n)_{n\in\mathbb{N}} \subseteq L^1 : \sum_{n=1}^{\infty}\mathbb{E}(|X_n|) < \infty$. Then $\sum_{n=1}^{\infty} X_n$ converges a.e., $\sum_{n=1}^{\infty} X_n \in L^1$ and

$$\mathbb{E}\Bigl(\sum_{n=1}^{\infty} X_n\Bigr) = \sum_{n=1}^{\infty}\mathbb{E}(X_n).$$

Proof.

- $\mathbb{E}(\sum_{n=1}^{\infty}|X_n|) \underset{\text{L. 5.8 3)}}{=} \sum_{n=1}^{\infty}\mathbb{E}(|X_n|) \underset{\text{ass.}}{<} \infty \Rightarrow Y := \sum_{n=1}^{\infty}|X_n| \in L^1$. By L. 5.13, $\exists\, N \in \mathcal{F} : \mu(N) = 0,\ Y(\omega) < \infty\ \forall\,\omega \in N^c$. Therefore, $\forall\,\omega \in N^c$, $\sum_{n=1}^{\infty} X_n(\omega)$ converges abs. and thus converges, so $\sum_{n=1}^{\infty} X_n$ converges a.e.
- Clearly, $(\sum_{n=1}^{N} X_n) \subseteq L^1$ and $\sum_{n=1}^{N} X_n \underset{N\to\infty}{\to} \sum_{n=1}^{\infty} X_n$ everywhere ($\Rightarrow$ 1st and 2nd ass. of DOM ✓)
- Domination holds, since $|\sum_{n=1}^{N} X_n| \underset{\triangle}{\le} \sum_{n=1}^{N}|X_n| \le Y \in L^1\ \forall\, N \in \mathbb{N}$ ($\Rightarrow$ 3rd ass. of DOM ✓).
- Applying DOM, we obtain that $\sum_{n=1}^{\infty} X_n \in L^1$ and

$$\mathbb{E}\Bigl(\sum_{n=1}^{\infty} X_n\Bigr) \underset{\text{DOM}}{=} \lim_{N\to\infty}\mathbb{E}\Bigl(\sum_{n=1}^{N} X_n\Bigr) \underset{\text{lin.}}{=} \lim_{N\to\infty}\sum_{n=1}^{N}\mathbb{E}(X_n) \underset{\text{def.}}{=} \sum_{n=1}^{\infty}\mathbb{E}(X_n). \qquad \square$$

5.1.4 L^p spaces

- For $p \in (0,\infty)$, let $L^p := L^p(\Omega,\mathcal{F},\mu)$ with
  $$L^p(\Omega,\mathcal{F},\mu) := \{X:\Omega\to\mathbb{R} : X \text{ measurable, } \|X\|_p := \mathbb{E}(|X|^p)^{1/p} < \infty\};$$
  note that $X \in L^p$ iff $|X|^p \in L^1$
- If $|X|^p \in L^1$, $\mathbb{E}(X^p)$ is the *pth moment* of X.
- For $p \ge 1$, L^p is a Banach space (complete vector space with norm $\|X\|_p = \mathbb{E}(|X|^p)^{1/p}$; complete = every Cauchy sequence converges in the space). L^2 is a Hilbert space (vector space with inner product $\langle X,Y\rangle = \mathbb{E}(XY)$ that is complete wrt the distance function $d(X,Y) = \|X-Y\|$ induced by $\|X\| = \sqrt{\langle X,X\rangle}$).
- For integration, it makes no difference if we alter measurable functions on null sets (see P. 5.14), e.g. by defining them to be 0 there. One therefore typically views two $X,Y \in L^p$ as equivalent iff $X = Y$ a.e..
- For $p=\infty$, one defines $\|X\|_\infty := \operatorname{ess\,sup}_\omega |X(\omega)| := \inf\{x \ge 0 : \mu(|X| > x) = 0\}$ ('essentially the supremum') and $L^\infty(\Omega,\mathcal{F},\mu) := \{X : \Omega \to \mathbb{R} : X \text{ measurable, } \|X\|_\infty < \infty\}$. One can show: $X \in L^\infty$ iff X is bounded a.e.
- Important inequalities in L^p spaces are:

- *Hölder's inequality*: For $p, q \in [1, \infty] : \frac{1}{p} + \frac{1}{q} = 1$ (conjugate indices), $\|XY\|_1 \le \|X\|_p \|Y\|_q$ for all measurable $X, Y : \Omega \to \mathbb{R}$. For $p = q = 2$, this is known as the *Cauchy–Schwarz inequality (CSI)*. If $p, q \in (1, \infty)$ and $X \in L^p$, $Y \in L^q$, then "$=$" iff $\exists\, \alpha, \beta \ge 0 : \alpha |X|^p = \beta |Y|^q$ μ-a.e..
- *Minkowski's inequality*: $\|X + Y\|_p \le \|X\|_p + \|Y\|_p \;\forall\, X, Y \in L^p$, $p \in [1, \infty]$.
- *Jensen's inequality*: If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $X \in L^1$ and φ convex (concave), then $\varphi(\mathbb{E}(X)) \le \mathbb{E}(\varphi(X))$ $(\varphi(\mathbb{E}(X)) \ge \mathbb{E}(\varphi(X)))$.

  Proof. Wlog suppose φ is convex (otherwise consider $-\varphi$). For any $\mu \in \mathbb{R}$ there is a supporting line $\varphi(\mu) + m(x - \mu)$ of the graph of φ in μ (if φ is differentiable in μ, take $m = \varphi'(\mu)$) with $\varphi(x) \underset{\varphi \text{ convex}}{\ge} \varphi(\mu) + m(x - \mu)$, $x \in \mathbb{R}$. With $\mu = \mathbb{E}(X)$, we thus have that $\varphi(x) \ge \varphi(\mathbb{E}(X)) + m(x - \mathbb{E}(X))$, $x \in \mathbb{R}$, and thus $\varphi(X) \ge \varphi(\mathbb{E}(X)) + m(X - \mathbb{E}(X))$. Since $\mathbb{E}(\mathbb{E}(X)) = \mathbb{E}(X)$, we have $\mathbb{E}(\varphi(X)) \underset{\text{mon.}}{\ge} \mathbb{E}\big(\varphi(\mathbb{E}(X)) + m(X - \mathbb{E}(X))\big) \underset{\text{lin.}}{=} \varphi(\mathbb{E}(X)) + 0$. □

  The convex $\varphi(x) = |x|$ implies $|\mathbb{E}(X)| \le \mathbb{E}(|X|)$ (Δ-inequality), and the concave $\varphi(x) = x^{\frac{p}{q}}$, $x > 0$, $0 < p < q < \infty$ implies $\mathbb{E}(|X|^p) = \mathbb{E}((|X|^q)^{\frac{p}{q}}) \underset{\text{Jensen}}{\le} (\mathbb{E}(|X|^q))^{\frac{p}{q}} < \infty$, so $\mathbb{E}(|X|^q) < \infty \Rightarrow \mathbb{E}(|X|^p) < \infty$.

$\mathbb{E}(|X|^q) < \infty \Rightarrow \mathbb{E}(|X|^p) < \infty$ also holds for general finite μ, as we now prove.

Proposition 5.19 ($L^q \subseteq L^p$, $p < q$)

If $\mu(\Omega) < \infty$ and $0 < p < q \le \infty$, then $L^q \subseteq L^p$ and $\|X\|_p \le \|X\|_q \cdot \mu(\Omega)^{\frac{1}{p}-\frac{1}{q}}$.

Proof. If $q = \infty$, $\|X\|_p^p = \mathbb{E}(|X|^p) \underset{\text{mon.}}{\le} \mathbb{E}(\|X\|_\infty^p) \le \|X\|_\infty^p \mathbb{E}(1) = \|X\|_\infty^p \mu(\Omega)$ $\underset{(\cdot)^{1/p}}{\Rightarrow} \|X\|_p \le \|X\|_\infty \mu(\Omega)^{1/p}$ ✓. And if $q < \infty$, then apply Hölder's inequality with $p \leftarrow \frac{q}{p} \ge 1$ and $q \leftarrow \frac{q}{q-p} \ge 1$ (conjugate ✓) to get $\|X\|_p^p = \mathbb{E}(|X|^p \cdot 1) \underset{\text{Hölder}}{\le}$ $\||X|^p\|_{\frac{q}{p}} \cdot \|1\|_{\frac{q}{q-p}} = \mathbb{E}(|X|^q)^{\frac{p}{q}} \mu(\Omega)^{\frac{q-p}{q}} = \|X\|_q^p \mu(\Omega)^{1-\frac{p}{q}} \underset{(\cdot)^{1/p}}{\Rightarrow} \|X\|_p \le \|X\|_q \cdot \mu(\Omega)^{\frac{1}{p}-\frac{1}{q}}$. □

Remark 5.20 (Integrability of complex-valued measurable functions)

- $X : \Omega \to \mathbb{C}$ is *measurable* if it is measurable wrt to $\mathcal{B}(\mathbb{C}) \underset{z = a + bi}{=} \mathcal{B}(\mathbb{R}^2) \underset{\text{R. 2.24 1)}}{=} \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$.
- A measurable $X : \Omega \to \mathbb{C}$ is *integrable* if $\mathbb{E}(|\operatorname{Re}(X)|) < \infty$ and $\mathbb{E}(|\operatorname{Im}(X)|) < \infty$ and then one defines $\mathbb{E}(X) := \mathbb{E}(\operatorname{Re}(X)) + i\mathbb{E}(\operatorname{Im}(X))$.
- Since $|X| = |\operatorname{Re}(X) + i\operatorname{Im}(X)| \underset{\Delta\text{-ineq.}}{\le} |\operatorname{Re}(X)| + |\operatorname{Im}(X)| \le |X| + |X| = 2|X|$, $X : \Omega \to \mathbb{C}$ is integrable iff $\mathbb{E}(|X|) < \infty$, just as for $X : \Omega \to \mathbb{R}$.

5.2 Calculating expectations

Although important for deriving general results, we rarely calculate expectations of X over Ω via the standard argument, but rather of F or f over $\mathbb{R}^d$, $d \ge 1$. The following result allows one to do that.

Theorem 5.21 (Change of variables)
Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, $(\Omega', \mathcal{F}')$ be a measurable space, $X : \Omega \to \Omega'$ measurable and $h : \Omega' \to \mathbb{R}$ measurable. If $\mathbb{E}(|h(X)|) < \infty$, then

$$\int_\Omega h(X)\,\mathrm{d}\mu = \int_{\Omega'} h\,\mathrm{d}\mu_X$$

(short: $\mathbb{E}_\mu(h(X)) = \mathbb{E}_{\mu_X}(h)$; verbose: $\int_\Omega h(X(\omega))\,\mathrm{d}\mu(\omega) = \int_{\Omega'} h(\omega')\,\mathrm{d}\mu_X(\omega')$).

Proof. We follow the standard argument, applied to h (not: $h(X)$).

1) Consider $h(\omega') = \mathbb{1}_{A'}(\omega')$, $A' \in \mathcal{F}'$, and note that $\mathbb{1}_{A'}(X(\omega)) \underset{(*)}{=} \mathbb{1}_{X^{-1}(A')}(\omega)$ since $\mathbb{1}_{A'}(X(\omega)) = 1$ iff $X(\omega) \in A'$ iff $\omega \in X^{-1}(A')$ iff $\mathbb{1}_{X^{-1}(A')}(\omega) = 1$. Therefore

$$\int_\Omega h(X(\omega))\,\mathrm{d}\mu(\omega) \underset{\text{def. } h}{=} \int_\Omega \mathbb{1}_{A'}(X(\omega))\,\mathrm{d}\mu(\omega) \underset{(*)}{=} \int_\Omega \mathbb{1}_{X^{-1}(A')}(\omega)\,\mathrm{d}\mu(\omega)$$

$$\underset{\text{def.}}{=} \mathbb{E}_\mu(\mathbb{1}_{X^{-1}(A')}) \underset{\text{L. 5.3 5)}}{=} \mu(X^{-1}(A')) \underset{\text{def.}}{=} \mu_X(A') \overset{\text{P. 3.14}}{\underset{\text{L. 5.3 5)}}{=}} \mathbb{E}_{\mu_X}(\mathbb{1}_{A'})$$

$$\underset{\text{def.}}{=} \int_{\Omega'} \mathbb{1}_{A'}(\omega')\,\mathrm{d}\mu_X(\omega') \underset{\text{def. } h}{=} \int_{\Omega'} h(\omega')\,\mathrm{d}\mu_X(\omega').$$

By linearity of $\int$ (see L. 5.11 1), 2)), the statement also holds for all simple h.

2) If $h \in L_+$, $\exists$ simple $h_n : \Omega' \to \mathbb{R}_+$ with $h_n \nearrow h$ (see L. 5.4), so

$$\mathbb{E}_\mu(h(X)) = \mathbb{E}_\mu(\lim_{n\to\infty} h_n(X)) \underset{\text{MON}}{=} \lim_{n\to\infty} \mathbb{E}_\mu(h_n(X)) \underset{1)}{=} \lim_{n\to\infty} \mathbb{E}_{\mu_X}(h_n)$$

$$\underset{\text{MON}}{=} \mathbb{E}_{\mu_X}(\lim_{n\to\infty} h_n) = \mathbb{E}_{\mu_X}(h).$$

3) If $h \in L^1$, then $\mathbb{E}_\mu(h(X)) \underset{\text{def.}}{=} \mathbb{E}_\mu(h^+(X)) - \mathbb{E}_\mu(h^-(X)) \underset{2)}{=} \mathbb{E}_{\mu_X}(h^+) - \mathbb{E}_{\mu_X}(h^-) \underset{\text{lin.}}{=} \mathbb{E}_{\mu_X}(h^+ - h^-) \underset{\text{def.}}{=} \mathbb{E}_{\mu_X}(h)$. □

Remark 5.22

1) We typically consider the case where $\mu = \mathbb{P}$ is a probability measure, $(\Omega', \mathcal{F}') = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $d \geq 1$, and $h : \mathbb{R}^d \to \mathbb{R}$. Due to $\mathbb{P}_{\boldsymbol{X}}$ being characterized by the df F of $\boldsymbol{X}$ (see R. 3.16 1)), one writes $\mathrm{d}F$ for $\mathrm{d}\mathbb{P}_{\boldsymbol{X}}$ (see R. 3.16 4)), so

$$\mathbb{E}(h(\boldsymbol{X})) \underset{\text{def.}}{=} \int_\Omega h(\boldsymbol{X})\,\mathrm{d}\mathbb{P} \underset{\text{T. 5.21}}{=} \int_{\mathbb{R}^d} h\,\mathrm{d}\mathbb{P}_{\boldsymbol{X}} = \int_{\mathbb{R}^d} h\,\mathrm{d}F \underset{\text{verbose}}{=} \int_{\mathbb{R}^d} h(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x});$$

in particular, if $\boldsymbol{X}_j \sim F_j$, $j = 1, 2$, then $F_1 = F_2 \Rightarrow \mathbb{E}(h(\boldsymbol{X}_1)) = \mathbb{E}(h(\boldsymbol{X}_2))$. $\int_{\mathbb{R}^d} h(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x})$ is also known as *Lebesgue–Stieltjes integral* of h wrt F.

2) With the standard argument, one can also show the following two special cases:
   - If F is discrete with pmf f with countable support $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}} \subseteq \mathbb{R}^d$ (so $F(B) = \sum_{i:\boldsymbol{x}_i\in B} f(\boldsymbol{x}_i)$, $B \in \mathcal{B}(\mathbb{R}^d)$), then
$$\int_\Omega h(\boldsymbol{X})\,\mathrm{d}\mathbb{P} = \sum_{i\in\mathbb{N}} h(\boldsymbol{x}_i) f(\boldsymbol{x}_i).$$
   - If F is absolutely continuous with density f (so $F(B) = \int_B f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, $B \in \mathcal{B}(\mathbb{R}^d)$), then
$$\int_\Omega h(\boldsymbol{X})\,\mathrm{d}\mathbb{P} = \int_{\mathbb{R}^d} h(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

   This gives us a way of interpreting $\int_{\mathbb{R}^d} h(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x})$ in the most common cases.

3) The integral $\int_{[\boldsymbol{a},\boldsymbol{b}]} h(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x})$ resembles a *Riemann–Stieltjes integral* with *integrand* $h : [\boldsymbol{a}, \boldsymbol{b}] \to \mathbb{R}$ and *integrator* $F : [\boldsymbol{a}, \boldsymbol{b}] \to \mathbb{R}$ defined as the limit of Riemann–Stieltjes sums
$$\int_{[\boldsymbol{a},\boldsymbol{b}]} h(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x}) := \lim_{\boldsymbol{n}\to\infty} \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} h(\boldsymbol{\xi_i})\Delta_{(\boldsymbol{x_{i-1}},\boldsymbol{x_i}]}F$$

$$= \lim_{n_1,\dots,n_d\to\infty} \sum_{i_d=1}^{n_d} \cdots \sum_{i_1=1}^{n_1} h(\xi_{1,i_1},\dots,\xi_{d,i_d}) \Delta_{\left(\begin{pmatrix} x_{1,i_1-1} \\ \vdots \\ x_{d,i_d-1}\end{pmatrix}, \begin{pmatrix} x_{1,i_1} \\ \vdots \\ x_{d,i_d}\end{pmatrix}\right]} F,$$

where $a_j = x_{j,0} < x_{j,1} < \dots < x_{j,n} = b_j$, $\xi_{j,i_j} \in [x_{j,i_j-1}, x_{j,i_j}]$, $\forall\, j = 1,\dots,d$.

One can show (see ter Horst (1984)):

- If $h : [\boldsymbol{a},\boldsymbol{b}] \to \mathbb{R}$ is Riemann–Stieltjes integrable wrt F on $[\boldsymbol{a},\boldsymbol{b}]$, then h is Lebesgue–Stieltjes integrable wrt λ_F on $[\boldsymbol{a},\boldsymbol{b}]$ and $\int_{[\boldsymbol{a},\boldsymbol{b}]} h \,\mathrm{d}\lambda_F = \int_{[\boldsymbol{a},\boldsymbol{b}]} h \,\mathrm{d}F$. This allows one to compute a Lebesgue–Stieltjes integral as a Riemann–Stieltjes integrals if the latter exists, in particular to compute $\int_{[\boldsymbol{a},\boldsymbol{b}]} h \,\mathrm{d}F$ as Riemann integral if F has density f.
- If $h : [\boldsymbol{a},\boldsymbol{b}] \to \mathbb{R}$ is bounded, then h is Riemann–Stieltjes integrable wrt F on $[\boldsymbol{a},\boldsymbol{b}]$ iff h is continuous λ_F-a.e. on $[\boldsymbol{a},\boldsymbol{b}]$.
- **Example:** $h(x) = \mathbb{1}_{\mathbb{Q}}(x)$, $x \in [0,1]$, is discontinuous on $[0,1]$ (with $\lambda([0,1]) = 1 > 0$), so h is not Riemann-integrable (also clear by definition). But it is Lebesgue integrable with value 0 (by L. 5.11 6)).

4) A change of variables formula for transformations is $\int_{\mathbb{R}^d} h(\boldsymbol{x}) \,\mathrm{d}F(\boldsymbol{x}) \underset{\text{def.}}{=} \mathbb{E}(h(\boldsymbol{X})) \underset{Y := h(\boldsymbol{X}) \sim F_Y}{\overset{\text{subs.}}{=}} \mathbb{E}(Y) \underset{\text{def.}}{=} \int_{\mathbb{R}} y \,\mathrm{d}F_Y(y)$.

Question: How can we compute multivariate (Lebesgue–Stieltjes) integrals?

Theorem 5.23 (Fubini–Tonelli)

Consider the product space $(\Omega, \mathcal{F}, \mu) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mu_1 \times \mu_2)$ of the σ-finite measure spaces $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, 2$; see E. 2.29. Let $h : \Omega \to \bar{\mathbb{R}}$ be measurable wrt $\mathcal{F}_1 \otimes \mathcal{F}_2$. If $h \in L_+$ or $h \in L^1$, then $\omega_1 \mapsto \int_{\Omega_2} h(\omega_1, \omega_2)\,\mathrm{d}\mu_2(\omega_2)$ is $\mathcal{F}_1$-measurable, $\omega_2 \mapsto \int_{\Omega_1} h(\omega_1, \omega_2)\,\mathrm{d}\mu_1(\omega_1)$ is $\mathcal{F}_2$-measurable and

$$\int_{\Omega} h\,\mathrm{d}\mu = \int_{\Omega_1} \left(\int_{\Omega_2} h(\omega_1, \omega_2)\,\mathrm{d}\mu_2(\omega_2) \right) \mathrm{d}\mu_1(\omega_1)$$
$$= \int_{\Omega_2} \left(\int_{\Omega_1} h(\omega_1, \omega_2)\,\mathrm{d}\mu_1(\omega_1) \right) \mathrm{d}\mu_2(\omega_2).$$

Proof. See Klenke (2008, T. 4.16). □

- $|h| \in L_+$, so one applies Tonelli's theorem to calculate $\int_{\mathbb{R}^d} |h(\boldsymbol{x})|\,\mathrm{d}F(\boldsymbol{x})$ to check whether $\int_{\mathbb{R}^d} h(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x})$ exists. If $\int_{\mathbb{R}^d} |h(\boldsymbol{x})|\,\mathrm{d}F(\boldsymbol{x}) < \infty$, then $h \in L^1$ and one applies Fubini's theorem to calculate $\int_{\mathbb{R}^d} h(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x})$.
- According to the Fubini–Tonelli theorem, the order of the iterated integrals does not matter, so changing the order of integration is allowed.

An important application of the Fubini–Tonelli theorem is the following.

Proposition 5.24 (Expectation of products under independence)
If $X_1, \dots, X_d \in L^1$ are independent, then $\mathbb{E}(X_1 \cdot \ldots \cdot X_d) = \mathbb{E}(X_1) \cdot \ldots \cdot \mathbb{E}(X_d)$.

Proof. We first check that $\mathbb{E}(X_1 \cdot \ldots \cdot X_d)$ exists via

$$\begin{aligned}
\mathbb{E}(|X_1 \cdot \ldots \cdot X_d|) = \mathbb{E}(|X_1| \cdot \ldots \cdot |X_d|) &\underset{\text{T. 5.21}}{=} \int_{\mathbb{R}^d} |x_1| \cdot \ldots \cdot |x_d| \,\mathrm{d}F(\boldsymbol{x}) \\
&\underset{\text{Tonelli}}{=} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} |x_1| \cdot \ldots \cdot |x_d| \,\mathrm{d}F_1(x_1) \cdot \ldots \cdot \mathrm{d}F_d(x_d) \\
&\underset{\text{lin.}}{=} \prod_{j=1}^d \int_{\mathbb{R}} |x_j| \,\mathrm{d}F_j(x_j) = \mathbb{E}(|X_1|) \cdot \ldots \cdot \mathbb{E}(|X_d|) \underset{\text{ass.}}{<} \infty.
\end{aligned}$$

Since $\mathbb{E}(|X_1 \cdot \ldots \cdot X_d|) < \infty$, $X_1 \cdot \ldots \cdot X_d \in L^1$ and Fubini now implies that

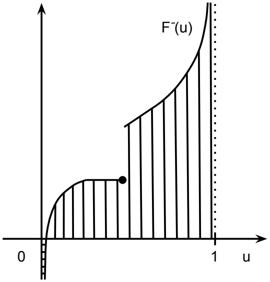
$$\begin{aligned}
\mathbb{E}(X_1 \cdot \ldots \cdot X_d) &\underset{\text{T. 5.21}}{=} \int_{\mathbb{R}^d} x_1 \cdot \ldots \cdot x_d \,\mathrm{d}F(\boldsymbol{x}) \\
&\underset{\text{Fubini}}{=} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} x_1 \cdot \ldots \cdot x_d \,\mathrm{d}F_1(x_1) \cdot \ldots \cdot \mathrm{d}F_d(x_d) \\
&\underset{\text{lin.}}{=} \prod_{j=1}^d \int_{\mathbb{R}} x_j \,\mathrm{d}F_j(x_j) = \mathbb{E}(X_1) \cdot \ldots \cdot \mathbb{E}(X_d).
\end{aligned}$$

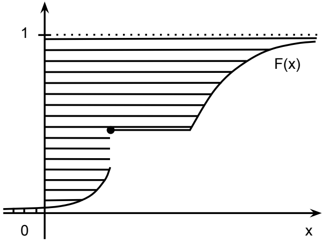
□

Another way of calculating expectations is the following.

Proposition 5.25 (Expectation via quantile or survival function)
Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and recall that $\bar{F}(x) = \mathbb{P}(X > x) = 1 - F(x)$ is the survival function of $X \sim F$. Then $\mathbb{E}(X) = \int_0^1 F^{-1}(u)\,\mathrm{d}u = \int_0^\infty \bar{F}(x)\,\mathrm{d}x - \int_{-\infty}^0 F(x)\,\mathrm{d}x$.

Proof. Quantile transform $\Rightarrow X \overset{\mathrm{d}}{=} F^{-1}(U)$ for $U \sim \mathrm{U}(0,1) \Rightarrow \mathbb{E}(X) = \mathbb{E}(F^{-1}(U))$ $= \int_0^1 F^{-1}(u)\,\mathrm{d}u$. The second equality can be seen geometrically as the graph of F is the one of F^{-1} mirrored at $y = x$ (this argument is much easier than an analytic proof based on integration by parts for Stieltjes integrals):





□

5.3 Variance, covariance and correlation

Definition 5.26 (Variance, standard deviation, covariance, correlation)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then

$$\operatorname{var}(X) = \mathbb{E}\big((X - \mathbb{E}(X))^2\big)$$

is the *variance* and $\operatorname{sd}(X) = \sqrt{\operatorname{var}(X)}$ the *standard deviation* of X (or its distribution or df F). The *covariance* of X, Y is

$$\operatorname{cov}(X, Y) = \mathbb{E}\big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big)$$

and the *correlation* of X, Y (or *Pearson's correlation coefficient*) is

$$\rho = \operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}.$$

Lemma 5.27 (Basic properties)

1) $\operatorname{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
2) $\operatorname{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
3) $\operatorname{cov}(X, X) = \operatorname{var}(X)$

4) $\operatorname{cov}(Y,X)=\operatorname{cov}(X,Y)$

5) $\operatorname{cov}(X,c)=0\ \forall\, c\in\mathbb{R}$

6) $\operatorname{var}(X)=0$ iff $X\overset{\text{a.s.}}{=}\mathbb{E}(X)$

7) $\operatorname{var}(aX+bY)=a^2\operatorname{var}(X)+2ab\operatorname{cov}(X,Y)+b^2\operatorname{var}(Y)$. For $a=b=1$, we see that $\operatorname{var}(X+Y)=\operatorname{var}(X)+\operatorname{var}(Y)+2\operatorname{cov}(X,Y)$. And for $Y\overset{\text{a.s.}}{=}1$, we see that $\operatorname{var}(aX+b)=a^2\operatorname{var}(X)$.

8) X,Y independent $\Rightarrow\operatorname{cov}(X,Y)=\operatorname{cor}(X,Y)=0$

Proof.

1) $\operatorname{var}(X)=\mathbb{E}\big((X-\mathbb{E}(X))^2\big)\underset{\text{out}}{\overset{\text{multiply}}{=}}\mathbb{E}(X^2-2X\mathbb{E}(X)+\mathbb{E}(X)^2)\underset{\text{lin.}}{=}\mathbb{E}(X^2)-2\mathbb{E}(X)\mathbb{E}(X)+\mathbb{E}(X)^2=\mathbb{E}(X^2)-\mathbb{E}(X)^2$.

2) $\operatorname{cov}(X,Y)=\mathbb{E}\big((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big)\underset{\text{out}}{\overset{\text{multiply}}{=}}\mathbb{E}(XY-\mathbb{E}(X)Y-X\mathbb{E}(Y)+\mathbb{E}(X)\mathbb{E}(Y))\underset{\text{lin.}}{=}\mathbb{E}(XY)-2\mathbb{E}(X)\mathbb{E}(Y)+\mathbb{E}(X)\mathbb{E}(Y)=\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)$.

3) $\operatorname{cov}(X,X)=\mathbb{E}\big((X-\mathbb{E}(X))(X-\mathbb{E}(X))\big)=\mathbb{E}\big((X-\mathbb{E}(X))^2\big)=\operatorname{var}(X)$.

4) $\operatorname{cov}(Y,X)=\mathbb{E}\big((Y-\mathbb{E}(Y))(X-\mathbb{E}(X))\big)=\mathbb{E}\big((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big)=\operatorname{cov}(X,Y)$.

5) $\operatorname{cov}(X,c) \underset{2)}{=} \mathbb{E}(cX) - c\mathbb{E}(X) \underset{\text{lin.}}{=} c\mathbb{E}(X) - c\mathbb{E}(X) = 0,\ c \in \mathbb{R}.$

6) $\operatorname{var}(X) = \mathbb{E}\big((X-\mathbb{E}(X))^2\big) = 0 \underset{\text{L. 5.8 1)}}{\Leftrightarrow} (X-\mathbb{E}(X))^2 \overset{\text{a.s.}}{=} 0 \Leftrightarrow X \overset{\text{a.s.}}{=} \mathbb{E}(X).$

7) $\operatorname{var}(aX+bY) = \mathbb{E}\big(((aX+bY) - \mathbb{E}(aX+bY))^2\big) = \mathbb{E}\Big(\big(a(X-\mathbb{E}(X)) + b(Y-\mathbb{E}(Y))\big)^2\Big) = \mathbb{E}\big(a^2(X-\mathbb{E}(X))^2\big) + \mathbb{E}\big(2ab(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big) + \mathbb{E}\big(b^2(Y-\mathbb{E}(Y))^2\big) \underset{\text{def.}}{\overset{\text{lin.}}{=}} a^2\operatorname{var}(X) + 2ab\operatorname{cov}(X,Y) + b^2\operatorname{var}(Y).$

8) $\mathbb{E}(XY) \underset{\text{P. 5.24}}{\overset{\text{ind.}}{=}} \mathbb{E}(X)\mathbb{E}(Y) \underset{2)}{\Rightarrow} \operatorname{cov}(X,Y) = 0 \underset{\text{def.}}{\Rightarrow} \operatorname{cor}(X,Y) = 0.$ □

From the last property, we see that independence implies *uncorrelatedness*. The converse is false in general, take e.g. $X \sim \mathrm{U}(-1,1)$, $Y = X^2 \Rightarrow \operatorname{cov}(X,Y) = \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(X^2) \underset{\text{about 0}}{\overset{\text{symm.}}{=}} 0 - 0 = 0$, but X, Y are not independent.

Proposition 5.28 (Cauchy–Schwarz inequality (CSI))

If $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, then

1) $|\mathbb{E}(XY)| \le (\mathbb{E}(X^2)\mathbb{E}(Y^2))^{1/2}$ (CSI) with equality iff $Y \overset{\text{a.s.}}{=} mX$.

2) $\rho = \operatorname{cor}(X,Y) \in [-1,1]$ with $\rho = \pm 1$ iff $Y \overset{\text{a.s.}}{=} mX + c$ (with $m \gtrless 0$ iff $\rho = \pm 1$).

Proof.

1) For $t \in \mathbb{R}$, let $Z_t = tX + Y$. Then $0 \le \mathbb{E}(Z_t^2) \overset{\text{multiply out}}{\underset{\text{lin.}}{=}} t^2\mathbb{E}(X^2) + 2t\mathbb{E}(XY) + \mathbb{E}(Y^2) =: at^2 + bt + c$, a polynomial in t with at most one root $\Rightarrow 0 \ge b^2 - 4ac = 4(\mathbb{E}(XY))^2 - 4\mathbb{E}(X^2)\mathbb{E}(Y^2)$ and thus $|\mathbb{E}(XY)| \le (\mathbb{E}(X^2)\mathbb{E}(Y^2))^{1/2}$. Furthermore, we have equality iff $b^2 - 4ac = 0 \Leftrightarrow \exists!\, t \in \mathbb{R} : at^2 + bt + c = 0 \Leftrightarrow \mathbb{E}(Z_t^2) = 0 \underset{\text{L. 5.8 1)}}{\Leftrightarrow} Z_t \overset{\text{a.s.}}{=} 0 \Leftrightarrow Y \overset{\text{a.s.}}{=} -tX$.

2) Applying the CSI to the centered rvs $\tilde{X} := X - \mathbb{E}(X)$, $\tilde{Y} := Y - \mathbb{E}(Y)$ gives $|\operatorname{cov}(X,Y)| \underset{\text{def.}}{=} |\mathbb{E}(\tilde{X}\tilde{Y})| \underset{\text{CSI}}{\le} (\mathbb{E}(\tilde{X}^2)\mathbb{E}(\tilde{Y}^2))^{1/2} \underset{\text{def.}}{=} (\operatorname{var}(X)\operatorname{var}(Y))^{1/2}$ with equality iff $\tilde{Y} \overset{\text{a.s.}}{=} m\tilde{X}$, so iff $Y \overset{\text{a.s.}}{=} mX + c$. Therefore, $|\rho| = \frac{|\operatorname{cov}(X,Y)|}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} \le 1$ with $\rho = \pm 1$ iff $Y \overset{\text{a.s.}}{=} mX + c$. In this case, $\rho = \frac{\operatorname{cov}(X, mX+c)}{\sqrt{\operatorname{var}(X)\operatorname{var}(mX+c)}} = \frac{m\operatorname{var}(X)}{\sqrt{m^2\operatorname{var}(X)^2}} = \frac{m}{|m|}$ which is ± 1 iff $m \gtrless 0$. ∎

- $\operatorname{cor}(X,Y)$ is only a measure of linear dependence ($Y = X^2$ is non-linear). Also, $\mathbb{E}(X^2) < \infty$, $\mathbb{E}(Y^2) < \infty$ are required for $\operatorname{cor}(X,Y)$ to exist.
- There are other *measures of association* between X, Y, e.g. Kendall's tau $\tau = \rho(\mathbb{1}_{\{F_1(X_1) \le F_1(X_1')\}}, \mathbb{1}_{\{F_2(X_2) \le F_2(X_2')\}})$, Spearman's rho $\rho_S = \rho(F_1(X_1), F_2(X_2))$.
- No summary statistic can characterize the dependence between X, Y (Sklar).

Lemma 5.29 (Hoeffding's lemma)
If $(X_1, X_2) \sim F$ with margins F_1, F_2 and $\mathbb{E}(X_j^2) < \infty$, $j \in \{1,2\}$, then $\operatorname{cov}(X_1, X_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(F(x_1,x_2) - F_1(x_1)F_2(x_2))\,\mathrm{d}x_1\mathrm{d}x_2$.

Proof. Let (X_1', X_2') be an iid-copy of (X_1, X_2). Then

$$
\begin{aligned}
&2\operatorname{cov}(X_1, X_2)\\
&= \mathbb{E}\Big((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\Big) + \mathbb{E}\Big((X_1' - \mathbb{E}(X_1'))(X_2' - \mathbb{E}(X_2'))\Big)\\
&\overset{\text{lin., multiply out}}{\underset{\text{ind.}}{=}} \mathbb{E}\Big(\big((X_1 - \mathbb{E}(X_1)) - (X_1' - \mathbb{E}(X_1'))\big)\cdot\big((X_2 - \mathbb{E}(X_2)) - (X_2' - \mathbb{E}(X_2'))\big)\Big)\\
&= \mathbb{E}((X_1 - X_1')(X_2 - X_2')).
\end{aligned}
$$

With $b - a = \int_{-\infty}^{\infty}(\mathbb{1}_{\{a\le x\}} - \mathbb{1}_{\{b\le x\}})\,\mathrm{d}x$ for all $a, b \in \mathbb{R}$, we obtain that

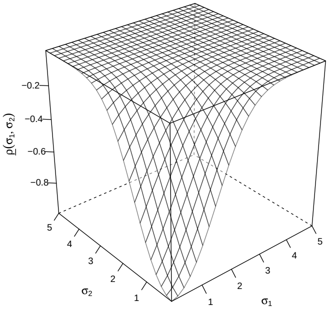
$$
\begin{aligned}
&2\operatorname{cov}(X_1, X_2)\\
&= \mathbb{E}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\mathbb{1}_{\{X_1'\le x_1\}} - \mathbb{1}_{\{X_1\le x_1\}})(\mathbb{1}_{\{X_2'\le x_2\}} - \mathbb{1}_{\{X_2\le x_2\}})\,\mathrm{d}x_1\,\mathrm{d}x_2\right]\\
&\underset{\text{Fubini}}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathbb{E}(\dots)\,\mathrm{d}x_1\,\mathrm{d}x_2 \overset{\text{multiply out}}{\underset{\text{ind.}}{=}} 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\big(F(x_1,x_2) - F_1(x_1)F_2(x_2)\big)\,\mathrm{d}x_1\,\mathrm{d}x_2.
\end{aligned}
$$

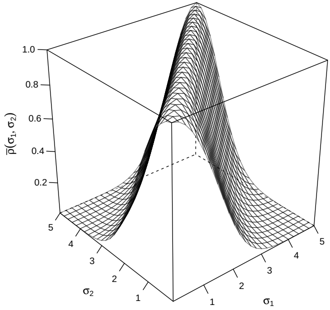
□

Example 5.30 (Correlation fallacies)

1) **Fallacy 1: F_1, F_2, ρ uniquely determine F.** Consider $X_1 \sim \mathrm{N}(0,1)$ and $X_2 = (-1)^{\mathbb{1}_{\{U \le 1/2\}}} X_1$ for $U \sim \mathrm{U}(0,1)$ independent of X_1. Then $X_2 \sim \mathrm{N}(0,1)$ and $\operatorname{cor}(X_1, X_2) \underset{\text{var.}}{\overset{\text{unit}}{=}} \operatorname{cov}(X_1, X_2) \underset{\text{centered}}{=} \mathbb{E}(X_1 X_2) \underset{\text{ind.}}{=} \mathbb{E}((-1)^{\mathbb{1}_{\{U \le 1/2\}}})\mathbb{E}(X_1^2) = 0$, but also $X_1', X_2' \overset{\text{ind.}}{\sim} \mathrm{N}(0,1)$ implies $\operatorname{cor}(X_1', X_2') = 0$.

2) **Fallacy 2: Given F_1, F_2 any $\rho \in [-1, 1]$ is attainable.** Let $X_j \sim \mathrm{LN}(0, \sigma_j^2)$, $j \in \{1, 2\}$. By Hoeffding's lemma, $\exists$ minimal ρ ($\underline{\rho}$; for W; left) and maximal ρ ($\overline{\rho}$; for M; right); for $\sigma_1^2 = 1$, $\sigma_2^2 = 16$, $\rho \in [-0.0003,\ 0.0137]$!





5.4 Multivariate notions

The following concepts generalize that of the mean and (co)variance to higher d.

Definition 5.31 (Mean vector, covariance matrix, correlation matrix, etc.) If $X_1, \dots, X_d \in L^1$, then $\mathbb{E}(\boldsymbol{X}) := (\mathbb{E}(X_1), \dots, \mathbb{E}(X_d))$ is the *mean vector* or *expectation* of $\boldsymbol{X}$ (or its df F). If $X_1, \dots, X_d, Y_1, \dots, Y_{d'} \in L^2$, then $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) = (\operatorname{cov}(X_i, Y_j))_{\substack{i=1,\dots,d \\ j=1,\dots,d'}} = \mathbb{E}\big((\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))(\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y}))^\top\big)$ is the *cross-covariance matrix* of $\boldsymbol{X} = (X_1, \dots, X_d)$ and $\boldsymbol{Y} = (Y_1, \dots, Y_{d'})$, $\operatorname{cov}(\boldsymbol{X}) := \operatorname{cov}(\boldsymbol{X}, \boldsymbol{X})$ is the *covariance matrix* and $\operatorname{cor}(\boldsymbol{X}) = (\operatorname{cor}(X_i, X_j))_{i,j=1}^d$ the *correlation matrix* of $\boldsymbol{X}$ (or its df F).

Lemma 5.32 (Properties)

1) $\mathbb{E}(A\boldsymbol{X} + \boldsymbol{b}) = A\mathbb{E}(\boldsymbol{X}) + \boldsymbol{b}$. In particular, $\mathbb{E}(\boldsymbol{a}^\top \boldsymbol{X}) = \boldsymbol{a}^\top \mathbb{E}(\boldsymbol{X})$.
2) $\operatorname{cov}(A\boldsymbol{X} + \boldsymbol{b}, C\boldsymbol{Y} + \boldsymbol{d}) = A \operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) C^\top$. In particular, $\operatorname{cov}(A\boldsymbol{X} + \boldsymbol{b}) = A \operatorname{cov}(\boldsymbol{X}) A^\top$ and $\operatorname{var}(\boldsymbol{a}^\top \boldsymbol{X}) = \operatorname{cov}(\boldsymbol{a}^\top \boldsymbol{X},\ \boldsymbol{a}^\top \boldsymbol{X}) = \boldsymbol{a}^\top \operatorname{cov}(\boldsymbol{X}) \boldsymbol{a}$.
3) $\operatorname{cov}(\boldsymbol{X} + \boldsymbol{Y}) = \operatorname{cov}(\boldsymbol{X}) + \operatorname{cov}(\boldsymbol{Y}) + \operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) + \operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})^\top$.

Proof. The given special cases follow immediately.

1) The jth component $\mathbb{E}(A\boldsymbol{X}+\boldsymbol{b})_j$ of $\mathbb{E}(A\boldsymbol{X}+\boldsymbol{b})$ satisfies

$$\mathbb{E}(A\boldsymbol{X}+\boldsymbol{b})_j = \mathbb{E}\biggl(\sum_{k=1}^d a_{jk}X_k + b_j\biggr) \underset{\text{lin.}}{=} \sum_{k=1}^d a_{jk}\mathbb{E}(X_k) + b_j = (A\mathbb{E}(\boldsymbol{X})+\boldsymbol{b})_j.$$

2) The (i,j)th component $\operatorname{cov}(A\boldsymbol{X}+\boldsymbol{b}, C\boldsymbol{Y}+\boldsymbol{d})_{i,j}$ of $\operatorname{cov}(A\boldsymbol{X}+\boldsymbol{b}, C\boldsymbol{Y}+\boldsymbol{d})$ is

$$\begin{aligned}
\operatorname{cov}(A\boldsymbol{X}+\boldsymbol{b}, C\boldsymbol{Y}+\boldsymbol{d})_{i,j} &= \operatorname{cov}((A\boldsymbol{X}+\boldsymbol{b})_i,\ (C\boldsymbol{Y}+\boldsymbol{d})_j)\\
&= \operatorname{cov}\biggl(\sum_{k=1}^d a_{ik}X_k + b_i,\ \sum_{l=1}^{d'} c_{jl}Y_l + d_j\biggr)\\
&\underset{\text{def.}}{=} \mathbb{E}\biggl(\biggl(\sum_{k=1}^d a_{ik}(X_k-\mathbb{E}(X_k))\biggr)\biggl(\sum_{l=1}^{d'} c_{jl}(Y_l-\mathbb{E}(Y_l))\biggr)\biggr)\\
&\overset{\text{multiply}}{\underset{\text{out}}{=}} \mathbb{E}\biggl(\sum_{k=1}^d\sum_{l=1}^{d'} a_{ik}c_{jl}(X_k-\mathbb{E}(X_k))(Y_l-\mathbb{E}(Y_l))\biggr)\\
&\overset{\text{lin.}}{\underset{\text{def.}}{=}} \sum_{k=1}^d\sum_{l=1}^{d'} a_{ik}c_{jl}\operatorname{cov}(X_k,Y_l) = (A\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})C^\top)_{i,j}.
\end{aligned}$$

3) The (i,j)th component $\operatorname{cov}(\boldsymbol{X}+\boldsymbol{Y})_{i,j}$ of $\operatorname{cov}(\boldsymbol{X}+\boldsymbol{Y})$ is

$$\begin{aligned}
\operatorname{cov}(\boldsymbol{X}+\boldsymbol{Y})_{i,j} &= \operatorname{cov}((\boldsymbol{X}+\boldsymbol{Y})_i,(\boldsymbol{X}+\boldsymbol{Y})_j) = \operatorname{cov}(X_i+Y_i, X_j+Y_j)\\
&\underset{\text{def.}}{=} \mathbb{E}\Big((X_i+Y_i-\mathbb{E}(X_i+Y_i))(X_j+Y_j-\mathbb{E}(X_j+Y_j))\Big)\\
&= \mathbb{E}((X_i-\mathbb{E}X_i+Y_i-\mathbb{E}Y_i)(X_j-\mathbb{E}X_j+Y_j-\mathbb{E}Y_j))\\
&\overset{\text{multiply}}{\underset{\text{out}}{=}} \mathbb{E}((X_i-\mathbb{E}X_i)(X_j-\mathbb{E}X_j))+\mathbb{E}((X_i-\mathbb{E}X_i)(Y_j-\mathbb{E}Y_j))\\
&\quad+\mathbb{E}((Y_i-\mathbb{E}Y_i)(X_j-\mathbb{E}X_j))+\mathbb{E}((Y_i-\mathbb{E}Y_i)(Y_j-\mathbb{E}Y_j))\\
&\underset{\text{def.}}{=} (\operatorname{cov}(\boldsymbol{X},\boldsymbol{X})+\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})+\operatorname{cov}(\boldsymbol{Y},\boldsymbol{X})+\operatorname{cov}(\boldsymbol{Y},\boldsymbol{Y}))_{i,j}\\
&\underset{\text{def.}}{=} (\operatorname{cov}(\boldsymbol{X})+\operatorname{cov}(\boldsymbol{Y})+\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})+\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})^\top)_{i,j}. \qquad \square
\end{aligned}$$

The following corollary lists immediate consequences of L. 5.32.

Corollary 5.33 (Implied properties)

1) For $\boldsymbol{a}=(1,\dots,1)$, we obtain from $\operatorname{var}(\boldsymbol{a}^\top\boldsymbol{X})=\boldsymbol{a}^\top\operatorname{cov}(\boldsymbol{X})\boldsymbol{a}$ that

$$\operatorname{var}\Bigg(\sum_{j=1}^d X_j\Bigg) \underset{\boldsymbol{a}=\mathbf{1}}{=} \sum_{j=1}^d\sum_{j'=1}^d \operatorname{cov}(X_j,X_{j'}) = \sum_{j=1}^d \operatorname{var}(X_j) + 2\sum_{1\le j<j'\le d}\operatorname{cov}(X_j,X_{j'}).$$

If $X_j, X_{j'}$ are uncorrelated $\forall\, j\neq j'$, then $\operatorname{var}(\sum_{j=1}^d X_j)=\sum_{j=1}^d \operatorname{var}(X_j)$.

2) For $\boldsymbol{a} = (1/d, \dots, 1/d)$ and uncorrelated $X_j, X_{j'}\ \forall\, j \neq j'$, we have

$$\operatorname{var}\Bigl(\frac{1}{d}\sum_{j=1}^d X_j\Bigr) \underset{\text{L. 5.27 7)}}{=} \frac{1}{d^2}\operatorname{var}\Bigl(\sum_{j=1}^d X_j\Bigr) \underset{\text{shown}}{=} \frac{1}{d^2}\sum_{j=1}^d \operatorname{var}(X_j).$$

If $X_1, \dots, X_d$ are id ($\Rightarrow$ all variances equal), then $\operatorname{var}(\frac{1}{d}\sum_{j=1}^d X_j) = \frac{\operatorname{var}(X_1)}{d}$.

- One can show that a symmetric, positive semi-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$ allows for the *Cholesky decomposition*

$$\Sigma = AA^\top,$$

where the *Cholesky factor* $A \in \mathbb{R}^{d\times d}$ is a lower triangular matrix with nonnegative (positive if Σ is positive definite) diagonal elements.

- The Cholesky factor A can be calculated row by row from $\Sigma = AA^\top$ via

$$A_{j,j} = \sqrt{\Sigma_{j,j} - \sum_{k=1}^{j-1} A_{j,k}^2}, \quad j = 1, \dots, d,$$

$$A_{i,j} = \frac{1}{A_{j,j}}\Bigl(\Sigma_{i,j} - \sum_{k=1}^{j-1} A_{i,k}A_{j,k}\Bigr), \quad i = j+1, \dots, d.$$

Based on the Cholesky decomposition, one can provide a characterization of covariance matrices.

Proposition 5.34 (Characterization of covariance matrices)
A real, symmetric matrix Σ is a covariance matrix iff it is positive semi-definite.

Proof.

"$\Rightarrow$": If Σ is the covariance matrix of a random vector $\boldsymbol{X}$, then $\boldsymbol{a}^\top\Sigma\boldsymbol{a} \underset{\text{L. 5.32 2)}}{=} \mathrm{var}(\boldsymbol{a}^\top\boldsymbol{X}) \ge 0\ \forall\,\boldsymbol{a} \in \mathbb{R}^d$, so covariance matrices Σ are positive semi-definite.

"$\Leftarrow$": Let $\Sigma \in \mathbb{R}^{d\times d}$ be positive semi-definite with Cholesky factor A. Let $\boldsymbol{X}$ be a random vector with $\mathrm{cov}(\boldsymbol{X}) = I_d = \mathrm{diag}(1,\dots,1)$ (e.g. $X_j \overset{\text{ind.}}{\sim} \mathrm{N}(0,1)$). Then $\mathrm{cov}(A\boldsymbol{X}) = A\,\mathrm{cov}(\boldsymbol{X})A^\top = AA^\top = \Sigma$, i.e. Σ is a covariance matrix (namely that of $A\boldsymbol{X}$). $\square$

As a direct consequence, a real, symmetric matrix P is a correlation matrix iff it is positive semi-definite and has $1s$ on its diagonal.

Lemma 5.35 (Concentration on a subspace of dimension at most $d-1$)
If X has $\mathrm{cov}(\boldsymbol{X}) = \Sigma$ and $\exists\, \boldsymbol{a} \in \mathbb{R}^d\backslash\{\boldsymbol{0}\} : \boldsymbol{a}^\top \Sigma \boldsymbol{a} = 0$, then $X_1, \dots, X_d$ are linearly dependent a.s.

Proof. $\mathrm{var}(\boldsymbol{a}^\top \boldsymbol{X}) = \mathrm{cov}(\boldsymbol{a}^\top \boldsymbol{X}, \boldsymbol{a}^\top \boldsymbol{X}) \underset{\text{L. 5.32 2)}}{=} \boldsymbol{a}^\top \Sigma \boldsymbol{a} \underset{\text{ass.}}{=} 0$, so that, by L. 5.27 6), $\boldsymbol{a}^\top \boldsymbol{X}$ is constant a.s. Therefore, $X_1, \dots, X_d$ are linearly dependent a.s. □

Invertibility of covariance matrices (including correlation matrices) can be characterized as follows.

Lemma 5.36 (Positive definite iff full rank)
$\Sigma \in \mathbb{R}^{d\times d}$ is positive definite iff $\mathrm{rank}(\Sigma) = d$ iff $\mathrm{rank}(A) = d$.

Proof. We show the two equivalences separately.

1) $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite $\Leftrightarrow \boldsymbol{x}^\top \Sigma \boldsymbol{x} > 0\ \forall\, \boldsymbol{x} \in \mathbb{R}^d\backslash\{\boldsymbol{0}\} \Leftrightarrow \Sigma \boldsymbol{x} \neq \boldsymbol{0}\ \forall\, \boldsymbol{x} \in \mathbb{R}^d\backslash\{\boldsymbol{0}\} \underset{\text{rank-nullity}}{\Leftrightarrow} \mathrm{rank}(\Sigma) = d - \dim(\ker(\Sigma)) = d - \dim(\{\boldsymbol{x} : \Sigma\boldsymbol{x} = \boldsymbol{0}\}) = d$.
2) We first show that $AA^\top = 0$ iff $A = 0$. "$\Leftarrow$" is clear. For "$\Rightarrow$", $AA^\top = (\sum_{k=1}^n a_{i,k} a_{j,k})_{i,j} \overset{!}{=} 0 \underset{j=i}{\Rightarrow} \sum_{k=1}^n a_{i,k}^2 = 0 \Rightarrow a_{i,j} = 0\ \forall\, i, j$. Now $\mathrm{rank}(A) \underset{\text{rank-nullity}}{=} d - \dim(\ker(A)) \underset{\text{shown}}{=} d - \dim(\ker(AA^\top)) \underset{\text{rank-nullity}}{=} \mathrm{rank}(AA^\top) = \mathrm{rank}(\Sigma)$. □

5.5 The Lebesgue–Radon–Nikodym theorem

Definition 5.37 (Absolutely continuous, equivalent, singular measures)

Let μ, ν be measures on a measurable space $(\Omega, \mathcal{F})$. Then

- ν is *absolutely continuous* wrt μ (notation: $\nu \ll \mu$) if $\nu(A) = 0 \;\forall\, A \in \mathcal{F} : \mu(A) = 0$;
- ν, μ are *equivalent* if $\nu \ll \mu$ and $\mu \ll \nu$ (share the same null sets); and
- ν, μ are *singular* wrt to each other (notation: $\nu \perp \mu$) if $\exists\, A \in \mathcal{F} : \nu(A) = 0$ (ν lives on A^c) and $\mu(A^c) = 0$ (μ lives on A).

Lemma 5.38 (Sums of measures)

Let ν_n, $n \in \mathbb{N}$, be measures on a measurable space $(\Omega, \mathcal{F})$ and thus $\nu := \sum_{n=1}^{\infty} \nu_n$ a measure. If $\nu_n \ll \mu \;\forall\, n \in \mathbb{N}$, then $\nu \ll \mu$. And if $\nu_n \perp \mu \;\forall\, n \in \mathbb{N}$, then $\nu \perp \mu$.

Proof. For $A \in \mathcal{F} : \mu(A) = 0$, we have $\nu(A) = \sum_{n=1}^{\infty} \nu_n(A) \underset{\nu_n \ll \mu}{=} 0$. For the second statement, $\nu_n \perp \mu$, $n \in \mathbb{N} \Rightarrow \exists\, A_n \in \mathcal{F} : \nu_n(A_n) = 0$, $\mu(A_n^c) = 0$, $n \in \mathbb{N}$. Let $A := \bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$. Then $A^c \underset{\text{De Morgan}}{=} \bigcup_{n=1}^{\infty} A_n^c \in \mathcal{F}$ and $0 \le \nu(A) \underset{\text{def.}}{=} \sum_{n=1}^{\infty} \nu_n(A) \underset{\text{mon.}}{\le} \sum_{n=1}^{\infty} \nu_n(A_n) \underset{\text{def. } A_n}{=} 0$ and $0 \le \mu(A^c) \underset{\sigma\text{-subadd.}}{\le} \sum_{n=1}^{\infty} \mu(A_n^c) \underset{\text{def. } A_n^c}{=} 0$. □

For the following result, we need the notion of signed measures. A *signed measure* on $(\Omega, \mathcal{F})$ is a function which satisfies

i) $\mu : \mathcal{F} \to [-\infty, \infty]$ and μ attains at most one of $\pm\infty$;

ii) $\mu(\emptyset) = 0$; and

iii) μ is σ-additive.

Lemma 5.39 (Relationship between finite measures)

If μ, ν are finite measures on $(\Omega, \mathcal{F})$, then either $\nu \perp \mu$ or $\exists\, \varepsilon > 0, A \in \mathcal{F} : \mu(A) > 0$, $\nu|_A \geq \varepsilon \mu|_A$.

Proof. $\forall\, n \in \mathbb{N}$, Hahn's decomposition theorem (see Folland (1999, T. 3.3)) applied to the signed measure $\nu - \frac{1}{n}\mu$ implies that $\exists\, A_n \in \mathcal{F} : \nu - \frac{1}{n}\mu|_{A_n} \geq 0$ and $\nu - \frac{1}{n}\mu|_{A_n^c} \leq 0$. Consider $A_\infty := \bigcup_{n=1}^\infty A_n \in \mathcal{F}$. Then

$$0 \underset{\nu \text{ meas.}}{\leq} \nu(A_\infty^c) \underset{A_\infty^c \subseteq A_n^c}{\overset{\text{mon.}}{\leq}} \nu(A_n^c) \underset{\text{def.}}{\leq} \frac{1}{n}\mu(A_n^c) \underset{A_n^c \subseteq \Omega}{\leq} \frac{1}{n}\mu(\Omega), \quad n \in \mathbb{N}.$$

$\underset{n \to \infty}{\Rightarrow} \nu(A_\infty^c) = 0$. If $\mu(A_\infty) = 0$, then $\nu \perp \mu$. And if $\mu(A_\infty) > 0$, $\exists\, n \in \mathbb{N}$: $\mu(A_n) > 0$ for some n and $\nu|_{A_n} \geq \frac{1}{n}\mu|_{A_n}$, so choose $A := A_n$ and $\varepsilon := 1/n$. $\square$

Theorem 5.40 (Lebesgue–Radon–Nikodym theorem (LRN))
Let μ, ν be σ-finite measures on a measurable space $(\Omega, \mathcal{F})$.

i) *Lebesgue decomposition*: $\exists!$ σ-finite measures ν_a, ν_s on $(\Omega, \mathcal{F})$ such that

$$\nu = \nu_a + \nu_s, \quad \nu_a \ll \mu, \ \nu_s \perp \mu.$$

ii) *Radon–Nikodym theorem*: $\exists$ a μ-a.e. unique integrable $f : \Omega \to [0, \infty)$:

$$\nu_a(A) = \int_A f \,\mathrm{d}\mu, \quad \forall\, A \in \mathcal{F},$$

so ν_a has *density* f wrt μ (notation: $\mathrm{d}\nu_a = f \mathrm{d}\mu$).

Proof. We prove both statements together, distinguishing two cases.

Case 1) $\mu(\Omega) < \infty$ and $\nu(\Omega) < \infty$.

1) *Maximum of densities.* Consider

$$\mathcal{A} = \left\{ g : \Omega \to [0, \infty] : g \text{ is } \mu\text{-integrable}, \int_A g \,\mathrm{d}\mu \le \nu(A) \ \forall\, A \in \mathcal{F} \right\}.$$

Then $0 \in \mathcal{A}$, so $\mathcal{A} \neq \emptyset$. And if $g_1, g_2 \in \mathcal{A}$, let $B := \{g_1 < g_2\}$ and note that $\int_A \max\{g_1, g_2\} \,\mathrm{d}\mu \underset{\text{tot. meas.}}{\overset{\text{L. 5.8 6)}}{=}} \int_{A \cap B} g_2 \,\mathrm{d}\mu + \int_{A \cap B^c} g_1 \,\mathrm{d}\mu \underset{g_1,\, g_2 \,\in\, \mathcal{A}}{\le}$

$\nu(A\cap B)+\nu(A\cap B^c) \underset{\text{meas.}}{\overset{\text{tot.}}{=}} \nu(A)$, so $\max\{g_1,g_2\}\in\mathcal{A}$. By induction, also $\max\{g_1,\dots,g_n\}\in\mathcal{A}$ for any number $n\in\mathbb{N}$ of such functions.

2) *Existence of the density f.* Let $S:=\sup_{g\in\mathcal{A}}\{\int_\Omega g\,\mathrm{d}\mu\} \underset{\text{for } A=\Omega}{\overset{\text{def. }\mathcal{A}}{\le}} \nu(\Omega) \underset{\text{ass.}}{<} \infty$. $\underset{\text{def. sup}}{\Rightarrow} \exists\,\{g_n\}_{n\in\mathbb{N}}\subseteq\mathcal{A}: \int_\Omega g_n\,\mathrm{d}\mu \underset{n\to\infty}{\to} S$. Let $f_n:=\max\{g_1,\dots,g_n\} \underset{1)}{\in} \mathcal{A}$ and $f:=\sup_{n\in\mathbb{N}} g_n \Rightarrow f_n\nearrow f$ pointwise $\underset{\text{MON}}{\overset{f_n\in L_+}{\Rightarrow}}$ f is integrable and $\int_\Omega f\,\mathrm{d}\mu=\lim_{n\to\infty}\int_\Omega f_n\,\mathrm{d}\mu$. Furthermore:

- $f_n\in\mathcal{A}\ \forall\, n \Rightarrow \int_A f_n\,\mathrm{d}\mu \underset{\text{def. }\mathcal{A}}{\le} \nu(A)\ \forall\, n \Rightarrow \int_A f\,\mathrm{d}\mu = \int_\Omega f\mathbb{1}_A\,\mathrm{d}\mu \underset{\text{MON}}{=} \lim_{n\to\infty}\int_\Omega f_n\mathbb{1}_A\,\mathrm{d}\mu = \lim_{n\to\infty}\int_A f_n\,\mathrm{d}\mu \le \nu(A)$, so $f\in\mathcal{A}$.
- Since $\int_\Omega f\,\mathrm{d}\mu \underset{A=\Omega}{\overset{\text{choose}}{=}} \lim_{n\to\infty}\int_\Omega f_n\,\mathrm{d}\mu \underset{f_n\ge g_n}{\ge} \lim_{n\to\infty}\int_\Omega g_n\,\mathrm{d}\mu \underset{\text{def. } g_n}{=}$ S and $\int_\Omega f\,\mathrm{d}\mu = \lim_{n\to\infty}\int_\Omega f_n\,\mathrm{d}\mu \underset{\int_\Omega f_n\,\mathrm{d}\mu \underset{f_n\in\mathcal{A}}{\le} S\ \forall\, n}{\le} S$, so $\int_\Omega f\,\mathrm{d}\mu = S$.
- Since f is integrable, $f \underset{\text{L. 5.85)}}{<} \infty$ a.e., so we may take f to be real-valued everywhere (redefine as 0 whenever ∞).

3) *Existence of ν_a,ν_s.* Define

$$\nu_a(A):=\int_A f\,\mathrm{d}\mu\ \forall\, A\in\mathcal{F} \text{ and } \nu_s(A):=\nu(A)-\nu_a(A) \underset{f\in\mathcal{A}}{\ge} 0.$$

$\underset{\text{L. 5.11 6)}}{\Rightarrow} \nu_a\ll\mu$. Suppose $\nu_s\not\perp\mu \underset{\text{L. 5.39}}{\Rightarrow} \exists\,\varepsilon>0,\ A\in\mathcal{F}: \mu(A)>0$,

$\nu_s|_A \geq \varepsilon\mu|_A \Rightarrow \forall\, B \in \mathcal{F},\ \int_B \varepsilon \mathbb{1}_A \,\mathrm{d}\mu = \varepsilon\mu(A\cap B) \underset{\varepsilon\mu|_A \leq \nu_s|_A}{\overset{A\cap B\subseteq A}{\leq}} \nu_s(A \cap B) \underset{\text{mon.}}{\leq} \nu_s(B) \underset{\text{def.}}{=} \nu(B) - \int_B f\,\mathrm{d}\mu \underset{\text{ineq.}}{\overset{\text{solve}}{\Rightarrow}} \int_B (f + \varepsilon\mathbb{1}_A)\,\mathrm{d}\mu \leq \nu(B) \Rightarrow f + \varepsilon\mathbb{1}_A \in \mathcal{A}$ and $\int_\Omega (f+\varepsilon\mathbb{1}_A)\,\mathrm{d}\mu = S + \varepsilon\mu(A) \underset{\text{ass.}}{>} S$ ↯ $\Rightarrow \nu_s \perp \mu$.

4) *Uniqueness of* ν_a, f, ν_s. If $\nu = \nu_a' + \nu_s'$ for $\mathrm{d}\nu_a' = f'\mathrm{d}\mu$, then $\nu_s - \nu_s' = \nu_a' - \nu_a$, so that $\mathrm{lhs}(B) := \nu_s(B) - \nu_s'(B) \underset{\text{lin.}}{=} \int_B (f' - f)\,\mathrm{d}\mu =: \mathrm{rhs}(B)$ $\forall\, B \in \mathcal{F}$. If $\mu(B) > 0$, then $\mathrm{lhs}(B) \underset{\nu_s'\perp\mu}{\overset{\nu_s\perp\mu}{=}} 0$, and if $\mu(B) = 0$, then $\mathrm{rhs}(B) \underset{\text{L. 5.11 6)}}{=} 0 \Rightarrow \mathrm{lhs}(B) = \mathrm{rhs}(B) = 0\ \forall\, B \in \mathcal{F} \underset{\text{P. 5.14 1)}\Rightarrow\text{3)}}{\Rightarrow} f' = f$ μ-a.e. $\Rightarrow \nu_a' = \nu_a \Rightarrow \nu_s' = \nu_s$.

Case 2) μ and ν σ-finite.

- By assumption, $\exists\, \{A_{\mu,m}\}_{m\in\mathbb{N}} \subseteq \mathcal{F} : A_{\mu,i} \cap A_{\mu,j} = \emptyset\ \forall\, i \neq j$, $\biguplus_{m=1}^\infty A_{\mu,m} = \Omega$ (wlog a partition), $\mu(A_{\mu,m}) < \infty$ and similiarly $\{A_{\nu,l}\}_{l\in\mathbb{N}}$ for ν. Then $\Omega = \Omega\cap\Omega = (\biguplus_{m=1}^\infty A_{\mu,m}) \cap (\biguplus_{l=1}^\infty A_{\nu,l}) \underset{\text{2x distr.}}{=} \biguplus_{m=1}^\infty \biguplus_{l=1}^\infty (A_{\mu,m} \cap A_{\nu,l})$. By relabeling, $A_n := A_{\mu,m} \cap A_{\nu,l}$ (Cantor's first diagonal argument), $\exists\, \{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F} : A_i \cap A_j = \emptyset\ \forall\, i \neq j$, $\biguplus_{n=1}^\infty A_n = \Omega$, $\mu(A_n) < \infty$ and $\nu(A_n) < \infty$, $n \in \mathbb{N}$.
- Define the finite measures $\mu_n(B) := \mu(B \cap A_n)$, $\nu_n(B) := \nu(B \cap A_n)$

$\forall\, B \in \mathcal{F}$. Case 1) applied to μ_n, ν_n implies that $\forall\, n \in \mathbb{N}$,

$$\nu_n(B) = \nu_{n,a}(B) + \nu_{n,s}(B) = \int_B f_n \,\mathrm{d}\mu_n + \nu_{n,s}(B) \quad \forall\, B \in \mathcal{F} \quad (*)$$

for $\nu_{n,a}, f_n, \nu_{n,s}$ as in Case 1), in particular, $\nu_{n,a}, \nu_{n,s}$ are finite.

- Since $\mu_n(A_n^c) \underset{\text{def. } \mu_n}{=} 0$ and $\nu_n(A_n^c) \underset{\text{def. } \nu_n}{=} 0$, we have $\nu_{n,s}(A_n^c) \underset{(*)}{=} \nu_n(A_n^c) - \int_{A_n^c} f_n \,\mathrm{d}\mu_n \underset{\text{L. 5.11 6)}}{=} 0$. We may thus assume that $f_n|_{A_n^c} = 0$.
- Let $\nu_a(B) := \int_B f \,\mathrm{d}\mu$ for $f := \sum_{n=1}^\infty f_n$ (f is integrable by L. 5.8 3)) and let $\nu_s := \sum_{n=1}^\infty \nu_{n,s}$. By L. 5.38 and Case 1), $\nu_s \perp \mu$. And ν_a, ν_s are σ-finite since $\nu_{n,a}, \nu_{n,s}$ are concentrated and finite on A_n and $\{A_n\}_{n\in\mathbb{N}}$ partitions Ω.
- $$\nu(B) \underset{\text{partition}}{=} \nu\Big(B \cap \biguplus_{n=1}^\infty A_n\Big) \underset{\text{distr.}}{=} \nu\Big(\biguplus_{n=1}^\infty (B \cap A_n)\Big) \underset{\sigma\text{-add.}}{=} \sum_{n=1}^\infty \nu(B \cap A_n)$$
$$\underset{\text{def. } \nu_n}{=} \sum_{n=1}^\infty \nu_n(B) \underset{(*)}{=} \sum_{n=1}^\infty \Big(\int_B f_n \,\mathrm{d}\mu_n + \nu_{n,s}(B)\Big)$$
$$\overset{\mu_n|_{A_n} \overset{\text{def.}}{=} \mu|_{A_n}}{\underset{f_n|_{A_n^c} = 0}{=}} \sum_{n=1}^\infty \Big(\int_B f_n \,\mathrm{d}\mu + \nu_{n,s}(B)\Big) \overset{\text{Tonelli or C. 5.18}}{\underset{\text{def. } f,\, \nu_s}{=}} \int_B f \,\mathrm{d}\mu + \nu_s(B)$$
$$\underset{\text{def. } \nu_a}{=} \nu_a(B) + \nu_s(B), \quad B \in \mathcal{F}.$$
- Uniqueness follows as in Case 1). □

- The converse is trivial: If μ, ν are σ-finite measures and $\exists$ an integrable $f: \Omega \to [0, \infty): \mathrm{d}\nu = f\,\mathrm{d}\mu$, then $\nu \underset{\text{L. 5.11 6)}}{\ll} \mu$.
- Because of the expressions $\nu(A) = \int_A f\,\mathrm{d}\mu$ or $\mathrm{d}\nu = f\,\mathrm{d}\mu$, the a.e. unique density f is also denoted

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$$

and called *Radon–Nikodym derivative* of ν wrt μ.

Corollary 5.41 (Radon–Nikodym theorem (RN); $\nu_s = 0$ in T. 5.40 i))
If μ, ν are σ-finite measures on a measurable space $(\Omega, \mathcal{F})$ such that $\nu \ll \mu$, then $\exists$ a μ-a.e. unique integrable $f: \Omega \to [0, \infty)$ ($f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$) such that

$$\nu(A) = \int_A f\,\mathrm{d}\mu, \quad \forall A \in \mathcal{F}.$$

The following result shows that formulas suggested by the notation $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ are often correct.

Proposition 5.42 (Formulas involving Radon–Nikodym derivatives)
Let μ, ν, λ be σ-finite measures on $(\Omega, \mathcal{F})$ such that $\nu \ll \mu$, $\mu \ll \lambda$.

1) If $g \in L^1(\Omega, \mathcal{F}, \nu)$ then $g\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^1(\Omega, \mathcal{F}, \mu)$ and $\int_\Omega g \,\mathrm{d}\nu \underset{(*)}{=} \int_\Omega g\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu$.
2) $\nu \ll \lambda$ and $\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ λ-a.e.

Proof.

1) We show both statements by applying the standard argument: First, $(*)$ holds for $g = \mathbb{1}_A$, $A \in \mathcal{F}$, since
$$\int_\Omega g\,\mathrm{d}\nu \underset{\text{def.}}{=} \mathbb{E}(\mathbb{1}_A) \underset{\text{simple}}{=} \nu(A) \overset{\nu \ll \mu}{\underset{\text{C. 5.41}}{=}} \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu \underset{\text{def.}}{=} \int_\Omega \mathbb{1}_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu \underset{\text{def.}}{=} \int_\Omega g\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu.$$
By linearity of the integral, $(*)$ also holds for all simple functions. By MON, it holds for all $g \in L_+$. Again by linearity, it holds for all $g \in L^1(\Omega, \mathcal{F}, \nu)$.

2) $\lambda(A) = 0 \underset{\mu \ll \lambda}{\Rightarrow} \mu(A) = 0 \underset{\nu \ll \mu}{\Rightarrow} \nu(A) = 0$, so $\nu \ll \lambda$. Applying $(*)$ to $\nu \leftarrow \mu$, $\mu \leftarrow \lambda$, $g \leftarrow \mathbb{1}_A\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$, we obtain
$$\int_A \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\,\mathrm{d}\lambda \overset{\nu \ll \lambda}{\underset{\text{RN}}{=}} \nu(A) \overset{\nu \ll \mu}{\underset{\text{RN}}{=}} \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu \underset{\text{def.}}{=} \int_\Omega \mathbb{1}_A\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu \underset{(*)}{=} \int_\Omega \mathbb{1}_A\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\cdot\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\,\mathrm{d}\lambda$$
$$\underset{\text{def.}}{=} \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\cdot\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\,\mathrm{d}\lambda \quad \forall\, A \in \mathcal{F}$$
and so the claim follows by P. 5.14 1) $\Rightarrow$ 3). □

Remark 5.43

1) If the distribution F of ν in the RN theorem is absolutely continuous wrt Lebesgue measure λ (μ in the RN theorem), then $F(B) \underset{\text{RN}}{=} \int_B f \,\mathrm{d}\lambda \;\forall\, B \in \mathcal{B}(\mathbb{R}^d)$, so the RN derivative f is the density of F; in differential form: $\mathrm{d}F = f\,\mathrm{d}\lambda$.

2) For $B = (-\infty, \boldsymbol{x}]$, we obtain $F(\boldsymbol{x}) \underset{\text{notation}}{=} F((-\infty, \boldsymbol{x}]) = \int_{(-\infty,\boldsymbol{x}]} f\,\mathrm{d}\lambda \underset{\text{notation}}{=} \int_{(-\infty,\boldsymbol{x}]} f(\tilde{\boldsymbol{x}})\,\mathrm{d}\tilde{\boldsymbol{x}}$; in differential form: $\mathrm{d}F(\boldsymbol{x}) = f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. Hence $\int_{\mathbb{R}} g(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x}) = \int_{\mathbb{R}} g(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$.

3) We recognize from P. 5.42 1) the measure change $\mathbb{E}_\nu(g) = \mathbb{E}_\mu(g\frac{\mathrm{d}\nu}{\mathrm{d}\mu})$ if $\nu \ll \mu$. This is applied in importance sampling in statistics. If $\boldsymbol{X} \sim f$ and $\tilde{\boldsymbol{X}} \sim h$, then

$$\mathbb{E}_f(g(\boldsymbol{X})) \underset{1)}{=} \int_{\mathbb{R}^d} g(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \underset{(**)}{=} \int_{\mathbb{R}^d} \frac{g(\boldsymbol{x})f(\boldsymbol{x})}{h(\boldsymbol{x})} h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mathbb{E}_h\biggl(\frac{g(\tilde{\boldsymbol{X}})f(\tilde{\boldsymbol{X}})}{h(\tilde{\boldsymbol{X}})}\biggr),$$

where $(**)$ holds for all densities h such that $h(\boldsymbol{x}) = 0$ implies $f(\boldsymbol{x}) = 0$ (so $f(\boldsymbol{x}) > 0 \Rightarrow h(\boldsymbol{x}) > 0$, i.e. $\mathrm{supp}(f) \subseteq \mathrm{supp}(h)$); note that for $h(\boldsymbol{x}) = 0$, we interpret $\frac{g(\boldsymbol{x})f(\boldsymbol{x})}{h(\boldsymbol{x})} \cdot h(\boldsymbol{x}) = \frac{0}{0} \cdot 0$ as 0.

References

Folland, G. B. (1999), Real Analysis: Modern Techniques and Their Applications, Wiley-Interscience.

Klenke, A. (2008), Probability Theory: A Comprehensive Course, Springer.

ter Horst, H. J. (1984), Riemann-Stieltjes and Lebesgue-Stieltjes Integrability, *The American Mathematical Monthly*, 91(9), 551–559.

6 Modes of convergence

6.1 Almost sure convergence and in probability

Definition 6.1 (Complete convergence, almost sure and in probability)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\boldsymbol{X}, (\boldsymbol{X}_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^d$, $d \geq 1$, be random vectors.

1) $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ *converges completely (c.c.)* to $\boldsymbol{X}$ (notation: $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{c.c.}}{\to}} \boldsymbol{X}$) if $\forall\,\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) < \infty.$$

2) $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ converges to $\boldsymbol{X}$ *almost surely (a.s.)* (notation: $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{X}$) if

$$\mathbb{P}(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}) = 1.$$

3) $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ converges to $\boldsymbol{X}$ *in probability (i.p.)* (notation: $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X}$) if $\forall\,\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) = 0.$$

Lemma 6.2 (Auxiliary results for characterization of a.s. convergence)
Let $\varepsilon > 0$.

1) Let $n \in \mathbb{N}$. Then $\sup_{k\ge n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\} = \{\sup_{k\ge n}\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}$. And $\limsup_{k\ge n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\} = \{\limsup_{k\ge n}\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}$.
2) $\lim_{n\to\infty}\mathbb{P}(\sup_{k\ge n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}) = \mathbb{P}(\limsup_{n\to\infty}\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon\})$.

Proof.

1) $\omega \in \sup_{k\ge n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\} \underset{\text{def.}}{=} \bigcup_{k\ge n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}$ iff $\exists\, k \ge n :$ $\|\boldsymbol{X}_k(\omega) - \boldsymbol{X}(\omega)\| > \varepsilon$ iff $\sup_{k\ge n}\|\boldsymbol{X}_k(\omega) - \boldsymbol{X}(\omega)\| > \varepsilon$ (note that the former $\sup$ is the set-theoretic one, the latter is the analytic one). Similarly for $\limsup$.
2) $A_{n,\varepsilon} := \sup_{k\ge n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\} \underset{\text{def.}}{=} \bigcup_{k\ge n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\} \searrow \Rightarrow A_{n,\varepsilon} \searrow \bigcap_{n=1}^{\infty} A_{n,\varepsilon} \underset{\text{def.}}{=} \limsup_{n\to\infty}\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon\} =: A_{\infty,\varepsilon} \Rightarrow \lim_{n\to\infty}\mathbb{P}(A_{n,\varepsilon}) \overset{\text{cont.}}{\underset{\text{above}}{=}} \mathbb{P}(A_{\infty,\varepsilon})$. □

Theorem 6.3 (Characterization of a.s. convergence by i.p.)
$\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{X}$ iff $\sup_{k\geq n}\|\boldsymbol{X}_k-\boldsymbol{X}\| \underset{n\to\infty}{\overset{\text{p}}{\to}} 0$.

Proof.

"⇒": $\exists\, N\in\mathcal{F}: \mathbb{P}(N)=0$ and $\boldsymbol{X}_n(\omega) \underset{n\to\infty}{\to} \boldsymbol{X}(\omega)\ \forall\,\omega\in N^c$. Let $\varepsilon>0$. $\underset{\text{def.}}{\Rightarrow}$ $\forall\,\omega\in N^c$, $\exists\, m\in\mathbb{N}: \|\boldsymbol{X}_n(\omega)-\boldsymbol{X}(\omega)\|<\varepsilon\ \forall\, n\geq m$. Let $A_{n,\varepsilon}:=\bigcup_{k\geq n}\{\|\boldsymbol{X}_k-\boldsymbol{X}\|>\varepsilon\}$ and $A_{\infty,\varepsilon}:=\bigcap_{n=1}^\infty A_{n,\varepsilon}$. Then $\forall\,\omega\in N^c\ \exists\, n\in\mathbb{N}$: $\omega\notin A_{n,\varepsilon}\Rightarrow \omega\notin A_{\infty,\varepsilon}$, so $A_{\infty,\varepsilon}\subseteq N$. Thus $\lim_{n\to\infty}\mathbb{P}(|\sup_{k\geq n}\|\boldsymbol{X}_k-\boldsymbol{X}\| -0|>\varepsilon)=\lim_{n\to\infty}\mathbb{P}(\sup_{k\geq n}\|\boldsymbol{X}_k-\boldsymbol{X}\|>\varepsilon) \underset{\text{L. 6.2 1)}}{=} \lim_{n\to\infty}\mathbb{P}(\sup_{k\geq n}\{\|\boldsymbol{X}_k-\boldsymbol{X}\|>\varepsilon\}) \underset{\text{L. 6.2 2)}}{=} \mathbb{P}(\limsup_{n\to\infty}\{\|\boldsymbol{X}_k-\boldsymbol{X}\|>\varepsilon\}) \underset{\limsup}{\overset{\text{def.}}{=}} \mathbb{P}(A_{\infty,\varepsilon})\leq\mathbb{P}(N)=0$.

"⇐": $\mathbb{P}(\lim_{n\to\infty}\boldsymbol{X}_n\neq\boldsymbol{X})=\mathbb{P}(\lim_{n\to\infty}\|\boldsymbol{X}_n-\boldsymbol{X}\|\neq 0)=\mathbb{P}(\limsup_{n\to\infty}\|\boldsymbol{X}_n-\boldsymbol{X}\|>0)$

$$=\mathbb{P}\Bigl(\bigcup_{m=1}^\infty\Bigl\{\limsup_{n\to\infty}\|\boldsymbol{X}_n-\boldsymbol{X}\|>\frac{1}{m}\Bigr\}\Bigr) \underset{\sigma\text{-subadd.}}{\leq} \sum_{m=1}^\infty\mathbb{P}\Bigl(\Bigl\{\limsup_{n\to\infty}\|\boldsymbol{X}_n-\boldsymbol{X}\|>\frac{1}{m}\Bigr\}\Bigr)$$

$$\underset{\text{L. 6.2 1)}}{=}\sum_{m=1}^\infty\mathbb{P}\Bigl(\limsup_{n\to\infty}\Bigl\{\|\boldsymbol{X}_n-\boldsymbol{X}\|>\frac{1}{m}\Bigr\}\Bigr) \underset{\text{L. 6.2 2)}}{=} \sum_{m=1}^\infty\lim_{n\to\infty}\mathbb{P}\Bigl(\sup_{k\geq n}\Bigl\{\|\boldsymbol{X}_n-\boldsymbol{X}\|>\frac{1}{m}\Bigr\}\Bigr)$$

$$\underset{\text{ass.}}{=}\sum_{m=1}^\infty 0=0.$$

□

Several important special cases appear from T. 6.3.

Corollary 6.4 (Complete convergence implies a.s.)

$\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{c.c.}}{\to}} \boldsymbol{X} \Rightarrow \boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{X}$.

Proof. $\sum_{n=1}^{\infty} \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) < \infty \Rightarrow \lim_{n\to\infty} \mathbb{P}(\sup_{k\ge n}\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon) \underset{\text{L. 6.2 1)}}{=} \lim_{n\to\infty} \mathbb{P}(\sup_{k\ge n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}) \underset{\text{L. 6.2 2)}}{=} \mathbb{P}(\limsup_{n\to\infty}\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon\}) \underset{\text{BC1}}{=} 0 \underset{\text{T. 6.3}}{\Rightarrow} \checkmark$. □

Corollary 6.5 (A.s. convergence implies i.p.)

$\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{X} \Rightarrow \boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X}$.

Proof. $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{X} \underset{\text{T. 6.3}}{\Rightarrow} \sup_{k\ge n}\|\boldsymbol{X}_k - \boldsymbol{X}\| \underset{n\to\infty}{\overset{\text{p}}{\to}} 0 \underset{\text{def.}}{\Rightarrow} \|\boldsymbol{X}_n - \boldsymbol{X}\| \underset{n\to\infty}{\overset{\text{p}}{\to}} 0 \underset{\text{def.}}{\Rightarrow} \boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X}$. □

Corollary 6.6 (A.s. monotone and convergence i.p. implies a.s.)

$\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X}$ and $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ a.s. monotone $\Rightarrow \boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{X}$.

Proof. $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X} \underset{\text{def.}}{\Rightarrow} \|\boldsymbol{X}_n - \boldsymbol{X}\| \underset{n\to\infty}{\overset{\text{p}}{\to}} 0$. Therefore, $\sup_{k\ge n}\|\boldsymbol{X}_k - \boldsymbol{X}\| \underset{\text{mon.}}{=} \|\boldsymbol{X}_n - \boldsymbol{X}\| \underset{n\to\infty}{\overset{\text{p}}{\to}} 0 \underset{\text{T. 6.3}}{\Rightarrow} \boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{X}$. □

Theorem 6.7 (Subsequence principle)

$\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X}$ iff $\forall\,(\boldsymbol{X}_{n_k})_{k\in\mathbb{N}} \subseteq (\boldsymbol{X}_n)_{n\in\mathbb{N}}\ \exists\,(\boldsymbol{X}_{n_{k_l}})_{l\in\mathbb{N}} \subseteq (\boldsymbol{X}_{n_k})_{k\in\mathbb{N}} : \boldsymbol{X}_{n_{k_l}} \underset{l\to\infty}{\overset{\text{c.c. or a.s.}}{\to}} \boldsymbol{X}$.

Proof.

"$\Rightarrow$": $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X} \Rightarrow \boldsymbol{X}_{n_k} \underset{k\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X}\ \forall\,(\boldsymbol{X}_{n_k})_{k\in\mathbb{N}} \subseteq (\boldsymbol{X}_n)_{n\in\mathbb{N}} \underset{\text{def.}}{\Rightarrow} \exists\,(n_{k_l})_{l\in\mathbb{N}} \subseteq (n_k)_{k\in\mathbb{N}} : \mathbb{P}(\|\boldsymbol{X}_{n_{k_l}} - \boldsymbol{X}\| > 1/l) \le 2^{-l}\ \forall\, l$. Thus, $\forall\,\varepsilon > 0$,

$$\sum_{l=1}^{\infty} \mathbb{P}(\|\boldsymbol{X}_{n_{k_l}} - \boldsymbol{X}\| > \varepsilon) = \underbrace{\sum_{\substack{l\in\mathbb{N}:\\ 1/l>\varepsilon}} \underbrace{\mathbb{P}(\|\boldsymbol{X}_{n_{k_l}} - \boldsymbol{X}\| > \varepsilon)}_{\le 1}}_{\le\ \lceil 1/\varepsilon\rceil\cdot 1\ \text{(finite)}} + \underbrace{\sum_{\substack{l\in\mathbb{N}:\\ 1/l\le\varepsilon}} \underbrace{\mathbb{P}(\|\boldsymbol{X}_{n_{k_l}} - \boldsymbol{X}\| > \varepsilon)}_{\le \mathbb{P}(\|\boldsymbol{X}_{n_{k_l}} - \boldsymbol{X}\| > 1/l) \le 2^{-l}}}_{\le \sum_{l=1}^{\infty} 2^{-l} = 1} < \infty.$$

$\underset{\text{def.}}{\Rightarrow} \boldsymbol{X}_{n_{k_l}} \underset{n\to\infty}{\overset{\text{c.c.}}{\to}} \boldsymbol{X} \underset{\text{C. 6.4}}{\Rightarrow} \boldsymbol{X}_{n_{k_l}} \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{X}$.

"$\Leftarrow$": Suppose $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\not\to}} \boldsymbol{X} \underset{\text{def.}}{\Rightarrow} \exists\,\varepsilon_1 > 0,\ \varepsilon_2 > 0, (\boldsymbol{X}_{n_k})_{k\in\mathbb{N}} \subseteq (\boldsymbol{X}_n)_{n\in\mathbb{N}}$: $\mathbb{P}(\|\boldsymbol{X}_{n_k} - \boldsymbol{X}\| > \varepsilon_1) \underset{(*)}{\ge} \varepsilon_2\ \forall\, k \underset{\text{ass.}}{\Rightarrow} \exists\,(\boldsymbol{X}_{n_{k_l}})_{l\in\mathbb{N}} \subseteq (\boldsymbol{X}_{n_k})_{k\in\mathbb{N}} : \boldsymbol{X}_{n_{k_l}} \underset{l\to\infty}{\overset{\text{c.c. or a.s.}}{\to}} \boldsymbol{X}$ $\underset{\text{C. 6.4, C. 6.5}}{\Rightarrow} \boldsymbol{X}_{n_{k_l}} \underset{l\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X}$ ↯ to $(*)$. □

Lemma (Equivalent condition for convergence)

$a_n \underset{n\to\infty}{\to} a$ iff $\forall\,(a_{n_k})_{k\in\mathbb{N}} \subseteq (a_n)_{n\in\mathbb{N}},\ \exists\,(a_{n_{k_l}})_{l\in\mathbb{N}} \subseteq (a_{n_k})_{k\in\mathbb{N}} : a_{n_{k_l}} \underset{n\to\infty}{\to} a$.

Proof. "$\Rightarrow$": ✓. "$\Leftarrow$": Suppose $a_n \underset{n\to\infty}{\not\to} a \Rightarrow \exists\,\varepsilon > 0$ and $(a_{n_k})_{k\in\mathbb{N}} \subseteq (a_n)_{n\in\mathbb{N}}$: $|a_{n_k} - a| > \varepsilon\ \forall\, k \in \mathbb{N}$. But by ass., $\exists\,(a_{n_{k_l}})_{l\in\mathbb{N}} \subseteq (a_{n_k})_{k\in\mathbb{N}} : a_{n_{k_l}} \underset{l\to\infty}{\to} a$ ↯ □

The subsequence principle allows us to extend DOM on probability spaces from almost sure convergence to convergence in probability.

Corollary 6.8 (DOM for convergence i.p.)

If $X_n \underset{n\to\infty}{\overset{\text{p}}{\to}} X$, $|X_n| \overset{\text{a.s.}}{\leq} Y\ \forall\, n \in \mathbb{N}$ for $Y \in L^1$, then $(X_n) \subseteq L^1$, $X \in L^1$ and $\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$.

Proof.

- $X_n \underset{n\to\infty}{\overset{\text{p}}{\to}} X \underset{\text{principle}}{\overset{\text{subseq.}}{\Rightarrow}} \forall\,(X_{n_k})_{k\in\mathbb{N}} \subseteq (X_n)_{n\in\mathbb{N}},\ \exists\,(X_{n_{k_l}})_{l\in\mathbb{N}} \subseteq (X_{n_k})_{k\in\mathbb{N}} : X_{n_{k_l}} \underset{l\to\infty}{\overset{\text{a.s.}}{\to}} X \underset{\text{DOM}}{\Rightarrow} (X_n) \subseteq L^1$, $X \in L^1$ and $\lim_{l\to\infty} \mathbb{E}(X_{n_{k_l}}) = \mathbb{E}(X)$.
- We have thus shown that $\forall\,(\mathbb{E}(X_{n_k}))_{k\in\mathbb{N}} \subseteq (\mathbb{E}(X_n))_{n\in\mathbb{N}},\ \exists\,(\mathbb{E}(X_{n_{k_l}}))_{l\in\mathbb{N}} \subseteq (\mathbb{E}(X_{n_k}))_{k\in\mathbb{N}} : \mathbb{E}(X_{n_{k_l}}) \underset{l\to\infty}{\to} \mathbb{E}(X)$. By the above lemma, we thus have $\mathbb{E}(X_n) \underset{n\to\infty}{\to} \mathbb{E}(X)$. □

Theorem 6.9 (Continuous Mapping Theorem (CMT) for a.s., i.p.)
Let $\boldsymbol{X}, (\boldsymbol{X}_n)_{n\in\mathbb{N}}$ be random vectors and $\boldsymbol{h}:\mathbb{R}^d\to\mathbb{R}^k$ be continuous. Then
1) $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{X}$ implies $\boldsymbol{h}(\boldsymbol{X}_n) \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{h}(\boldsymbol{X})$; and
2) $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X}$ implies $\boldsymbol{h}(\boldsymbol{X}_n) \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{h}(\boldsymbol{X})$.

Proof.

1) $\forall\,\omega\in\Omega: \lim_{n\to\infty}\boldsymbol{X}_n(\omega)=\boldsymbol{X}(\omega)$, we have $\lim_{n\to\infty}\boldsymbol{h}(\boldsymbol{X}_n(\omega)) \underset{\boldsymbol{h}\text{ cont.}}{=} \boldsymbol{h}(\boldsymbol{X}(\omega))$.
Hence

$$\{\omega\in\Omega: \lim_{n\to\infty}\boldsymbol{X}_n(\omega)=\boldsymbol{X}(\omega)\} \subseteq \{\omega\in\Omega: \lim_{n\to\infty}\boldsymbol{h}(\boldsymbol{X}_n(\omega))=\boldsymbol{h}(\boldsymbol{X}(\omega))\},$$

so

$$\begin{aligned}\mathbb{P}(\lim_{n\to\infty}\boldsymbol{h}(\boldsymbol{X}_n)=\boldsymbol{h}(\boldsymbol{X})) &\underset{\text{def.}}{=} \mathbb{P}(\{\omega\in\Omega: \lim_{n\to\infty}\boldsymbol{h}(\boldsymbol{X}_n(\omega))=\boldsymbol{h}(\boldsymbol{X}(\omega))\})\\ &\geq \mathbb{P}(\{\omega\in\Omega: \lim_{n\to\infty}\boldsymbol{X}_n(\omega)=\boldsymbol{X}(\omega)\})\\ &\underset{\text{def.}}{=} \mathbb{P}(\lim_{n\to\infty}\boldsymbol{X}_n=\boldsymbol{X}) \underset{\text{ass.}}{=} 1.\end{aligned}$$

2) **Advanced:** By the subsequence principle, $\forall\,(\boldsymbol{X}_{n_k})_{k\in\mathbb{N}} \subseteq (\boldsymbol{X}_n)_{n\in\mathbb{N}}$, $\exists\,(\boldsymbol{X}_{n_{k_l}})_{l\in\mathbb{N}} \subseteq (\boldsymbol{X}_{n_k})_{k\in\mathbb{N}} : \boldsymbol{X}_{n_{k_l}} \underset{l\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{X} \underset{1)}{\Rightarrow} \boldsymbol{h}(\boldsymbol{X}_{n_{k_l}}) \underset{l\to\infty}{\overset{\text{a.s.}}{\to}} \boldsymbol{h}(\boldsymbol{X}) \underset{\text{applied to } (h(\boldsymbol{X}_n))_{n\in\mathbb{N}}}{\overset{\text{subsequence principle}}{\Rightarrow}} \checkmark$.

Elementary: Fix $\varepsilon > 0$. $\forall\,\delta > 0$, let $E_\delta := \{\boldsymbol{x} \in \mathbb{R}^d : \exists\,\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{y}\| < \delta,\ \|\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}(\boldsymbol{y})\| > \varepsilon\}$. By continuity of $\boldsymbol{h}$, $\lim_{\delta\to 0} E_\delta = \emptyset$. Also, $\forall\,\delta_1 < \delta_2$, $E_{\delta_1} \subseteq E_{\delta_2}$. Therefore, $\lim_{\delta\to 0} \mathbb{P}(\boldsymbol{X} \in E_\delta) \underset{\text{above}}{\overset{\text{cont.}}{=}} \mathbb{P}(\boldsymbol{X} \in \emptyset) = 0$, so

$$\begin{aligned}\mathbb{P}(\|\boldsymbol{h}(\boldsymbol{X}_n) - \boldsymbol{h}(\boldsymbol{X})\| > \varepsilon) &\underset{\text{prob.}}{\overset{\text{tot.}}{=}} \mathbb{P}(\|\boldsymbol{h}(\boldsymbol{X}_n) - \boldsymbol{h}(\boldsymbol{X})\| > \varepsilon, \|\boldsymbol{X}_n - \boldsymbol{X}\| < \delta) \\ &\quad + \mathbb{P}(\|\boldsymbol{h}(\boldsymbol{X}_n) - \boldsymbol{h}(\boldsymbol{X})\| > \varepsilon, \|\boldsymbol{X}_n - \boldsymbol{X}\| \ge \delta) \\ &\le \mathbb{P}(\boldsymbol{X} \in E_\delta) + \underbrace{\mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| \ge \delta)}_{\underset{n\to\infty}{\overset{\text{ass.}}{\to}} 0 \text{ for any fixed } \delta > 0}.\end{aligned}$$

Therefore,

$$\lim_{n\to\infty} \mathbb{P}(\|\boldsymbol{h}(\boldsymbol{X}_n) - \boldsymbol{h}(\boldsymbol{X})\| > \varepsilon) \le \mathbb{P}(\boldsymbol{X} \in E_\delta) \underset{\delta\to 0+}{\to} 0 \qquad \square$$

Similarly as in 1), involving a null set of discontinuities of $\boldsymbol{h}$, one can easily see that the CMT extends to almost surely continuous $\boldsymbol{h}$.

6.2 Convergence in L^p

The following result often provides useful bounds on tail probabilities.

Lemma 6.10 (Tail probability bounds)
Let $h : [0,\infty) \to [0,\infty)$, $\uparrow$ and X be a rv. Then

$$\mathbb{P}(|X| \ge x) \le \frac{\mathbb{E}(h(|X|))}{h(x)}, \quad x > 0.$$

Proof. For all $x > 0$, we have

$$\begin{aligned}\mathbb{P}(|X| \ge x) &\underset{h\uparrow}{=} \mathbb{P}(h(|X|) \ge h(x)) = \mathbb{E}(\mathbb{1}_{\{h(|X|)\ge h(x)\}}) \\ &\underset{\text{mon.}}{\le} \mathbb{E}\Bigl(\underbrace{\frac{h(|X|)}{h(x)}}_{\ge 1 \text{ if } \mathbb{1}_{\{\cdot\}}=1} \mathbb{1}_{\{h(|X|)\ge h(x)\}}\Bigr) \underset{\mathbb{1}_{\{\cdot\}}\le 1}{\le} \mathbb{E}\Bigl(\frac{h(|X|)}{h(x)}\Bigr) \underset{\text{lin.}}{=} \frac{\mathbb{E}(h(|X|))}{h(x)}.\end{aligned}$$

□

- For $h(x) = x$, $\mathbb{P}(|X| \ge x) \le \frac{\mathbb{E}(|X|)}{x}$ is known as *Markov's inequality*.
- For $h(x) = x^2$, $\mathbb{P}(|X| \ge x) \le \frac{\mathbb{E}(X^2)}{x^2}$ is known as *Chebyshev's inequality*.

Definition 6.11 (Convergence in L^p)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X, (X_n)_{n\in\mathbb{N}} \in L^p$, $p \in [1, \infty]$, be rvs. Then $(X_n)_{n\in\mathbb{N}}$ converges to X *in L^p* (or *in the pth mean*) (notation: $X_n \underset{n\to\infty}{\overset{L^p}{\to}} X$) if

$$\lim_{n\to\infty} \|X_n - X\|_p = 0$$

(recall: $\|X_n - X\|_p = \mathbb{E}(|X_n - X|^p)^{1/p}$, $p \in [1, \infty)$ and $\|X\|_\infty := \operatorname{ess\,sup}|X| := \inf\{x \geq 0 : \mathbb{P}(|X| > x) = 0\}$).

Although some results also hold for $p \in (0, 1)$, this case is often excluded as $\|\cdot\|_p$ does not define a norm anymore since the triangle inequality fails for $p \in (0, 1)$.

Lemma 6.12 (Higher order convergence in mean implies lower order)
$\forall\, 1 \leq p < q \leq \infty$, $X_n \underset{n\to\infty}{\overset{L^q}{\to}} X \Rightarrow X_n \underset{n\to\infty}{\overset{L^p}{\to}} X$.

Proof. $\|X_n - X\|_p \underset{\text{P. 5.19}}{\leq} \|X_n - X\|_q \cdot 1^{\frac{1}{p}-\frac{1}{q}} \underset{n\to\infty}{\overset{\text{ass.}}{\to}} 0.$ □

Convergence in L^p is often used to show convergence i.p.

Lemma 6.13 (Convergence in mean implies i.p.)

$\forall\, p \in [1, \infty]$, $X_n \underset{n\to\infty}{\overset{L^p}{\to}} X \Rightarrow X_n \underset{n\to\infty}{\overset{\text{p}}{\to}} X$.

Proof.

$p \in [1, \infty)$: $\forall\, \varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \underset{\text{L. 6.10}}{\leq} \frac{\mathbb{E}(|X_n - X|^p)}{\varepsilon^p} \underset{n\to\infty}{\to} 0$.

$p = \infty$: $\forall\, \varepsilon > 0\ \exists\, n_\varepsilon \in \mathbb{N} : \|X_n - X\|_\infty < \varepsilon\ \forall\, n \geq n_\varepsilon$. Since $|X_n - X| \underset{\text{def. }\|\cdot\|_\infty}{\overset{\text{a.s.}}{\leq}} \|X_n - X\|_\infty$ (the rhs is that level which the lhs only exceeds with probability 0), we also have $|X_n - X| \overset{\text{a.s.}}{\leq} \|X_n - X\|_\infty < \varepsilon\ \forall\, n \geq n_\varepsilon$, so $\mathbb{P}(|X_n - X| > \varepsilon) \leq \mathbb{P}(\|X_n - X\|_\infty > \varepsilon) = 0\ \forall\, n \geq n_\varepsilon$. □

Proposition 6.14 (Convergence of pth moment)

$\forall\, p \in [1, \infty]$, $X_n \underset{n\to\infty}{\overset{L^p}{\to}} X \Rightarrow \|X_n\|_p \underset{n\to\infty}{\to} \|X\|_p$.

Proof.

$$\|X_n\|_p = \|X_n - X + X\|_p \underset{\text{Minkowski}}{\leq} \|X_n - X\|_p + \|X\|_p$$

$$\|X\|_p = \|X - X_n + X_n\|_p \underset{\text{Minkowski}}{\leq} \|X_n - X\|_p + \|X_n\|_p$$

imply $\big|\|X_n\|_p - \|X\|_p\big| \leq \|X_n - X\|_p \underset{n\to\infty}{\overset{\text{ass.}}{\to}} 0$. □

6.3 Convergence in distribution

Definition 6.15 (Convergence in distribution)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\boldsymbol{X}, (\boldsymbol{X}_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^d$, $d \ge 1$, be random vectors with $\boldsymbol{X}_n \sim F_n$, $n \in \mathbb{N}$, $\boldsymbol{X} \sim F$. Then $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ converges to a random vector $\boldsymbol{X}$ *in distribution* (or *weakly*) (notation: $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{d}}{\to}} \boldsymbol{X}$ or $F_n \underset{n\to\infty}{\overset{\text{d}}{\to}} F$) if $\lim_{n\to\infty} F_n(\boldsymbol{x}) = F(\boldsymbol{x})\ \forall\, \boldsymbol{x} \in C(F)$ $:= \{\boldsymbol{x} \in \mathbb{R}^d : F$ is continuous at $\boldsymbol{x}\}$.

- Recall that $C(F)^c$ (discontinuity points) is countable, so a Lebesgue null set.
- **Uniqueness of limiting distributions.** Let $\lim_{n\to\infty} F_n(\boldsymbol{x}) = F(\boldsymbol{x})\ \forall\, \boldsymbol{x} \in C(F)$ and $\lim_{n\to\infty} F_n(\boldsymbol{x}) = \tilde{F}(\boldsymbol{x})\ \forall\, \boldsymbol{x} \in C(\tilde{F})$. Then $F(\boldsymbol{x}) = \lim_{n\to\infty} F_n(\boldsymbol{x}) = \tilde{F}(\boldsymbol{x})\ \forall\, \boldsymbol{x} \in C(F) \cap C(\tilde{F}) = (C(F)^c \cup C(\tilde{F})^c)^c =: N^c$, so that $F = \tilde{F}$ except possibly on the Lebesgue null set N. But for $\boldsymbol{x} \in N$, right-continuity of $F, \tilde{F}$ implies that $F(\boldsymbol{x}) = \lim_{\substack{\boldsymbol{z}\to\boldsymbol{x}+,\\ \boldsymbol{z}\in N^c}} F(\boldsymbol{z}) \underset{\text{on } N^c}{\overset{F=\tilde{F}}{=}} \lim_{\substack{\boldsymbol{z}\to\boldsymbol{x}+,\\ \boldsymbol{z}\in N^c}} \tilde{F}(\boldsymbol{z}) = \tilde{F}(\boldsymbol{x})$, so $F = \tilde{F}$ on $\mathbb{R}^d$.
- **Example:** Let $\boldsymbol{X} \sim F$ for continuous F, not symmetric about $\boldsymbol{0}$. Then $\boldsymbol{X}_n := (-1)^n \boldsymbol{X} \sim F_n$ for $F_n(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}) = F(\boldsymbol{x})$ for even n and $F_n(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \ge -\boldsymbol{x}) \underset{\text{cont.}}{=} \bar{F}(-\boldsymbol{x}) \ne F(\boldsymbol{x})$ for odd n, so $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{d}}{\not\to}} \boldsymbol{X}$.

- Convergence in distribution is a weaker form of convergence than a.s., i.p. or in pth mean, the $\boldsymbol{X}_n$'s do not even have to be defined on the same prob. space.

The following result provides an important characterization of convergence in distribution; see van der Vaart (2000, pp. L. 2.2) for more equivalences.

Theorem 6.16 (Portmanteau theorem)

$\boldsymbol{X}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{X}$ iff $\lim_{n\to\infty} \mathbb{E}(h(\boldsymbol{X}_n)) = \mathbb{E}(h(\boldsymbol{X}))\ \forall$ bounded and cont. $h : \mathbb{R}^d \to \mathbb{R}$.

Proof. Let $\boldsymbol{X}_n \sim F_n$, $\boldsymbol{X} \sim F$.

"$\Rightarrow$": 1) Let h be continuous and $|h(\boldsymbol{x})| \le M$, $\boldsymbol{x} \in \mathbb{R}^d$.

2) By R. 4.24 4) ($\Rightarrow F$ only jumps where its margins jump) and R. 3.16 3) ($\forall\, j$, F_j has at most countably many jumps), the set of discontinuities of F is of the form $D = \prod_{j=1}^d D_j$, where D_j is the (at most countable) set of discontinuities of F_j, a Lebesgue null set. Since $\Delta_{(\boldsymbol{a},\boldsymbol{b}]}F \to 1$ for $\boldsymbol{a} \to -\infty$ and $\boldsymbol{b} \to \infty$, we find for any $\varepsilon > 0$, $\boldsymbol{a}, \boldsymbol{b} \in D^c : \Delta_{[\boldsymbol{a},\boldsymbol{b}]}F \ge 1 - \frac{\varepsilon}{6M}$; note that $\Delta_{[\boldsymbol{a},\boldsymbol{b}]}F = \Delta_{(\boldsymbol{a},\boldsymbol{b}]}F$ as $\boldsymbol{a}$ is a continuity point of F. For $I := [\boldsymbol{a}, \boldsymbol{b}]$, we thus have $\mathbb{P}(\boldsymbol{X} \in I^c) = 1 - \Delta_{[\boldsymbol{a},\boldsymbol{b}]}F \le \frac{\varepsilon}{6M}$.

3) h cont. on the compact $I \Rightarrow h$ uniformly continuous on I, so $\exists\, m \in \mathbb{N}$ and a partition $I = \biguplus_{k=1}^m I_k$ for rectangles I_k with endpoints in D^c such that $\sup_{\boldsymbol{x},\boldsymbol{y}\in I_k} |h(\boldsymbol{x}) - h(\boldsymbol{y})| \le \frac{\varepsilon}{6}$. For $k = 1, \dots, m$, pick $\boldsymbol{x}_k \in I_k$ and define the piecewise constant $h_\varepsilon(\boldsymbol{x}) := \sum_{k=1}^m h(\boldsymbol{x}_k)\mathbb{1}_{I_k}(\boldsymbol{x})$. Then $|h(\boldsymbol{x}) - h_\varepsilon(\boldsymbol{x})| \le \frac{\varepsilon}{6}\ \forall\, \boldsymbol{x} \in I$.

4) $$|\mathbb{E}(h(\boldsymbol{X})) - \mathbb{E}(h_\varepsilon(\boldsymbol{X}))| \underset{\text{lin.}}{=} |\mathbb{E}(h(\boldsymbol{X}) - h_\varepsilon(\boldsymbol{X}))| \underset{\Delta}{\le} \mathbb{E}(|h(\boldsymbol{X}) - h_\varepsilon(\boldsymbol{X})|)$$
$$\underset{\text{lin.}}{=} \mathbb{E}(|h(\boldsymbol{X}) - h_\varepsilon(\boldsymbol{X})|\mathbb{1}_I(\boldsymbol{X})) + \mathbb{E}(|h(\boldsymbol{X}) - h_\varepsilon(\boldsymbol{X})|\mathbb{1}_{I^c}(\boldsymbol{X}))$$
$$\underset{3)}{\le} \mathbb{E}\Bigl(\frac{\varepsilon}{6}\mathbb{1}_I(\boldsymbol{X})\Bigr) + \mathbb{E}(|h(\boldsymbol{X}) - 0|\mathbb{1}_{I^c}(\boldsymbol{X})) \underset{1)}{\le} \frac{\varepsilon}{6}\cdot 1 + M\underbrace{\mathbb{P}(\boldsymbol{X} \in I^c)}_{\underset{2)}{\le}\ \frac{\varepsilon}{6M}} \le \frac{\varepsilon}{3}.$$

5) Recall that $\mathbb{P}(\boldsymbol{X}_n \in I^c) \underset{\text{ass.}}{\overset{\text{endpoints of } I \text{ in } D^c}{\longrightarrow}} \mathbb{P}(\boldsymbol{X} \in I^c) \underset{2)}{\le} \frac{\varepsilon}{6M}$. For all n sufficiently large such that $\mathbb{P}(\boldsymbol{X}_n \in I^c) \le \frac{\varepsilon}{3M}$, we thus obtain similarly as in 4) that $|\mathbb{E}(h(\boldsymbol{X}_n)) - \mathbb{E}(h_\varepsilon(\boldsymbol{X}_n))| \le \frac{\varepsilon}{6} + M\mathbb{P}(\boldsymbol{X}_n \in I^c) \le \frac{\varepsilon}{6} + \frac{\varepsilon}{3} = \frac{\varepsilon}{2}$.

6) For all n sufficiently large, we thus have
$$|\mathbb{E}(h_\varepsilon(\boldsymbol{X}_n)) - \mathbb{E}(h_\varepsilon(\boldsymbol{X}))| \underset{\text{lin.}}{=} \Bigl|\sum_{k=1}^m h(\boldsymbol{x}_k)(\mathbb{P}(\boldsymbol{X}_n \in I_k) - \mathbb{P}(\boldsymbol{X} \in I_k))\Bigr|$$

$$\underset{\triangle}{\leq} \sum_{k=1}^{m} |h(\boldsymbol{x}_k)| \cdot \underbrace{|\mathbb{P}(\boldsymbol{X}_n \in I_k) - \mathbb{P}(\boldsymbol{X} \in I_k)|}_{\underset{\text{ass.}}{\overset{\text{endpoints of } I_k \text{ in } D^c}{\longrightarrow}} 0} \underset{n \text{ large}}{\leq} \frac{\varepsilon}{6}.$$

7) For n sufficiently large, we thus have

$$\begin{aligned}
&|\mathbb{E}(h(\boldsymbol{X}_n)) - \mathbb{E}(h(\boldsymbol{X}))| \\
&= |\mathbb{E}(h(\boldsymbol{X}_n)) - \mathbb{E}(h_\varepsilon(\boldsymbol{X}_n)) + \mathbb{E}(h_\varepsilon(\boldsymbol{X}_n)) - \mathbb{E}(h_\varepsilon(\boldsymbol{X})) + \mathbb{E}(h_\varepsilon(\boldsymbol{X})) - \mathbb{E}(h(\boldsymbol{X}))| \\
&\underset{\triangle}{\leq} |\mathbb{E}(h(\boldsymbol{X}_n)) - \mathbb{E}(h_\varepsilon(\boldsymbol{X}_n))| + |\mathbb{E}(h_\varepsilon(\boldsymbol{X}_n)) - \mathbb{E}(h_\varepsilon(\boldsymbol{X}))| + |\mathbb{E}(h_\varepsilon(\boldsymbol{X})) - \mathbb{E}(h(\boldsymbol{X}))| \\
&\underset{5),6),4)}{\leq} \frac{\varepsilon}{2} + \frac{\varepsilon}{6} + \frac{\varepsilon}{3} = \varepsilon
\end{aligned}$$

"$\Leftarrow$": i) Let $\boldsymbol{x} \in C(F)$. For $\boldsymbol{\varepsilon} > \boldsymbol{0}$, consider the multilinear function $h_{\boldsymbol{\varepsilon}}(\boldsymbol{z}) = \prod_{j=1}^{d} \max\{\min\{\frac{x_j - z_j}{\varepsilon_j}, 1\}, 0\}$, $\boldsymbol{z} \in \mathbb{R}^d$, which satisfies

$$\mathbb{1}_{(-\infty, \boldsymbol{x}-\boldsymbol{\varepsilon}]}(\boldsymbol{z}) \leq h_{\boldsymbol{\varepsilon}}(\boldsymbol{z}) \leq \mathbb{1}_{(-\infty, \boldsymbol{x}]}(\boldsymbol{z}), \quad \boldsymbol{z} \in \mathbb{R}^d.$$

Therefore, $F(\boldsymbol{x} - \boldsymbol{\varepsilon}) = \mathbb{E}(\mathbb{1}_{(-\infty, \boldsymbol{x}-\boldsymbol{\varepsilon}]}(\boldsymbol{X})) \leq \mathbb{E}(h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}))$ and $\mathbb{E}(h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}_n)) \leq \mathbb{E}(\mathbb{1}_{(-\infty, \boldsymbol{x}]}(\boldsymbol{X}_n)) = F_n(\boldsymbol{x})$. This implies that

$$\liminf_{n\to\infty} F_n(\boldsymbol{x}) \geq \liminf_{n\to\infty} \mathbb{E}(h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}_n)) \underset{\text{ass.}}{=} \lim_{n\to\infty} \mathbb{E}(h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}_n)) \underset{\text{ass.}}{=} \mathbb{E}(h_{\boldsymbol{\varepsilon}}(\boldsymbol{X})) \geq F(\boldsymbol{x} - \boldsymbol{\varepsilon}).$$

Hence $F(\boldsymbol{x}) \underset{\boldsymbol{x}\,\in\,C(F)}{=} \lim_{\boldsymbol{\varepsilon}\to\boldsymbol{0}} F(\boldsymbol{x}-\boldsymbol{\varepsilon}) \le \liminf_{n\to\infty} F_n(\boldsymbol{x})$.

ii) Similarly, the multilinear function $h_{\boldsymbol{\varepsilon}}(\boldsymbol{z}) = \prod_{j=1}^d \max\{\min\{\frac{x_j+\varepsilon_j-z_j}{\varepsilon_j},1\},0\}$, $\boldsymbol{z}\in\mathbb{R}^d$, satisfies

$$\mathbb{1}_{(-\boldsymbol{\infty},\boldsymbol{x}]}(\boldsymbol{z}) \le h_{\boldsymbol{\varepsilon}}(\boldsymbol{z}) \le \mathbb{1}_{(-\boldsymbol{\infty},\boldsymbol{x}+\boldsymbol{\varepsilon}]}(\boldsymbol{z}), \quad \boldsymbol{z}\in\mathbb{R}^d.$$

Therefore, $F_n(\boldsymbol{x}) = \mathbb{E}(\mathbb{1}_{(-\boldsymbol{\infty},\boldsymbol{x}]}(\boldsymbol{X}_n)) \le \mathbb{E}(h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}_n))$ and $\mathbb{E}(h_{\boldsymbol{\varepsilon}}(\boldsymbol{X})) \le \mathbb{E}(\mathbb{1}_{(-\boldsymbol{\infty},\boldsymbol{x}+\boldsymbol{\varepsilon}]}(\boldsymbol{X})) = F(\boldsymbol{x}+\boldsymbol{\varepsilon})$. This implies that

$$\begin{aligned}\limsup_{n\to\infty} F_n(\boldsymbol{x}) &\le \limsup_{n\to\infty} \mathbb{E}(h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}_n)) \underset{\text{ass.}}{=} \lim_{n\to\infty} \mathbb{E}(h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}_n)) \underset{\text{ass.}}{=} \mathbb{E}(h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}))\\ &\le F(\boldsymbol{x}+\boldsymbol{\varepsilon}).\end{aligned}$$

Hence $\limsup_{n\to\infty} F_n(\boldsymbol{x}) \le \lim_{\boldsymbol{\varepsilon}\to\boldsymbol{0}} F(\boldsymbol{x}+\boldsymbol{\varepsilon}) \underset{\text{right-cont.}}{=} F(\boldsymbol{x})$.

We thus obtain that

$$F(\boldsymbol{x}) \underset{\text{i)}}{\le} \liminf_{n\to\infty} F_n(\boldsymbol{x}) \le \limsup_{n\to\infty} F_n(\boldsymbol{x}) \underset{\text{ii)}}{\le} F(\boldsymbol{x}).$$

□

Corollary 6.17 (Convergence i.p. implies in distribution)

$\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X} \Rightarrow \boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{d}}{\to}} \boldsymbol{X}$.

Proof.

Advanced:

- We verify the ass. of C. 6.8 (DOM). Let $h : \mathbb{R}^d \to \mathbb{R}$ be continuous and $|h| \le M < \infty$.
  - i) h continuous $\underset{\text{P. 3.7}}{\Rightarrow}$ h is measurable $\underset{\text{P. 3.6}}{\Rightarrow}$ $h(\boldsymbol{X}_n)$ is measurable $\forall\, n \in \mathbb{N}$, and $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{X} \underset{\text{CMT}}{\Rightarrow} h(\boldsymbol{X}_n) \underset{n\to\infty}{\overset{\text{p}}{\to}} h(\boldsymbol{X})$.
  - ii) $|h(\boldsymbol{X}_n)| \le M$, $n \in \mathbb{N}$.
  - iii) The constant function M is integrable.
- By C. 6.8, $(h(\boldsymbol{X}_n)) \subseteq L^1$, $h(\boldsymbol{X}) \in L^1$ and $\lim_{n\to\infty} \mathbb{E}(h(\boldsymbol{X}_n)) = \mathbb{E}(h(\boldsymbol{X}))$ $\underset{\text{portmanteau}}{\Rightarrow}$ ✓

Elementary: Let $\boldsymbol{X}_n \sim F_n$, $\boldsymbol{X} \sim F$, $\boldsymbol{x} \in C(F)$, $\varepsilon > 0$, $\boldsymbol{\varepsilon} = (\varepsilon, \dots, \varepsilon)$.

i) $F_n(\boldsymbol{x}) \underset{\text{prob.}}{\overset{\text{tot.}}{=}} \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x},\ \|\boldsymbol{X}_n - \boldsymbol{X}\| \le \varepsilon) + \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x},\ \|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon)$

$$\le \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x} + \boldsymbol{\varepsilon}) + \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) = F(\boldsymbol{x} + \boldsymbol{\varepsilon}) + \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon),$$

hence $\limsup_{n\to\infty} F_n(\boldsymbol{x}) \le F(\boldsymbol{x} + \boldsymbol{\varepsilon})$.

ii) And

$$F(\boldsymbol{x}-\boldsymbol{\varepsilon}) \underset{\text{prob.}}{\overset{\text{tot.}}{=}} \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}-\boldsymbol{\varepsilon},\ \|\boldsymbol{X}_n-\boldsymbol{X}\| \le \varepsilon) + \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}-\boldsymbol{\varepsilon},\ \|\boldsymbol{X}_n-\boldsymbol{X}\| > \varepsilon)$$

$$\le \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x}) + \mathbb{P}(\|\boldsymbol{X}_n-\boldsymbol{X}\| > \varepsilon) = F_n(\boldsymbol{x}) + \mathbb{P}(\|\boldsymbol{X}_n-\boldsymbol{X}\| > \varepsilon),$$

hence $F(\boldsymbol{x}-\boldsymbol{\varepsilon}) \le \liminf_{n\to\infty} F_n(\boldsymbol{x})$.

We thus obtain that

$$F(\boldsymbol{x}-\boldsymbol{\varepsilon}) \underset{\text{ii)}}{\le} \liminf_{n\to\infty} F_n(\boldsymbol{x}) \le \limsup_{n\to\infty} F_n(\boldsymbol{x}) \underset{\text{i)}}{\le} F(\boldsymbol{x}+\boldsymbol{\varepsilon}).$$

Now let $\varepsilon \to 0+$ to see that $\lim_{n\to\infty} F_n(\boldsymbol{x}) = F(\boldsymbol{x})$. □

Corollary 6.18 (Convergence in distribution to a constant implies i.p.)

$\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{d}}{\to}} \boldsymbol{c} \in \mathbb{R}^d \Rightarrow \boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{p}}{\to}} \boldsymbol{c}$.

Proof.

- For $r \in (0,1]$, $\frac{\sum_{j=1}^d |x_j|^r}{(\sum_{j=1}^d |x_j|)^r} = \sum_{j=1}^d \Bigl(\frac{|x_j|}{\sum_{k=1}^d |x_k|}\Bigr)^r \ge \sum_{j=1}^d \frac{|x_j|}{\sum_{k=1}^d |x_k|} = 1$, so $(\sum_{j=1}^d |x_j|)^r \overset{(*)}{\le} \sum_{j=1}^d |x_j|^r$. Therefore, for $1 \le p \le q < \infty$, $\|\boldsymbol{x}\|_q =$

$\left(\sum_{j=1}^d |x_j|^q\right)^{\frac{p}{qp}} \underset{r=p/q\in(0,1]}{\overset{(*)}{\leq}} \left(\sum_{j=1}^d |x_j|^{q\frac{p}{q}}\right)^{\frac{1}{p}} = \left(\sum_{j=1}^d |x_j|^p\right)^{\frac{1}{p}} = \|\boldsymbol{x}\|_p$. In particular, $\|\cdot\|_2 \leq \|\cdot\|_1$.

- Note that the df of the degenerate random vector $\boldsymbol{c}$ is $F(\boldsymbol{x}) = \mathbb{1}_{[\boldsymbol{c},\boldsymbol{\infty})}(\boldsymbol{x})$ with $C(F) = \mathbb{R}^d \backslash \{\boldsymbol{x} : \exists\, j : x_j = c_j,\ x_k \geq c_k\ \forall\, k \neq j\}$. Let $n \in \mathbb{N}$, $\varepsilon > 0$, $\boldsymbol{\varepsilon} = (\varepsilon, \dots, \varepsilon)$ and $\boldsymbol{c} \in \mathbb{R}^d$.
- Since $\{\omega : \max_j\{|X_{n,j}(\omega) - c_j|\} \leq \frac{\varepsilon}{d}\} \subseteq \{\omega : \|\boldsymbol{X}_n(\omega) - \boldsymbol{c}\|_1 \leq \varepsilon\} \underset{\|\cdot\|_2 \leq \|\cdot\|_1}{\subseteq} \{\omega : \|\boldsymbol{X}_n(\omega) - \boldsymbol{c}\| \leq \varepsilon\}$ we have

$$\mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{c}\| > \varepsilon) = 1 - \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{c}\| \leq \varepsilon) \leq 1 - \mathbb{P}\left(\max_j\{|X_{n,j} - c_j|\} \leq \frac{\varepsilon}{d}\right)$$
$$= 1 - \mathbb{P}\left(|X_{n,j} - c_j| \leq \frac{\varepsilon}{d}\ \forall\, j\right) = 1 - \Delta_{\left[\boldsymbol{c}-\frac{\boldsymbol{\varepsilon}}{d}, \boldsymbol{c}+\frac{\boldsymbol{\varepsilon}}{d}\right]} F_n.$$

As all endpoints of $\left[\boldsymbol{c} - \frac{\boldsymbol{\varepsilon}}{d}, \boldsymbol{c} + \frac{\boldsymbol{\varepsilon}}{d}\right]$ are in $C(F)$, we get

$$\lim_{n\to\infty} \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{c}\| > \varepsilon) = 1 - \Delta_{\left[\boldsymbol{c}-\frac{\boldsymbol{\varepsilon}}{d}, \boldsymbol{c}+\frac{\boldsymbol{\varepsilon}}{d}\right]} F \underset{\text{otherwise}}{\overset{\text{zero}}{=}} 1 - F\left(c_1 + \frac{\varepsilon}{d}, \dots, c_d + \frac{\varepsilon}{d}\right)$$
$$= 1 - 1 = 0. \qquad \square$$

Example 6.19 (Convergence i.p. to a constant $\not\Rightarrow$ a.s.)

Consider independent $(X_n)_{n\in\mathbb{N}}$ with $\mathbb{P}(X_n = 0) = 1 - 1/n$, $\mathbb{P}(X_n = n) = 1/n$.

- $\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(X_n = n) = \frac{1}{n} \underset{n\to\infty}{\longrightarrow} 0$, so $X_n \underset{n\to\infty}{\overset{\text{p}}{\longrightarrow}} 0$.
- But $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - 0| > \varepsilon) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty \underset{\text{BC2}}{\Rightarrow} \mathbb{P}(X_n = n \text{ io}) = 1$. Therefore,

$$\mathbb{P}(\lim_{n\to\infty} X_n = 0) \underset{X_n \in \{0,n\}}{=} \mathbb{P}(X_n = 0 \text{ abfm}) = 1 - \mathbb{P}(X_n = n \text{ io}) = 1 - 1 = 0,$$

so $X_n \underset{n\to\infty}{\overset{\text{a.s.}}{\not\longrightarrow}} 0$.

Theorem 6.20 (Continuous Mapping Theorem (CMT) for conv. in d.)

Let $\boldsymbol{X}, (\boldsymbol{X}_n)_{n\in\mathbb{N}}$ be random vectors and $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^k$ be continuous. Then $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{d}}{\longrightarrow}} \boldsymbol{X}$ implies $\boldsymbol{h}(\boldsymbol{X}_n) \underset{n\to\infty}{\overset{\text{d}}{\longrightarrow}} \boldsymbol{h}(\boldsymbol{X})$.

Proof. For all bounded and continuous $g : \mathbb{R}^k \to \mathbb{R}$, the composition $g \circ \boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}$ is also bounded and continuous $\underset{g\circ \boldsymbol{h} \text{ applied to } (\boldsymbol{X}_n)_n}{\overset{\text{portmanteau “}\Rightarrow\text{” with}}{\Longrightarrow}} \mathbb{E}\big(g(\boldsymbol{h}(\boldsymbol{X}_n))\big) \underset{n\to\infty}{\longrightarrow} \mathbb{E}\big(g(\boldsymbol{h}(\boldsymbol{X}))\big)$ $\underset{g \text{ applied to } (\boldsymbol{h}(\boldsymbol{X}_n))_n}{\overset{\text{portmanteau “}\Leftarrow\text{” with}}{\Longrightarrow}} \boldsymbol{h}(\boldsymbol{X}_n) \underset{n\to\infty}{\overset{\text{d}}{\longrightarrow}} \boldsymbol{h}(\boldsymbol{X})$. □

6.4 Uniform integrability

Under a.s. convergence or convergence i.p., mass can escape to $\pm\infty$ such that there is no convergence in mean (see later). Combined with uniform integrability, this can be avoided and one can obtain convergence in mean.

Definition 6.21 (Uniform integrability)

$(X_i)_{i\in I} \subseteq L^1$ is *uniformly integrable (u.i.)* if $\sup_{i\in I} \mathbb{E}(|X_i| \mathbb{1}_{\{|X_i|>a\}}) \underset{a\to\infty}{\to} 0$.

Example 6.22 (Uniform integrability)

1) If $X \in L^1$, then $|X|\mathbb{1}_{\{|X|>a\}} \underset{a\to\infty}{\to} 0$ a.e. (everywhere where X is finite, which is a.e. by L. 5.13), $|X|\mathbb{1}_{\{|X|>a\}} \le |X| \; \forall\, a \ge 0$ and $|X| \in L^1 \underset{\text{DOM}}{\Rightarrow} \lim_{a\to\infty} \mathbb{E}(|X|\mathbb{1}_{\{|X|>a\}}) = 0$, so a single $X \in L^1$ is u.i.

2) If $|X_i| \le Y \in L^1 \; \forall\, i \in I$, then $\sup_{i\in I} \mathbb{E}(|X_i|\mathbb{1}_{\{|X_i|>a\}}) \underset{\mathbb{1}_{\{|X_i|>a\}} \le \mathbb{1}_{\{Y>a\}}}{\overset{|X_i|\le Y}{\le}} \sup_{i\in I} \mathbb{E}(Y\mathbb{1}_{\{Y>a\}}) \underset{a\to\infty}{\overset{1)}{\to}} 0$, so a sequence dominated by the same $Y \in L^1$ is u.i.

3) Finitely many $(X_i)_{i=1}^n \subseteq L^1$ are u.i. by 2) with $|X_j| \le \sum_{i=1}^n |X_i| \in L^1 \; \forall\, j$.

Theorem 6.23 (Characterization of u.i.)
$(X_i)_{i\in I} \subseteq L^1$ is u.i. iff
i) $\sup_{i\in I} \mathbb{E}(|X_i|) < \infty$ (uniform bounded first absolute moments); and
ii) $\forall\, \varepsilon > 0\ \exists\, \delta > 0 : \sup_{i\in I} \mathbb{E}(|X_i|\mathbb{1}_A) < \varepsilon\ \forall\, A \in \mathcal{F} : \mathbb{P}(A) < \delta$ (uniform absolute continuity).

Proof.

"$\Rightarrow$" For any $i \in I$ and $a \in (0,\infty)$, we have $\mathbb{E}(|X_i|\mathbb{1}_A) \underset{\text{lin.}}{=} \mathbb{E}(|X_i|\mathbb{1}_{A\cap\{|X_i|\le a\}}) + \mathbb{E}(|X_i|\mathbb{1}_{A\cap\{|X_i|>a\}}) \le a\mathbb{P}(A)+\mathbb{E}(|X_i|\mathbb{1}_{\{|X_i|>a\}})$, so that $\sup_{i\in I} \mathbb{E}(|X_i|\mathbb{1}_A) \overset{(*)}{\le} a\mathbb{P}(A) + \sup_{i\in I} \mathbb{E}(|X_i|\mathbb{1}_{\{|X_i|>a\}})$. With $A = \Omega$, we get i). To get ii) for $\varepsilon > 0$, apply L. 5.13 to choose a sufficiently large so that $\sup_{i\in I} \mathbb{E}(|X_i|\mathbb{1}_{\{|X_i|>a\}}) < \frac{\varepsilon}{2}$ and let $\delta := \frac{\varepsilon/2}{a}$. Then for $\mathbb{P}(A) < \delta$, $\sup_{i\in I} \mathbb{E}(|X_i|\mathbb{1}_A) \underset{(*)}{<} a\frac{\varepsilon/2}{a} + \frac{\varepsilon}{2} = \varepsilon$.

"$\Leftarrow$" By ii) with $A = \{|X_i| > a\}$, we have that $\forall\, \varepsilon > 0\ \exists\, \delta > 0 : \mathbb{E}(|X_i|\mathbb{1}_{\{|X_i|>a\}}) < \varepsilon\ \forall\, i : \mathbb{P}(|X_i| > a) < \delta$. Since $\sup_{i\in I} \mathbb{P}(|X_i| > a) \underset{\text{Markov}}{\le} \frac{\sup_{i\in I} \mathbb{E}(|X_i|)}{a} =: \frac{c}{a} \underset{\text{i)}}{<} \infty$ choose a sufficiently large such that $\mathbb{P}(|X_i| > a) < \delta\ \forall\, i \in I$. Therefore, $\mathbb{E}(|X_i|\mathbb{1}_{\{|X_i|>a\}}) < \varepsilon\ \forall\, i \in I$, so that $\sup_{i\in I} \mathbb{E}(|X_i|\mathbb{1}_{\{|X_i|>a\}}) \le \varepsilon$. □

Lemma 6.24 (Probability of integration domain going to 0)
If $X \in L^1$ and $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{F} : \mathbb{P}(A_n) \underset{n\to\infty}{\to} 0$, then $\lim_{n\to\infty} \mathbb{E}(X\mathbb{1}_{A_n}) = 0$.

Proof. Let $\varepsilon > 0$. By E. 6.22 1) $\exists\, a > 0 : \mathbb{E}(|X|\mathbb{1}_{\{|X|>a\}}) < \varepsilon/2$. And by assumption, $\exists\, n_\varepsilon \in \mathbb{N} : \mathbb{P}(A_n) < \frac{\varepsilon}{2a}\ \forall\, n \geq n_\varepsilon$. Then

$$|\mathbb{E}(X\mathbb{1}_{A_n})| \underset{\text{Jensen}}{\leq} \mathbb{E}(|X|\mathbb{1}_{A_n}) \underset{\text{lin.}}{=} \mathbb{E}(\underbrace{|X|}_{\leq a}\mathbb{1}_{A_n\cap\{|X|\leq a\}}) + \mathbb{E}(|X|\mathbb{1}_{A_n\cap\{|X|>a\}})$$
$$\leq a\mathbb{P}(A_n) + \mathbb{E}(|X|\mathbb{1}_{\{|X|>a\}}) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \qquad \square$$

Theorem 6.25 (Convergence i.p. and u.i. imply in L^p)
Let $X, (X_n)_{n\in\mathbb{N}}$ be rvs. If $X_n \underset{n\to\infty}{\overset{\text{p}}{\to}} X$ and $(|X_n|^p)_{n\in\mathbb{N}}$ is u.i. for some $p \in [1,\infty)$, then $X_n \underset{n\to\infty}{\overset{L^p}{\to}} X$.

Proof.

1) $X_n \underset{n\to\infty}{\overset{\text{p}}{\to}} X \underset{\text{principle}}{\overset{\text{subseq.}}{\Rightarrow}} \forall\, (X_{n_k})_{k\in\mathbb{N}} \subseteq (X_n)_{n\in\mathbb{N}}\ \exists\, (X_{n_{k_l}})_{l\in\mathbb{N}} \subseteq (X_{n_k})_{k\in\mathbb{N}} : X_{n_{k_l}} \underset{l\to\infty}{\overset{\text{a.s.}}{\to}} X$.

So $\mathbb{E}(|X|^p) = \mathbb{E}(\liminf_{l\to\infty} |X_{n_{k_l}}|^p) \underset{\text{Fatou}}{\leq} \liminf_{l\to\infty} \mathbb{E}(|X_{n_{k_l}}|^p) \leq \sup_{n\in\mathbb{N}} \mathbb{E}(|X_n|^p) \underset{\text{T. 6.23 i)}}{<} \infty$ and therefore $X \in L^p$.

2) For $a, b \in \mathbb{R}$ and $p > 0$, $|a+b|^p \underset{\Delta}{\leq} (|a|+|b|)^p \leq (2\max\{|a|,|b|\})^p = 2^p \max\{|a|^p, |b|^p\} \leq 2^p(|a|^p + |b|^p)$. Therefore, $|X_n - X|^p \leq 2^p(|X_n|^p + |X|^p)$.

3) Let $\varepsilon > 0$. Then

$$\mathbb{E}(|X_n - X|^p) \underset{\text{lin.}}{=} \mathbb{E}(\underbrace{|X_n - X|^p}_{\leq \varepsilon^p} \mathbb{1}_{\{|X_n - X| \leq \varepsilon\}}) + \mathbb{E}(\underbrace{|X_n - X|^p}_{\underset{2)}{\leq} 2^p(|X_n|^p + |X|^p)} \mathbb{1}_{\{|X_n - X| > \varepsilon\}})$$

$$\underset{\text{lin.}}{\leq} \varepsilon^p \cdot 1 + 2^p \mathbb{E}(|X_n|^p \mathbb{1}_{\{|X_n - X| > \varepsilon\}}) + 2^p \mathbb{E}(|X|^p \mathbb{1}_{\{|X_n - X| > \varepsilon\}}).$$

4) $X \underset{1)}{\in} L^p \underset{\text{L. 6.24}}{\Rightarrow} \exists\, n_\varepsilon \in \mathbb{N} : \mathbb{E}(|X|^p \mathbb{1}_{\{|X_n - X| > \varepsilon\}}) < \varepsilon \;\forall\, n \geq n_\varepsilon$.

5) By ass., $(|X_i|^p)_{i \in \mathbb{N}}$ is u.i. $\underset{\text{T. 6.23}}{\Rightarrow}$ For $\varepsilon > 0$ as above, $\exists\, \delta > 0 : \sup_{i \in \mathbb{N}} \mathbb{E}(|X_i|^p \mathbb{1}_A) < \varepsilon \;\forall\, n : \mathbb{P}(A) < \delta$. Since $X_n \underset{n \to \infty}{\overset{\text{p}}{\to}} X$, $\exists\, \tilde{n}_\varepsilon \in \mathbb{N} : \mathbb{P}(|X_n - X| > \varepsilon) < \delta$ $\forall\, n \geq \tilde{n}_\varepsilon$. For such n, we can take $A = \{|X_n - X| > \varepsilon\}$ and obtain that $\mathbb{E}(|X_n|^p \mathbb{1}_{\{|X_n - X| > \varepsilon\}}) \leq \sup_{i \in \mathbb{N}} \mathbb{E}(|X_i|^p \mathbb{1}_{\{|X_n - X| > \varepsilon\}}) < \varepsilon \;\forall\, n \geq \tilde{n}_\varepsilon$.

6) Therefore, $\forall\, n \geq \max\{n_\varepsilon, \tilde{n}_\varepsilon\}$, we have

$$\mathbb{E}(|X_n - X|^p) \underset{3),4),5)}{\leq} \varepsilon^p + 2^p \varepsilon + 2^p \varepsilon. \qquad \square$$

6.5 Slutsky's theorem

According to the CMT, if $(\boldsymbol{X}_n, \boldsymbol{Y}_n) \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} (\boldsymbol{X}, \boldsymbol{Y})$ (joint convergence in distribution), then $\boldsymbol{h}(\boldsymbol{X}_n, \boldsymbol{Y}_n) \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Y})$ $\forall$ continuous h, in particular, if of the same dimension, we would also have that $\boldsymbol{X}_n + \boldsymbol{Y}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{X} + \boldsymbol{Y}$.

Question: If $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{X}$ and $\boldsymbol{Y}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{Y}$ (individual convergence in distribution), does $(\boldsymbol{X}_n, \boldsymbol{Y}_n) \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} (\boldsymbol{X}, \boldsymbol{Y})$ hold?

In general no, as there is no guarantee that the dependence structure of $(\boldsymbol{X}_n, \boldsymbol{Y}_n)$ converges to that of $(\boldsymbol{X}, \boldsymbol{Y})$; see also the following example.

Example 6.26 (Convergence in distribution not robust under summation)

Let $X \sim F$ with F symmetric about 0 (e.g. Φ), $\mathbb{P}(X = 0) \in [0, 1)$ so that $F(x) \neq \mathbb{1}_{[0,\infty)}(x)$. Then $X \overset{\mathrm{d}}{=} -X$. If $X_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} X$, then also $X_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} -X$, but $X_n + X_n = 2X_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} F(x/2) \neq \mathbb{1}_{[0,\infty)}(x) \sim X + (-X)$.

However, it holds if $\boldsymbol{X}_n, \boldsymbol{Y}_n$ and $\boldsymbol{X}, \boldsymbol{Y}$ are independent (with characteristic functions later: $\phi_{\boldsymbol{X}_n+\boldsymbol{Y}_n}(\boldsymbol{t}) \underset{\text{ind.}}{=} \phi_{\boldsymbol{X}_n}(\boldsymbol{t})\phi_{\boldsymbol{Y}_n}(\boldsymbol{t}) \underset{n\to\infty}{\overset{\text{ass.}}{\to}} \phi_{\boldsymbol{X}}(\boldsymbol{t})\phi_{\boldsymbol{Y}}(\boldsymbol{t}) \underset{\text{ind.}}{=} \phi_{\boldsymbol{X}+\boldsymbol{Y}}(\boldsymbol{t}) \underset{\text{unique}}{\Rightarrow} \checkmark$).

The next result provides another case where individual implies joint convergence.

Lemma 6.27 (Condition for joint convergence in distribution)
If $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{X}$ and $\boldsymbol{Y}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{c} \in \mathbb{R}^d$, then $(\boldsymbol{X}_n, \boldsymbol{Y}_n) \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} (\boldsymbol{X}, \boldsymbol{c})$.

Proof. Let $(\boldsymbol{X}_n, \boldsymbol{Y}_n) \sim F_n$, $(\boldsymbol{X}, \boldsymbol{Y}) \sim F(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}, \boldsymbol{c} \le \boldsymbol{y}) = F_{\boldsymbol{X}}(\boldsymbol{x}) \cdot \mathbb{1}_{[\boldsymbol{c},\boldsymbol{\infty})}(\boldsymbol{y})$, and let $(\boldsymbol{x}, \boldsymbol{y}) \in C(F)$, $\varepsilon > 0$, $\boldsymbol{\varepsilon} = (\varepsilon, \dots, \varepsilon)$.

i) $$F_n(\boldsymbol{x}, \boldsymbol{y}) \underset{\text{prob.}}{\overset{\text{tot.}}{=}} \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x}, \boldsymbol{Y}_n \le \boldsymbol{y}, \|\boldsymbol{Y}_n - \boldsymbol{c}\| \le \varepsilon) + \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x}, \boldsymbol{Y}_n \le \boldsymbol{y}, \|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon)$$
$$\le \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x},\ \boldsymbol{c} \le \boldsymbol{y} + \boldsymbol{\varepsilon}) + \mathbb{P}(\|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon)$$
$$= F_{\boldsymbol{X}_n}(\boldsymbol{x}) \mathbb{1}_{[\boldsymbol{c},\boldsymbol{\infty})}(\boldsymbol{y} + \boldsymbol{\varepsilon}) + \mathbb{P}(\|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon),$$

hence $\limsup_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) \underset{\text{C. 6.18}}{\overset{\text{ass.}}{\le}} F_{\boldsymbol{X}}(\boldsymbol{x}) \mathbb{1}_{[\boldsymbol{c},\boldsymbol{\infty})}(\boldsymbol{y} + \boldsymbol{\varepsilon}) + 0 = F(\boldsymbol{x}, \boldsymbol{y} + \boldsymbol{\varepsilon})$.

ii) As $\boldsymbol{x} \in C(F_{\boldsymbol{X}})$, $\forall\, \delta > 0\ \exists\, n_0 \in \mathbb{N} : F_{\boldsymbol{X}_n}(\boldsymbol{x}) \ge F_{\boldsymbol{X}}(\boldsymbol{x}) - \delta\ \forall\, n \ge n_0$, so that
$$F(\boldsymbol{x}, \boldsymbol{y} - \boldsymbol{\varepsilon}) = F_{\boldsymbol{X}}(\boldsymbol{x}) \mathbb{1}_{[\boldsymbol{c},\boldsymbol{\infty})}(\boldsymbol{y} - \boldsymbol{\varepsilon}) \underset{n \ge n_0}{\le} (F_{\boldsymbol{X}_n}(\boldsymbol{x}) + \delta) \mathbb{1}_{[\boldsymbol{c},\boldsymbol{\infty})}(\boldsymbol{y} - \boldsymbol{\varepsilon})$$
$$\le F_{\boldsymbol{X}_n}(\boldsymbol{x}) \mathbb{1}_{[\boldsymbol{c},\boldsymbol{\infty})}(\boldsymbol{y} - \boldsymbol{\varepsilon}) + \delta = \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x},\ \boldsymbol{c} \le \boldsymbol{y} - \boldsymbol{\varepsilon}) + \delta$$
$$\underset{\text{prob.}}{\overset{\text{tot.}}{=}} \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x},\ \boldsymbol{c} \le \boldsymbol{y} - \boldsymbol{\varepsilon},\ \|\boldsymbol{Y}_n - \boldsymbol{c}\| \le \varepsilon)$$
$$+ \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x},\ \boldsymbol{c} \le \boldsymbol{y} - \boldsymbol{\varepsilon},\ \|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon) + \delta$$

$$\le \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x},\ \boldsymbol{Y}_n \le \boldsymbol{y}) + \mathbb{P}(\|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon) + \delta$$

$$= F_n(\boldsymbol{x}, \boldsymbol{y}) + \mathbb{P}(\|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon) + \delta,$$

hence $F(\boldsymbol{x}, \boldsymbol{y} - \boldsymbol{\varepsilon}) \underset{\text{C. 6.18}}{\overset{\text{ass.}}{\le}} \liminf_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) + 0 + \delta$.

So

$$F(\boldsymbol{x}, \boldsymbol{y} - \boldsymbol{\varepsilon}) - \delta \underset{\text{ii)}}{\le} \liminf_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) \le \limsup_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) \underset{\text{i)}}{\le} F(\boldsymbol{x}, \boldsymbol{y} + \boldsymbol{\varepsilon}).$$

Let $\delta \to 0+$, we obtain

$$F(\boldsymbol{x}, \boldsymbol{y} - \boldsymbol{\varepsilon}) \le F(\boldsymbol{x}, \boldsymbol{y} + \boldsymbol{\varepsilon}).$$

Letting $\varepsilon \to 0+$ and using $(\boldsymbol{x}, \boldsymbol{y}) \in C(F)$, we obtain $\lim_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) = F(\boldsymbol{x}, \boldsymbol{y})$.

□

Theorem 6.28 (Slutsky's theorem)

If $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{d}}{\to}} \boldsymbol{X}$ and $\boldsymbol{Y}_n \underset{n\to\infty}{\overset{\text{d}}{\to}} \boldsymbol{c} \in \mathbb{R}^d$, then $\boldsymbol{X}_n + \boldsymbol{Y}_n \underset{n\to\infty}{\overset{\text{d}}{\to}} \boldsymbol{X} + \boldsymbol{c}$ and $\boldsymbol{X}_n\boldsymbol{Y}_n \underset{n\to\infty}{\overset{\text{d}}{\to}} \boldsymbol{c}\boldsymbol{X}$.

Proof. Apply L. 6.27 and the CMT with the continuous $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} + \boldsymbol{y}$ and $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}\boldsymbol{y} = (x_1y_1, \dots, x_dy_d)$. □

6.6 Counterexamples

Lemma 6.29 (Independent two-point distributions)

If $(X_n)_{n\in\mathbb{N}}$ are independent with $\mathbb{P}(X_n = 0) = 1 - 1/n^\alpha$ and $\mathbb{P}(X_n = n) = 1/n^\alpha$ $\forall\, n \in \mathbb{N}$ and $\alpha > 0$, then

1) $X_n \overset{\text{a.s.}}{\to} 0$ if and only if $\alpha > 1$;
2) $X_n \overset{\text{p}}{\to} 0 \;\forall\, \alpha > 0$; and
3) $X_n \overset{L^p}{\to} 0$ if and only if $\alpha > p$.

Proof.

1) Since $\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(X_n > \varepsilon) = \mathbb{P}(X_n = n) = \frac{1}{n^\alpha}$, we have $\sum_{n=1}^\infty \mathbb{P}(|X_n - 0| > \varepsilon) = \sum_{n=1}^\infty \frac{1}{n^\alpha} < \infty \Leftrightarrow \alpha > 1$. Then $\mathbb{P}(X_n = n \text{ io}) \overset{\text{BC1}}{\underset{\text{BC2}}{=}} \mathbb{1}_{(0,1]}(\alpha)$ and therefore $\mathbb{P}(\lim_{n\to\infty} X_n = 0) = \mathbb{P}(X_n = 0 \text{ abfm}) = 1 - \mathbb{P}(X_n = n \text{ io}) = \mathbb{1}_{\{\alpha>1\}}$ which is 1 iff $\alpha > 1$.
2) $\forall\, \alpha > 0$, $\forall\, \varepsilon > 0$, $\mathbb{P}(|X_n - 0| > \varepsilon) \overset{\text{as in}}{\underset{1)}{=}} \frac{1}{n^\alpha} \underset{n\to\infty}{\to} 0$.
3) $\forall\, \alpha > 0$, $\forall\, p > 0$, $\lim_{n\to\infty} \mathbb{E}(|X_n - 0|^p) = \lim_{n\to\infty} \mathbb{E}(X_n^p) = \lim_{n\to\infty}(n^p \cdot \frac{1}{n^\alpha}) = \lim_{n\to\infty} n^{p-\alpha}$, which is 0 iff $\alpha > p$ $\big(1$ iff $\alpha = p$ and ∞ iff $\alpha \in (0,p)\big)$. $\square$

Example 6.30 (Counterexamples for modes of convergence)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of rvs defined on $(\Omega, \mathcal{F}, \mathbb{P})$.

1) Let $(X_n)_{n\in\mathbb{N}}$ with $\mathbb{P}(X_n = 0) = 1 - 1/n^\alpha$ and $\mathbb{P}(X_n = n) = 1/n^\alpha$, and $\alpha \in (0, 1]$. Then $(X_n)_{n\in\mathbb{N}}$ converges i.p. but not a.s..

   Proof. By L. 6.29 2), $X_n \underset{n\to\infty}{\overset{\text{p}}{\to}} 0$ but, by L. 6.29 1), $X_n \underset{n\to\infty}{\overset{\text{a.s.}}{\not\to}} 0$. Note that we cannot have a.s. convergence to any other limit, say Y, either, since then $X_n \underset{n\to\infty}{\overset{\text{p}}{\to}} Y \neq 0$ ↯ □

2) Let $(X_n)_{n\in\mathbb{N}}$ with $\mathbb{P}(X_n = 0) = 1 - 1/n^\alpha$ and $\mathbb{P}(X_n = n) = 1/n^\alpha$, and $\alpha \in (0, p)$. Then $(X_n)_{n\in\mathbb{N}}$ converges i.p. but not in L^p, $p \in [1, \infty)$.

   Proof. By L. 6.29 2), $X_n \underset{n\to\infty}{\overset{\text{p}}{\to}} 0$ but, by L. 6.29 3), $X_n \underset{n\to\infty}{\overset{L^p}{\not\to}} 0$. As before, $(X_n)_{n\in\mathbb{N}}$ can also not converge to any other limit in L^p either. □

3) Let $(X_n)_{n\in\mathbb{N}}$ with $\mathbb{P}(X_n = 0) = 1 - 1/n^\alpha$ and $\mathbb{P}(X_n = n) = 1/n^\alpha$, and $\alpha \in (1, p]$ for some $p > 1$. Then $(X_n)_{n\in\mathbb{N}}$ converges a.s. but not in L^p, $p \in [1, \infty)$.

   Proof. By L. 6.29 1), $X_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} 0$ but, by L. 6.29 3), $X_n \underset{n\to\infty}{\overset{L^p}{\not\to}} 0$. As before, $(X_n)_{n\in\mathbb{N}}$ can also not converge to any other limit in L^p either. □

4) Consider the *typewriter sequence* $X_{\frac{n(n-1)}{2}+k}=\mathbb{1}_{[\frac{k-1}{n},\frac{k}{n}]}$, $n\in\mathbb{N}$, $k=1,\dots,n$, (or: $X_n=\mathbb{1}_{[\frac{n-2^k}{2^k},\frac{n-2^k+1}{2^k}]}$, $k\in\mathbb{N}_0$, $n\in[2^k,2^{k+1})$) on $(\Omega,\mathcal{F},\mathbb{P})=([0,1],\bar{\mathcal{B}}([0,1]),\lambda)$, so $X_1=\mathbb{1}_{[0,1]}$, $X_2=\mathbb{1}_{[0,\frac{1}{2}]}$, $X_3=\mathbb{1}_{[\frac{1}{2},1]}$, $X_4=\mathbb{1}_{[0,\frac{1}{3}]}$, $X_5=\mathbb{1}_{[\frac{1}{3},\frac{2}{3}]}$, $X_6=\mathbb{1}_{[\frac{2}{3},1]}$, etc. Then $(X_n)_{n\in\mathbb{N}}$ converges in L^p, $p\in[1,\infty)$, but not a.s..

Proof. $\forall\, k\in\{1,\dots,n\}$, $\mathbb{E}(|X_{\frac{n(n-1)}{2}+k}-0|^p)=\mathbb{E}(\mathbb{1}_{[\frac{k-1}{n},\frac{k}{n}]})=1\cdot\lambda([\frac{k-1}{n},\frac{k}{n}])=\frac{1}{n}\underset{n\to\infty}{\to}0$ $\forall\, p\in[1,\infty)$ so that $X_m\underset{m\to\infty}{\overset{L^p}{\to}}0$, but $\forall\,\omega\in[0,1]$ and all $n\in\mathbb{N}$, $\exists!\, k\in\{1,\dots,n\}:X_{\frac{n(n-1)}{2}+k}(\omega)=1$ (so $\limsup_{m\to\infty}X_m(\omega)=1$) and $X_{\frac{n(n-1)}{2}+k}(\omega)=0$ $\forall\, k$ except one (so $\liminf_{m\to\infty}X_m(\omega)=0$), hence $(X_m)_{m\in\mathbb{N}}$ does not converge in any $\omega\in\Omega$, so also not a.s.. □

5) Let X be *Rademacher* ($\mathbb{P}(X=-1)=\mathbb{P}(X=1)=1/2$) and $X_n:=(-1)^nX$, $n\in\mathbb{N}$. Then $(X_n)_{n\in\mathbb{N}}$ converges in distribution, but not i.p.

Proof. Clearly, $X\overset{\mathrm{d}}{=}X_1\overset{\mathrm{d}}{=}X_2\overset{\mathrm{d}}{=}\dots$, so $X_n\underset{n\to\infty}{\overset{\mathrm{d}}{\to}}X$. However, $\forall\,\varepsilon\in(0,2)$,

$$\mathbb{P}(|X_n-X|>\varepsilon)=\mathbb{P}(|((-1)^n-1)X|>\varepsilon)=\begin{cases}\mathbb{P}(0>\varepsilon)=0, & n\in 2\mathbb{N},\\ \mathbb{P}(2>\varepsilon)=1, & n\in 2\mathbb{N}-1,\end{cases}$$

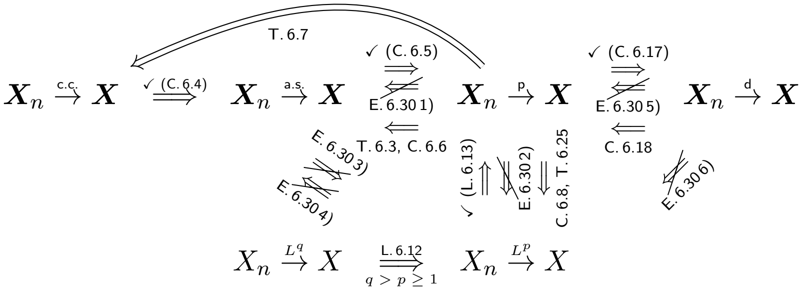
so that $\lim_{n\to\infty}\mathbb{P}(|X_n-X|>\varepsilon)$ does not exist. □

6) Let $(X_n)_{n\in\mathbb{N}}$ be as in 5). Then $(X_n)_{n\in\mathbb{N}}$ converges in distribution, but not in L^p, $p \in [1,\infty)$.

Proof. We showed in 5) that $X_n \underset{n\to\infty}{\overset{\text{d}}{\to}} X$, but $X_n \underset{n\to\infty}{\overset{\text{p}}{\not\to}} X$. If $X_n \underset{n\to\infty}{\overset{L^p}{\to}} X$, then, by L. 6.13, $X_n \underset{n\to\infty}{\overset{\text{p}}{\to}} X$ ↯. □

6.7 Relationships between modes of convergence

We have shown:



6.8 Convergence of quantile functions

Proposition 6.31 (Convergence of quantile functions)
If $F_n(x) \underset{n\to\infty}{\to} F(x)$ $\forall\, x \in C(F)$, then $F_n^{-1}(u) \underset{n\to\infty}{\to} F^{-1}(u)$ $\forall\, u \in (0,1) \cap C(F^{-1})$.

Proof. We show $F^{-1}(u) \underset{\text{i)}}{\leq} \liminf_{n\to\infty} F_n^{-1}(u) \leq \limsup_{n\to\infty} F_n^{-1}(u) \underset{\text{ii)}}{\leq} F^{-1}(u)$ $\forall\, u \in (0,1) \cap C(F^{-1})$. So let $u \in (0,1) \cap C(F^{-1})$.

i) $C(F)^c$ at most countable $\Rightarrow \forall\, \varepsilon > 0,\ \exists\, x \in C(F) : F^{-1}(u) - \varepsilon < x < F^{-1}(u)$ and $\lim_{n\to\infty} F_n(x) \underset{x\,\in\, C(F)}{=} F(x) \overset{\text{def. } F^{-1}(u) \text{ smallest}}{\underset{x\,:\,F(x)\,\geq\, u}{<}} u \Rightarrow F_n(x) < u$ $\forall\, n$ sufficiently large $\underset{\text{def. } F^{-1}(u)}{\Rightarrow} F_n^{-1}(u) \geq x > F^{-1}(u) - \varepsilon$ $\forall\, n$ sufficiently large $\Rightarrow$ $\liminf_{n\to\infty} F_n^{-1}(u) \geq x > F^{-1}(u) - \varepsilon \underset{\varepsilon\,\to\, 0+}{\Rightarrow} \liminf_{n\to\infty} F_n^{-1}(u) \geq F^{-1}(u)$.

ii) $C(F)^c$ at most countable $\Rightarrow \forall\, \varepsilon > 0,\ \forall\, u' > u,\ \exists\, x' \in C(F) : F^{-1}(u') < x' < F^{-1}(u') + \varepsilon$ and $\lim_{n\to\infty} F_n(x') \underset{x'\,\in\, C(F)}{=} F(x') \underset{F^{-1}(u')\,<\,x'}{\geq} u' \underset{\text{ass.}}{>} u \Rightarrow$ $F_n(x') > u$ $\forall\, n$ sufficiently large $\underset{\text{def. } F_n^{-1}}{\Rightarrow} F_n^{-1}(u) \leq x' < F^{-1}(u') + \varepsilon$ $\forall\, n$ sufficiently large $\Rightarrow \limsup_{n\to\infty} F_n^{-1}(u) \leq x' < F^{-1}(u') + \varepsilon \underset{\varepsilon\,\to\, 0+}{\Rightarrow} \limsup_{n\to\infty} F_n^{-1}(u) \leq F^{-1}(u') \underset{u'\,\to\, u+\,\in\, C(F^{-1})}{\Rightarrow} \limsup_{n\to\infty} F_n^{-1}(u) \leq F^{-1}(u)$. $\square$

6.9 Strong law of large numbers

Theorem 6.32 (Strong law of large numbers (SLLN))

If $(X_n)_{n\in\mathbb{N}} \subseteq L^1$ is iid with mean $\mu = \mathbb{E}(X_1)$, then $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} \mu$.

Proof. Not covered. Instead, we give the proof of the weak law of large numbers (WLLN) under the assumption $(X_n)_{n\in\mathbb{N}} \subseteq L^2$ with $\sigma^2 = \operatorname{var}(X_1)$ and with convergence i.p.: $\frac{1}{n}\sum_{i=1}^n X_i \underset{n\to\infty}{\overset{\text{p}}{\to}} \mu$. $\forall\,\varepsilon > 0$, we have

$$0 \le \mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \underset{\text{mon.}}{\le} \mathbb{P}(|\bar{X}_n - \mu| \ge \varepsilon) \underset{\text{Cheby.}}{\le} \frac{\mathbb{E}((\bar{X}_n - \mu)^2)}{\varepsilon^2} = \frac{\operatorname{var}(\bar{X}_n)}{\varepsilon^2}$$

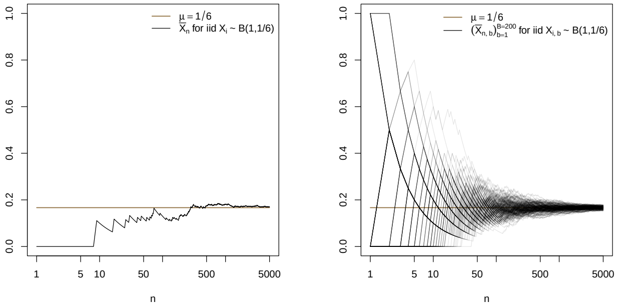
$$\underset{\text{C. 5.33 2)}}{\overset{\text{iid}}{=}} \frac{\sigma^2}{n\varepsilon^2} \underset{n\to\infty}{\to} 0.$$ □

- We already knew from E. 4.12 2) that $\mathbb{P}(\lim_{n\to\infty} \bar{X}_n = c \in \mathbb{R}) \underset{\text{K. 0-1}}{\in} \{0,1\}$.
- **Statistics:**
  - If $X_1, X_2, \ldots \overset{\text{ind.}}{\sim} F$ with $\mathbb{E}(|X_1|) < \infty$, the *sample mean* $\bar{X}_n$ converges a.s. to the (true) mean $\mu = \mathbb{E}(X_1)$ of (the underlying) F. Under these assumptions, we can thus 'see' μ even though we don't know F.

- Although for realizations, $\bar{X}_n$ is a real number, statistical theory is concerned with properties of $\bar{X}_n$ as an approximation to μ and thus studies $\bar{X}_n$ as a rv.
- *Monte Carlo simulation* for approximating means: $\mu = \mathbb{E}(h(\boldsymbol{X})) \underset{n \text{ large}}{\overset{\text{SLLN}}{\approx}} \overline{h(\boldsymbol{X})}_n$.
- For iid $(\boldsymbol{X}_n)_{n \in \mathbb{N}}$, the mean of $\mathbb{1}_{\{\boldsymbol{X}_i \le \boldsymbol{x}\}}$ always exists for fixed $\boldsymbol{x} \in \mathbb{R}^d$, so by the SLLN, we have pointwise a.s. convergence $F_n(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{X}_i \le \boldsymbol{x}\}} \underset{n \to \infty}{\overset{\text{a.s.}}{\to}} \mathbb{E}(\mathbb{1}_{\{\boldsymbol{X}_1 \le \boldsymbol{x}\}}) = \mathbb{P}(\boldsymbol{X}_1 \le \boldsymbol{x}) = F(\boldsymbol{x}) \ \forall \boldsymbol{x} \in \mathbb{R}^d$ of the edf F_n.

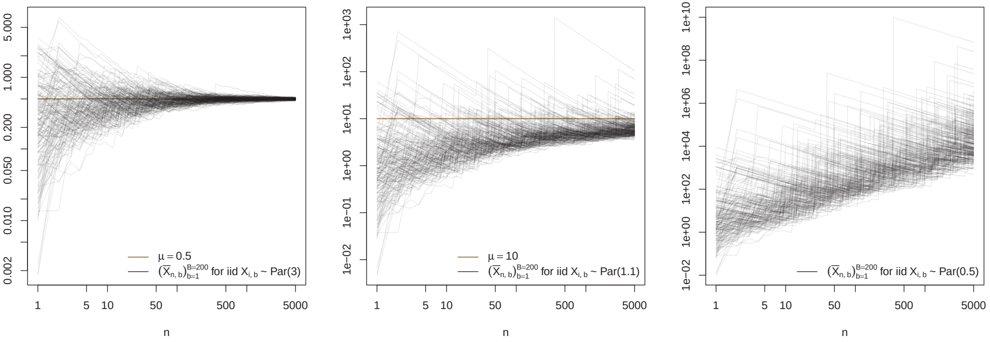
Example 6.33 (SLLN for the probability of rolling a 6)

Let $X_1, \dots, X_n \overset{\text{ind.}}{\sim} \mathrm{B}(1, 1/6)$. By the SLLN, $\bar{X}_n \underset{n \to \infty}{\overset{\text{a.s.}}{\to}} \mu = \mathbb{E}(X_1) = \mathbb{P}(X_1 = 1) = 1/6$. The lhs (rhs) shows 1 (200) simulated paths $n \mapsto \bar{X}_n$, all converging to μ.



Example 6.34 (SLLN for Pareto rvs)

Let $X, X_1, \dots, X_n \overset{\text{ind.}}{\sim} \mathrm{Par}(\theta)$, $\theta > 0$, with $F(x) = 1 - (1+x)^{-\theta}$, $x \geq 0$, mean $\mathbb{E}(X) = \frac{1}{\theta-1}$, $\theta > 1$, and ∞ otherwise. The three figures show 200 simulated paths $n \mapsto \bar{X}_n$ for $\theta = 3$ ($\mathbb{E}(X) = \frac{1}{2}$), $\theta = 1.1$ ($\mathbb{E}(X) = 10$) and $\theta = 0.5$ ($\mathbb{E}(X) = \infty$), from which we see the necessary requirement of an existing mean.



The pointwise convergence $F_n(\boldsymbol{x}) \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, of edfs is even uniform.

Theorem 6.35 (Glivenko–Cantelli)
If $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \overset{\text{ind.}}{\sim} F$, then $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |F_n(\boldsymbol{x}) - F(\boldsymbol{x})| \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} 0$.

Proof. We consider $d = 1$, the proof for $d \geq 2$ is given in Kiefer and Wolfowitz (1958). We show that the supremum is bounded above by the maximal distance at finitely-many points and apply the SLLN finitely-many times.

1) For $x \in [a, b)$, monotonicity of the dfs F_n and F implies that

$$F_n(x) - F(x) \geq F_n(a) - F(b-) = (F_n(a) - F(a)) - (F(b-) - F(a))$$
$$F_n(x) - F(x) \leq F_n(b-) - F(a) = (F_n(b-) - F(b-)) + (F(b-) - F(a))$$

$$\underset{|z|\leq\max\{|c|,|d|\}}{\overset{z\in[c,d]\text{ implies}}{\Longrightarrow}} |F_n(x) - F(x)| \underset{\Delta}{\leq} \max\{|F_n(a) - F(a)|,\ |F_n(b-) - F(b-)|\} + (F(b-) - F(a)).$$

2) $\forall\, \varepsilon \in (0, 3]$, $\exists\, n_\varepsilon \in \mathbb{N}$ and a partition $-\infty =: z_0 < z_1 < \dots < z_{n_\varepsilon} := \infty$ such that $F(z_k-) - F(z_{k-1}) \leq \frac{\varepsilon}{3}$ for all $k = 1, \dots, n_\varepsilon$; if F is continuous we can take $z_k = F^{-1}(\frac{\varepsilon}{3}k)$, $k = 0, 1, 2, \dots, \lfloor 3/\varepsilon \rfloor$, $n_\varepsilon := \lfloor 3/\varepsilon \rfloor + 1$, and otherwise, F

can jump at most $\lceil 1/y \rceil$ times $> y$, so include those $x : F(x) - F(x-) > \frac{\varepsilon}{3}$ into the partition.

3) Since each $x \in \mathbb{R}$ lies in precisely one partition element, we apply 1) with $a = z_{k-1}$ and $b = z_k$ for $k = 1, \dots, n_\varepsilon$ to get

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

$$\underset{1)}{\leq} \max_{1\leq k\leq n_\varepsilon} \{\max\{|F_n(z_{k-1}) - F(z_{k-1})|, \; |F_n(z_k-) - F(z_k-)|\}\} + \frac{\varepsilon}{3}$$

$$\underset{\max_k\{c_k\}+\max_k\{d_k\}}{\overset{\max_k\{\max\{c_k,d_k\}\}\leq}{\leq}} \max_{1\leq k\leq n_\varepsilon} \{|\underbrace{F_n(z_{k-1}) - F(z_{k-1})}_{\underset{n\to\infty}{\overset{\text{a.s.}}{\to}} 0 \text{ (SLLN)}}|\} + \max_{1\leq k\leq n_\varepsilon} \{|\underbrace{F_n(z_k-) - F(z_k-)}_{\underset{n\to\infty}{\overset{\text{a.s.}}{\to}} 0 \text{ (SLLN)}}|\} + \frac{\varepsilon}{3}$$

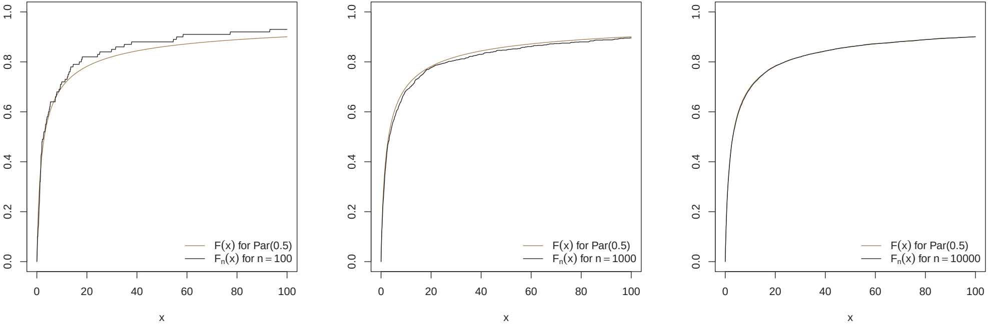
$$\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \quad \text{a.s. for } n \text{ sufficiently large.}$$ □

The rate of convergence of $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ is quantified by the *Dvoretzky–Kiefer–Wolfowitz (DKW) inequality*

$$\mathbb{P}(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}, \quad \varepsilon > 0,$$

for $d = 1$ and $\mathbb{P}(\sup_{\boldsymbol{x}\in\mathbb{R}^d} |F_n(\boldsymbol{x}) - F(\boldsymbol{x})| > \varepsilon) \leq (n+1)de^{-2n\varepsilon^2}$ for $d \geq 2$.

Example 6.36 (Glivenko–Cantelli theorem for $\mathrm{Par}(1/2)$ rvs)
Let $X_1, \dots, X_n \overset{\text{ind.}}{\sim} \mathrm{Par}(1/2)$ (no mean). The three figures show $F(x) = 1 - (1 + x)^{-1/2}$, $x \geq 0$, and F_n for $n \in \{100, 1000, 10\,000\}$.



References

Kiefer, J. and Wolfowitz, J. (1958), On the Deviations of the Empirical Distribution Function of Vector Chance Variables, *Transactions of the American Mathematical Society*, 87(1), 173–186.

Van der Vaart, A. W. (2000), Asymptotic Statistics, Cambridge University Press.

7 Characteristic functions

7.1 Basics

7.2 The central limit theorem

7.1 Basics

- **Motivation:** For $X_i \sim F_i$, $i = 1, \dots, n$, we are often interested in

$$S_n := \sum_{i=1}^n X_i \quad \text{and} \quad \bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i.$$

Example: Total/average loss of n observations.

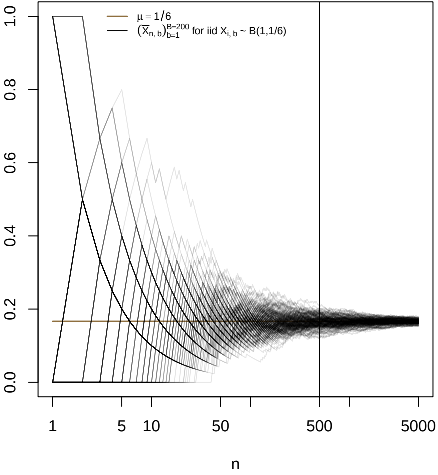
- **Problem:** Even if $X_1, \dots, X_n$ are independent, we only get a *convolution* formula for $\mathbb{P}(S_n \le x)$:

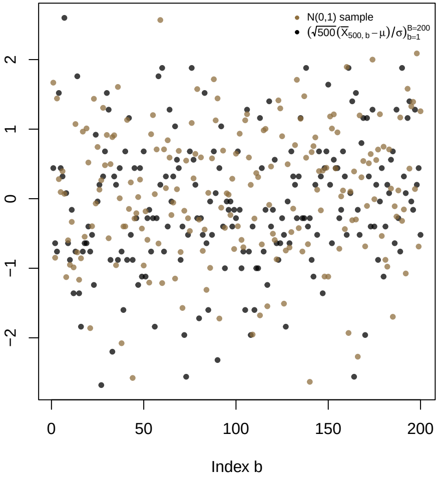
$$\begin{aligned}
F_{S_n}(x) &= \mathbb{P}(X_1 + \dots + X_n \le x) = \mathbb{E}(\mathbb{1}_{\{X_1+\dots+X_n \le x\}}) \\
&\underset{\text{variables}}{\overset{\text{change of}}{=}} \int_{\mathbb{R}^n} \mathbb{1}_{\{x_1+\dots+x_n \le x\}} \, \mathrm{d}F(x_1, \dots, x_n) \\
&\underset{\text{Tonelli}}{=} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{1}_{\{x_1 \le x - x_2 - \dots - x_n\}} \, \mathrm{d}F_1(x_1) \mathrm{d}F_2(x_2) \dots \mathrm{d}F_n(x_n) \\
&= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \mathbb{E}(\mathbb{1}_{\{X_1 \le x - x_2 - \dots - x_n\}}) \, \mathrm{d}F_2(x_2) \dots \mathrm{d}F_n(x_n) \\
&= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} F_1(x - x_2 - \dots - x_n) \, \mathrm{d}F_2(x_2) \dots \mathrm{d}F_n(x_n).
\end{aligned}$$

- **Question:** How can we approximate F_{S_n} or $F_{\bar{X}_n}$?

Example 7.1 (Normality appearing when rolling a 6)

The lhs figure shows the rhs figure of E. 6.33 with $n = 500$ indicated. For this n, we have $B = 200$ realizations $(\bar{X}_{n,b})_{b=1}^{B}$ of $\bar{X}_n$. Standardizing them with $\mu = 1/6$, $\sigma = 5/36$ leads to $(\sqrt{n}\frac{\bar{X}_{n,b}-\mu}{\sigma})_{b=1}^{B}$, plotted on the rhs together with a $\mathrm{N}(0,1)$ sample of size B. We see a rough similarity of the two samples ($\Rightarrow$ idea).





- **Idea:** Use a transform that turns sums of independent rvs into products, then use $\mathbb{E}(\prod_{j=1}^n X_j) \underset{\text{ind.}}{=} \prod_{j=1}^n \mathbb{E}(X_j) \underset{\text{id}}{=} \mathbb{E}(X_1)^n$. One is the moment-generating function $M(\boldsymbol{t}) = \mathbb{E}(e^{\boldsymbol{t}^\top \boldsymbol{X}})$ for $\boldsymbol{t}$ in some open neighbourhood of $\boldsymbol{0}$.
- **Problem:** Does not always exist (e.g. not for Cauchy distributed X).

Definition 7.2 (Characteristic function)

The *characteristic function (cf)* $\phi_{\boldsymbol{X}} : \mathbb{R}^d \to \mathbb{C}$ of $\boldsymbol{X}$ is $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}(e^{i\boldsymbol{t}^\top \boldsymbol{X}})$, $\boldsymbol{t} \in \mathbb{R}^d$.

Lemma 7.3 (Properties of cfs)

1) $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}(\cos(\boldsymbol{t}^\top \boldsymbol{X})) + i\mathbb{E}(\sin(\boldsymbol{t}^\top \boldsymbol{X})) = \mathrm{Re}(\phi_{\boldsymbol{X}}(\boldsymbol{t})) + i\,\mathrm{Im}(\phi_{\boldsymbol{X}}(\boldsymbol{t}))$, $\boldsymbol{t} \in \mathbb{R}^d$.
2) $\mathbb{E}(|e^{i\boldsymbol{t}^\top \boldsymbol{X}}|) = 1$, $\boldsymbol{t} \in \mathbb{R}^d$ (integrability).
3) Cfs exists $\forall\, \boldsymbol{t} \in \mathbb{R}^d$.
4) $|\phi_{\boldsymbol{X}}(\boldsymbol{t})| \le 1$, $\boldsymbol{t} \in \mathbb{R}^d$, and $\phi_{\boldsymbol{X}}(\boldsymbol{0}) = 1$.
5) $\phi_{\boldsymbol{X}}$ is uniformly continuous.
6) $\phi_{A\boldsymbol{X}+\boldsymbol{b}}(\boldsymbol{t}) = e^{i\boldsymbol{t}^\top \boldsymbol{b}} \phi_{\boldsymbol{X}}(A^\top \boldsymbol{t})$, $\boldsymbol{t} \in \mathbb{R}^d$.
7) If $X_1, \dots, X_d$ are independent, then $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \prod_{j=1}^d \phi_{X_j}(t_j)$, $\boldsymbol{t} \in \mathbb{R}^d$. In particular, $\phi_{\sum_{j=1}^d X_j}(t) = \prod_{j=1}^d \phi_{X_j}(t)$, $t \in \mathbb{R}$.

Proof.

1) *Euler's formula* $e^{ix} = \cos(x) + i\sin(x)$, $x \in \mathbb{R}$, and linearity imply $\phi_{\boldsymbol{X}}(\boldsymbol{t}) \underset{\text{def.}}{=} \mathbb{E}(\cos(\boldsymbol{t}^\top \boldsymbol{X})) + i\mathbb{E}(\sin(\boldsymbol{t}^\top \boldsymbol{X}))$, $\boldsymbol{t} \in \mathbb{R}^d$.
2) Since $|x + iy| = \sqrt{x^2 + y^2}$,

$$|e^{i\boldsymbol{t}^\top \boldsymbol{X}}| = |\cos(\boldsymbol{t}^\top \boldsymbol{X}) + i\sin(\boldsymbol{t}^\top \boldsymbol{X})| = \sqrt{\cos^2(\boldsymbol{t}^\top \boldsymbol{X}) + \sin^2(\boldsymbol{t}^\top \boldsymbol{X})} = 1 \quad (5)$$

and thus $\mathbb{E}(|e^{i\boldsymbol{t}^\top \boldsymbol{X}}|) = 1$, $\boldsymbol{t} \in \mathbb{R}^d$.
3) By 2) and Remark 5.20, $\phi_{\boldsymbol{X}}(\boldsymbol{t})$ exists $\forall\, \boldsymbol{t} \in \mathbb{R}^d$.
4) $|\phi_{\boldsymbol{X}}(\boldsymbol{t})| \underset{\Delta}{\leq} \mathbb{E}(|e^{i\boldsymbol{t}^\top \boldsymbol{X}}|) \underset{2)}{=} 1$, $\boldsymbol{t} \in \mathbb{R}^d$, and $\phi_{\boldsymbol{X}}(\boldsymbol{0}) = \mathbb{E}(e^0) = 1$.
5) $|\phi_{\boldsymbol{X}}(\boldsymbol{t} + \boldsymbol{h}) - \phi_{\boldsymbol{X}}(\boldsymbol{t})| = |\mathbb{E}(e^{i\boldsymbol{t}^\top \boldsymbol{X}}(e^{i\boldsymbol{h}^\top \boldsymbol{X}} - 1))| \underset{\Delta}{\leq} \mathbb{E}(|e^{i\boldsymbol{t}^\top \boldsymbol{X}}(e^{i\boldsymbol{h}^\top \boldsymbol{X}} - 1)|) \underset{(5)}{=} \mathbb{E}(1 \cdot |e^{i\boldsymbol{h}^\top \boldsymbol{X}} - 1|)$. Now $e^{i\boldsymbol{h}^\top \boldsymbol{X}} - 1 \overset{\text{pointwise}}{\underset{\boldsymbol{h} \to \boldsymbol{0}+}{\longrightarrow}} 0$ and $|e^{i\boldsymbol{h}^\top \boldsymbol{X}} - 1| \underset{\Delta}{\leq} |e^{i\boldsymbol{h}^\top \boldsymbol{X}}| + 1 \leq 2 \in L^1 \underset{\text{DOM}}{\Rightarrow} \lim_{\boldsymbol{h} \to \boldsymbol{0}+} \mathbb{E}(|e^{i\boldsymbol{h}^\top \boldsymbol{X}} - 1|) = 0$, so $\phi_{\boldsymbol{X}}$ is uniformly continuous.
6) $\phi_{A\boldsymbol{X}+\boldsymbol{b}}(\boldsymbol{t}) = \mathbb{E}(e^{i\boldsymbol{t}^\top (A\boldsymbol{X}+\boldsymbol{b})}) = e^{i\boldsymbol{t}^\top \boldsymbol{b}}\mathbb{E}(e^{i\boldsymbol{t}^\top (A\boldsymbol{X})}) = e^{i\boldsymbol{t}^\top \boldsymbol{b}}\phi_{\boldsymbol{X}}(A^\top \boldsymbol{t})$, $\boldsymbol{t} \in \mathbb{R}^d$.
7) $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}(e^{i\boldsymbol{t}^\top \boldsymbol{X}}) = \mathbb{E}(e^{i\sum_{j=1}^d t_j X_j}) = \mathbb{E}(\prod_{j=1}^d e^{it_j X_j}) \underset{\text{P. 5.24}}{\overset{\text{ind.}}{=}} \prod_{j=1}^d \mathbb{E}(e^{it_j X_j}) = \prod_{j=1}^d \phi_{X_j}(t_j)$. The second statement follows from $\phi_{\boldsymbol{X}}(\boldsymbol{t}) \underset{\boldsymbol{t} = t\boldsymbol{1}}{=} \phi_{\sum_{j=1}^d X_j}(t)$. □

Example 7.4 (Cf of normal distribution)

1) For $Z \sim \mathrm{N}(0,1)$, point-symmetry of $\sin$ about 0 implies that $\mathbb{E}(\sin(tZ)) = 0$ and thus $\phi_Z(t) = \mathbb{E}(\cos(tZ)) = \int_{\mathbb{R}} \cos(tz)\varphi(z)\,\mathrm{d}z$. By the Leibniz integral rule (general: if $f(t,z)$ and $\frac{\partial}{\partial t} f(t,z)$ are continuous in z and a, b are continuously differentiable, then $\frac{\mathrm{d}}{\mathrm{d}t} \int_{a(t)}^{b(t)} f(t,z)\,\mathrm{d}z = f(t,b(t))b'(t) - f(t,a(t))a'(t) + \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} f(t,z)\,\mathrm{d}z$), so that

$$\phi'_Z(t) \underset{\text{Leibniz}}{=} \int_{\mathbb{R}} (-z)\sin(tz)\varphi(z)\,\mathrm{d}z \underset{-z\varphi(z) = \varphi'(z)}{=} \int_{\mathbb{R}} \sin(tz)\varphi'(z)\,\mathrm{d}z$$

$$\underset{\text{by parts}}{=} [\sin(tz)\varphi(z)]_{-\infty}^{\infty} - t\int_{\mathbb{R}} \cos(tz)\varphi(z)\,\mathrm{d}z = 0 - 0 - t\phi_Z(t).$$

$\Rightarrow \phi'_Z(t) = -t\phi_Z(t) \Rightarrow \big(\log(\phi_Z(t))\big)' = \frac{\phi'_Z(t)}{\phi_Z(t)} = -t \Rightarrow \log(\phi_Z(t)) = -\frac{t^2}{2} + c$, $c \in \mathbb{R} \Rightarrow \phi_Z(t) = e^{-\frac{t^2}{2}+c} \underset{\phi_Z(0) \overset{!}{=} 1}{\Rightarrow} \phi_Z(t) = e^{-\frac{t^2}{2}}$, $t \in \mathbb{R}$.

2) If $\boldsymbol{X} = \boldsymbol{\mu} + A\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$ for $\Sigma = AA^\top$, then

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = e^{i\boldsymbol{t}^\top\boldsymbol{\mu}}\mathbb{E}(e^{i\boldsymbol{t}^\top A\boldsymbol{Z}}) \underset{\tilde{\boldsymbol{t}} = A^\top \boldsymbol{t}}{=} e^{i\boldsymbol{t}^\top\boldsymbol{\mu}}\phi_{\boldsymbol{Z}}(\tilde{\boldsymbol{t}}) \underset{\text{ind.}}{=} e^{i\boldsymbol{t}^\top\boldsymbol{\mu}} \prod_{j=1}^{d} \phi_{Z_j}(\tilde{t}_j)$$

$$\underset{1)}{=} e^{i\boldsymbol{t}^\top\boldsymbol{\mu}} \prod_{j=1}^{d} e^{-\frac{1}{2}\tilde{t}_j^2} = e^{i\boldsymbol{t}^\top\boldsymbol{\mu}-\frac{1}{2}\sum_{j=1}^d \tilde{t}_j^2} = e^{i\boldsymbol{t}^\top\boldsymbol{\mu}-\frac{1}{2}\tilde{\boldsymbol{t}}^\top\tilde{\boldsymbol{t}}}$$

$$\underset{\tilde{\boldsymbol{t}}=A^\top\boldsymbol{t}}{=} e^{i\boldsymbol{t}^\top\boldsymbol{\mu}-\frac{1}{2}(\boldsymbol{t}^\top A)A^\top\boldsymbol{t}} = e^{i\boldsymbol{t}^\top\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t}}, \quad \boldsymbol{t}\in\mathbb{R}^d.$$

For $d=1$, we obtain the cf $\phi_X(t) = e^{it\mu-\frac{1}{2}\sigma^2t^2}$, $t\in\mathbb{R}$, of $\mathrm{N}(\mu,\sigma^2)$.

Theorem 7.5 (Lévy's continuity theorem)

1) $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{X} \Rightarrow \phi_{\boldsymbol{X}_n}(\boldsymbol{t}) \underset{n\to\infty}{\to} \phi_{\boldsymbol{X}}(\boldsymbol{t})\ \forall\,\boldsymbol{t}\in\mathbb{R}^d$.
2) If $\phi(\boldsymbol{t}) := \lim_{n\to\infty}\phi_{\boldsymbol{X}_n}(\boldsymbol{t})$ exists $\forall\,\boldsymbol{t}\in\mathbb{R}^d$ and is continuous at $\boldsymbol{0}$, then $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{X}$ for a random vector $\boldsymbol{X}$ with cf ϕ.

Proof.

1) $\forall\,\boldsymbol{t}\in\mathbb{R}^d$, $h(\boldsymbol{x}) = e^{i\boldsymbol{t}^\top\boldsymbol{x}} \underset{\text{Euler}}{=} \cos(\boldsymbol{t}^\top\boldsymbol{x}) + i\sin(\boldsymbol{t}^\top\boldsymbol{x})$ is bounded (by 1) and continuous (composition) $\underset{\text{portmanteau}}{\Rightarrow} \phi_{\boldsymbol{X}_n}(\boldsymbol{t}) = \mathbb{E}(e^{i\boldsymbol{t}^\top\boldsymbol{X}_n}) \underset{n\to\infty}{\to} \mathbb{E}(e^{i\boldsymbol{t}^\top\boldsymbol{X}}) = \phi_{\boldsymbol{X}}(\boldsymbol{t})$.
2) Needs more work. □

Theorem 7.6 (Cramér–Wold device)

If $\boldsymbol{X}, \boldsymbol{X}_n$, $n \in \mathbb{N}$, are random vectors, $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{X}$ iff $\boldsymbol{a}^\top \boldsymbol{X}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{a}^\top \boldsymbol{X}\ \forall\, \boldsymbol{a} \in \mathbb{R}^d$.

Proof.

"$\Rightarrow$": Apply the CMT with $h(\boldsymbol{x}) := \boldsymbol{a}^\top \boldsymbol{x}$.

"$\Leftarrow$": $\phi_{\boldsymbol{X}_n}(\boldsymbol{t}) = \mathbb{E}(e^{i\boldsymbol{t}^\top \boldsymbol{X}_n}) = \phi_{\boldsymbol{t}^\top \boldsymbol{X}_n}(1) \underset{\text{T. 7.5 1)}}{\overset{\text{ass.}}{\to}} \phi_{\boldsymbol{t}^\top \boldsymbol{X}}(1) = \phi_{\boldsymbol{X}}(\boldsymbol{t})\ \forall\, \boldsymbol{t} \in \mathbb{R}^d$. By L. 7.3 5), $\phi_{\boldsymbol{X}}$ is continuous at $\boldsymbol{0}$ (even uniformly) $\underset{\text{T. 7.5 2)}}{\Rightarrow} \boldsymbol{X}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \boldsymbol{X}$ □

Corollary 7.7

Let $\boldsymbol{X}, \boldsymbol{Y}$ be two random vectors. Then $\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{Y}$ iff $\boldsymbol{a}^\top \boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{a}^\top \boldsymbol{Y}\ \forall\, \boldsymbol{a} \in \mathbb{R}^d$.

Lemma 7.8 (Analytic auxiliary results)

1) $\forall\, m \in \mathbb{N}_0$, $\left|e^{ix} - \sum_{k=0}^m \frac{(ix)^k}{k!}\right| \le \min\left\{\frac{|x|^{m+1}}{(m+1)!}, \frac{2|x|^m}{m!}\right\}$, $x \in \mathbb{R}$.

2) $\forall\, \alpha \in \mathbb{R}$, let $I_\alpha(T) := \int_0^T \frac{\sin(\alpha t)}{t}\,\mathrm{d}t$, $T > 0$, and $I_\alpha(0) := 0$. Then $I_{-\alpha}(T) = -I_\alpha(T)$ and $I_\alpha(T) = I_1(\alpha T)$. Furthermore, $I_\alpha(T)$, $T \ge 0$, is continuous. And

$$\lim_{T\to\infty} I_\alpha(T) \underset{T\to\infty}{\to} \begin{cases} -\frac{\pi}{2}, & \alpha < 0, \\ 0, & \alpha = 0, \\ \frac{\pi}{2}, & \alpha > 0. \end{cases}$$

3) $\forall\, T > 0$ and $\alpha \in \mathbb{R}$, we have $|I_\alpha(T)| \le \sup_{\tilde{T}\ge 0} |I_1(\tilde{T})| < \infty$.

Proof.

1) a) Taylor's theorem with integral remainder states that $f(x)=\sum_{k=0}^m \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k+\int_{x_0}^x \frac{f^{(m+1)}(t)}{m!}(x-t)^m\,\mathrm{d}t$. Applied to $f(x)=e^{ix}$ about $x_0=0$, we have $e^{ix}=\sum_{k=0}^m \frac{i^k}{k!}x^k+\frac{i^{m+1}}{m!}\int_0^x e^{it}(x-t)^m\,\mathrm{d}t$.

b) Therefore, $\forall\, x>0$, $\left|e^{ix}-\sum_{k=0}^m \frac{(ix)^k}{k!}\right| = \left|\frac{i^{m+1}}{m!}\int_0^x e^{it}(x-t)^m\,\mathrm{d}t\right| \underset{x>0}{\overset{\Delta,\,|e^{it}|=1}{\leq}} \frac{1}{m!}\int_0^x (x-t)^m\,\mathrm{d}t=\frac{x^{m+1}}{(m+1)!}$.

c) $e^{ix}\underset{m\leftarrow m-1}{=}\sum_{k=0}^{m-1}\frac{(ix)^k}{k!}+\frac{i^m}{(m-1)!}\int_0^x e^{it}(x-t)^{m-1}\,\mathrm{d}t \underset{\pm\frac{(ix)^m}{m!}}{=} \sum_{k=0}^m \frac{(ix)^k}{k!}+\frac{i^m}{(m-1)!}\cdot\left(\int_0^x e^{it}(x-t)^{m-1}\,\mathrm{d}t-\frac{x^m}{m}\right) \underset{\frac{x^m}{m}=\int_0^x (x-t)^{m-1}\mathrm{d}t}{=} \sum_{k=0}^m \frac{(ix)^k}{k!}+\frac{i^m}{(m-1)!}\int_0^x (e^{it}-1)(x-t)^{m-1}\,\mathrm{d}t$,

so $\left|e^{ix}-\sum_{k=0}^m \frac{(ix)^k}{k!}\right| \underset{\Delta}{\leq} \frac{|i|^m}{(m-1)!}\int_0^x |e^{it}-1|(x-t)^{m-1}\,\mathrm{d}t \leq \frac{2}{(m-1)!}\frac{x^m}{m}=\frac{2x^m}{m!}$.

d) $\underset{\text{b),c)}}{\Rightarrow} \left|e^{ix}-\sum_{k=0}^m \frac{(ix)^k}{k!}\right| \leq \min\left\{\frac{x^{m+1}}{(m+1)!},\frac{2x^m}{m!}\right\}$, $x>0$. Similarly for $x<0$. Pulling the two cases together, we obtain the result as stated.

2) The first two identities trivially hold for $T=0$. And for $T>0$, $I_{-\alpha}(T)=\int_0^T \frac{\sin(-\alpha t)}{t}\,\mathrm{d}t \underset{\text{odd}}{\overset{\sin()}{=}} -\int_0^T \frac{\sin(\alpha t)}{t}\,\mathrm{d}t=-I_\alpha(T)$ and $I_\alpha(T)\underset{x=\alpha t}{=}\int_0^{\alpha T}\frac{\sin(x)}{x}\,\mathrm{d}x=I_1(\alpha T)$. For the remaining parts, see the additional slides. □

Theorem 7.9 (Uniqueness)

$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \phi_{\boldsymbol{Y}}(\boldsymbol{t})\ \forall\,\boldsymbol{t}\in\mathbb{R}^d$ iff $\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{Y}$, so cfs uniquely characterize distributions.

Proof. "$\Leftarrow$" ✓ by definition. Consider "$\Rightarrow$". For $d=1$, we derive an inversion formula; the case $d>1$ follows thereafter.

i) $|e^{ix}-1| \underset{\text{for } m=0}{\overset{\text{L. 7.8 1)}}{\leq}} \min\{|x|,2\} \underset{(*)}{\leq} |x| \Rightarrow \forall\, a<b,\ \left|\frac{e^{-ita}-e^{-itb}}{it}\right| = |e^{-itb}|\cdot\left|\frac{e^{-it(a-b)}-1}{it}\right| = \frac{|e^{-it(a-b)}-1|}{|t|} \underset{(*)}{\leq} \frac{|-t(a-b)|}{|t|} = b-a$, so $\int_{-T}^{T}\left|\frac{e^{-ita}-e^{-itb}}{it}\phi_X(t)\right|\mathrm{d}t \underset{|\phi_X|\leq 1}{\leq} 2T|b-a| < \infty\ \forall\, T>0$, which justifies the following application of Fubini.

ii) We have

$$\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\phi_X(t)\,\mathrm{d}t \underset{\text{Fubini}}{\overset{\text{def. }\phi_X}{=}} \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-T}^{T}\frac{e^{it(x-a)}-e^{it(x-b)}}{it}\,\mathrm{d}t\,\mathrm{d}F(x)$$

$$\underset{\frac{\cos(tc)}{t}\ \text{point-symm.}}{\overset{\text{Euler's formula}}{=}} \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-T}^{T}\frac{\sin(t(x-a))-\sin(t(x-b))}{t}\,\mathrm{d}t\,\mathrm{d}F(x)$$

$$\underset{\frac{\sin(tc)}{t}\ \text{symm.}}{=} \frac{1}{\pi}\int_{-\infty}^{\infty}\int_{0}^{T}\frac{\sin(t(x-a))-\sin(t(x-b))}{t}\,\mathrm{d}t\,\mathrm{d}F(x)$$

$$= \frac{1}{\pi}\int_{-\infty}^{\infty}(I_{x-a}(T)-I_{x-b}(T))\,\mathrm{d}F(x).$$

iii) We have

$$\lim_{T\to\infty}(I_{x-a}(T)-I_{x-b}(T)) \underset{\text{L. 7.8 2)}}{=} \begin{cases} -\frac{\pi}{2}-(-\frac{\pi}{2})=0, & x<a,\\ 0-(-\frac{\pi}{2})=\frac{\pi}{2}, & x=a,\\ \frac{\pi}{2}-(-\frac{\pi}{2})=\pi, & x\in(a,b),\\ \frac{\pi}{2}-0=\frac{\pi}{2}, & x=b,\\ \frac{\pi}{2}-\frac{\pi}{2}=0, & x>b,\end{cases}$$

so $\mathbb{E}(\lim_{T\to\infty}\frac{I_{x-a}(T)-I_{x-b}(T)}{\pi})=\frac{1}{2}\mathbb{P}(X=a)+1\cdot\mathbb{P}(X\in(a,b))+\frac{1}{2}\mathbb{P}(X=b)$.

iv) Therefore

$$\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\phi_X(t)\,\mathrm{d}t \underset{\text{ii)}}{=} \lim_{T\to\infty}\frac{1}{\pi}\int_{-\infty}^{\infty}(I_{x-a}(T)-I_{x-b}(T))\,\mathrm{d}F(x)$$

$$\overset{\text{iii), L. 7.8 3)}}{\underset{\text{DOM}}{=}} \int_{-\infty}^{\infty}\lim_{T\to\infty}\frac{I_{x-a}(T)-I_{x-b}(T)}{\pi}\,\mathrm{d}F(x) \underset{\text{iii)}}{=} F(b-)-F(a)+\frac{\mathbb{P}(X=a)}{2}+\frac{\mathbb{P}(X=b)}{2}.$$

Adding and subtracting $\mathbb{P}(X=b)$, we obtain the inversion formula

$$F(b)-F(a)+\frac{\mathbb{P}(X=a)}{2}-\frac{\mathbb{P}(X=b)}{2}=\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\phi_X(t)\,\mathrm{d}t.$$

For $d>1$, $\phi_{\boldsymbol{a}^\top\boldsymbol{X}}(t)=\mathbb{E}(e^{it\boldsymbol{a}^\top\boldsymbol{X}})=\phi_{\boldsymbol{X}}(t\boldsymbol{a}) \underset{\text{ass.}}{=} \phi_{\boldsymbol{Y}}(t\boldsymbol{a}) \overset{\text{same}}{\underset{\text{backwards}}{=}} \phi_{\boldsymbol{a}^\top\boldsymbol{Y}}(t)\ \forall\, t\in\mathbb{R}$
$\overset{\text{uniqueness}}{\underset{\text{for } d=1}{\Rightarrow}} \boldsymbol{a}^\top\boldsymbol{X}\overset{\text{d}}{=}\boldsymbol{a}^\top\boldsymbol{Y}\ \forall\,\boldsymbol{a}\in\mathbb{R}^d \underset{\text{C. 7.7}}{\Rightarrow} \checkmark$. □

If F is continuous, the inversion formula for general d can be shown to be

$$\Delta_{[\boldsymbol{a},\boldsymbol{b}]}F = \lim_{T\to\infty}\frac{1}{(2\pi)^d}\int_{[-T,T]^d}\biggl(\prod_{j=1}^d \frac{e^{-it_ja_j}-e^{-it_jb_j}}{it_j}\biggr)\phi_{\boldsymbol{X}}(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t};$$

if F has a density, then $f(\boldsymbol{x}) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\boldsymbol{t}^\top\boldsymbol{x}}\phi_{\boldsymbol{X}}(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}$.

Corollary 7.10

$\phi_{\boldsymbol{X}}$ is real iff $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \phi_{\boldsymbol{X}}(-\boldsymbol{t}) = \phi_{-\boldsymbol{X}}(\boldsymbol{t})$, $\boldsymbol{t}\in\mathbb{R}^d$, so iff $\boldsymbol{X} \overset{\mathrm{d}}{=} -\boldsymbol{X}$, i.e. $\boldsymbol{X}$ is radially symmetric about $\boldsymbol{0}$.

Proof. $\phi_{\boldsymbol{X}}$ is real iff

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \overline{\phi_{\boldsymbol{X}}}(\boldsymbol{t}) = \mathbb{E}(\cos(\boldsymbol{t}^\top\boldsymbol{X})) - i\mathbb{E}(\sin(\boldsymbol{t}^\top\boldsymbol{X})) \underset{\text{odd}}{\overset{\sin()}{=}} \phi_{\boldsymbol{X}}(-\boldsymbol{t}) = \mathbb{E}(e^{i(-\boldsymbol{t})^\top\boldsymbol{X}})$$
$$= \mathbb{E}(e^{i\boldsymbol{t}^\top(-\boldsymbol{X})}) = \phi_{-\boldsymbol{X}}(\boldsymbol{t}),\quad \boldsymbol{t}\in\mathbb{R}^d,$$

so, by uniqueness, iff $\boldsymbol{X} \overset{\mathrm{d}}{=} -\boldsymbol{X}$. □

Proposition 7.11 (Characterization of $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$)
$\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$ iff $\boldsymbol{a}^\top \boldsymbol{X} \sim \mathrm{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \Sigma \boldsymbol{a})$ for all $\boldsymbol{a} \in \mathbb{R}^d$.

Proof.

"$\Rightarrow$": $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma) \Rightarrow \phi_{\boldsymbol{a}^\top \boldsymbol{X}}(t) = \mathbb{E}(e^{i(t\boldsymbol{a})^\top \boldsymbol{X}}) = \phi_{\boldsymbol{X}}(t\boldsymbol{a}) \underset{\text{E. 7.4 2)}}{=} e^{i(t\boldsymbol{a})^\top \boldsymbol{\mu} - \frac{1}{2}(t\boldsymbol{a})^\top \Sigma (t\boldsymbol{a})}$ $= e^{it(\boldsymbol{a}^\top \boldsymbol{\mu}) - \frac{1}{2}t^2(\boldsymbol{a}^\top \Sigma \boldsymbol{a})}$, $t \in \mathbb{R} \underset{\text{E. 7.4 2)}}{\Rightarrow} \phi_{\boldsymbol{a}^\top \boldsymbol{X}}$ is the cf of $\mathrm{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \Sigma \boldsymbol{a}) \underset{\text{uniqueness}}{\Rightarrow}$ $\boldsymbol{a}^\top \boldsymbol{X} \sim \mathrm{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \Sigma \boldsymbol{a})$.

"$\Leftarrow$": Let $\boldsymbol{a}^\top \boldsymbol{X} \sim \mathrm{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \Sigma \boldsymbol{a})$. If $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, "$\Rightarrow$" implies that $\boldsymbol{a}^\top \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \Sigma \boldsymbol{a})$, so $\boldsymbol{a}^\top \boldsymbol{Y} \overset{\mathrm{d}}{=} \boldsymbol{a}^\top \boldsymbol{X}$, and that $\forall\, \boldsymbol{a} \in \mathbb{R}^d \underset{\text{C. 7.7}}{\Rightarrow} \boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{Y}$. Since $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, we obtain that $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$. $\square$

Corollary 7.12 (Special cases)

- Margins: $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma) \overset{\boldsymbol{a} = \boldsymbol{e}_j}{\underset{\not\Leftarrow}{\Rightarrow}} X_j \sim \mathrm{N}(\mu_j, \Sigma_{jj})$, $j \in \{1, \dots, d\}$.
- Sums: $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma) \overset{\boldsymbol{a} = \boldsymbol{1}}{\underset{\not\Leftarrow}{\Rightarrow}} S_d \sim \mathrm{N}(\sum_{j=1}^d \mu_j,\ \sum_{i,j=1}^d \Sigma_{ij})$. If $X_1, \dots, X_d$ are uncorrelated, then $S_d \sim \mathrm{N}(\sum_{j=1}^d \mu_j,\ \sum_{j=1}^d \Sigma_{jj}) \overset{\text{same marginal}}{\underset{\text{means, var.}}{=}} \mathrm{N}(d\mu, d\sigma^2)$.
- Means: $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma) \overset{\boldsymbol{a} = \boldsymbol{1}/d}{\underset{\not\Leftarrow}{\Rightarrow}} \bar{X}_d \sim \mathrm{N}(\frac{1}{d} \sum_{j=1}^d \mu_j,\ \frac{1}{d^2} \sum_{i,j=1}^d \Sigma_{ij})$. If $X_1, \dots, X_d$ are uncorrelated, then $\bar{X}_d \sim \mathrm{N}(\frac{1}{d} \sum_{j=1}^d \mu_j,\ \frac{1}{d^2} \sum_{j=1}^d \Sigma_{jj}) \overset{\text{same marginal}}{\underset{\text{means, var.}}{=}} \mathrm{N}(\mu, \frac{\sigma^2}{d})$.

7.2 The central limit theorem

$f_n \to f$, $a_n \to a \not\Rightarrow f_n(a_n) \to f(a)$. Example: $f_n(x) := x^n \to \mathbb{1}_{\{x=1\}} =: f(x)$, $x \in [0,1]$, and $a_n := 1-1/n \to 1 =: a \Rightarrow f_n(a_n) = (1-1/n)^n \to e^{-1} \neq 1 = f(a)$.

Lemma 7.13 (Convergence to the exponential)
If $a_n \underset{n\to\infty}{\to} a$, then $\lim_{n\to\infty}(1+\frac{a_n}{n})^n = e^a$.

Proof. $\log(1+x)' = \frac{1}{1-(-x)} \underset{|x|<1}{=} \sum_{k=0}^{\infty}(-x)^k$, so $\log(1+x) = -\sum_{k=1}^{\infty}\frac{(-x)^k}{k}$.

1) $|\log(1+x)-x| \overset{\text{Taylor}}{\underset{|x|<1}{=}} \left|\sum_{k=2}^{\infty}\frac{(-x)^k}{k}\right| \underset{\Delta}{\leq} x^2\sum_{k=2}^{\infty}\frac{|x|^{k-2}}{k} = x^2\sum_{k=0}^{\infty}\frac{|x|^k}{k+2} \leq$
   $x^2\frac{1}{2}\sum_{k=0}^{\infty}|x|^k \underset{|x|\leq 1/2}{\leq} x^2\frac{1}{2}\sum_{k=0}^{\infty}(1/2)^k \underset{\text{geom.}}{=} x^2 \ \forall\,|x|<1/2$.
2) $\frac{a_n}{n} \to 0$, so $\exists\, n_0 : |\frac{a_n}{n}| \leq 1/2 \ \forall\, n \geq n_0$. Also, $|a_n| \underset{(a_n)_{n\in\mathbb{N}}\text{ conv.}}{\leq} M \ \forall\, n$ for some M.
3) Hence $\forall\, n \geq n_0$,
$$\left|n\log\left(1+\frac{a_n}{n}\right)-a_n\right| = n\left|\log\left(1+\frac{a_n}{n}\right)-\frac{a_n}{n}\right| \underset{1)}{\leq} n\left(\frac{a_n}{n}\right)^2 \underset{2)}{\leq} \frac{M^2}{n} \underset{n\to\infty}{\to} 0.$$
4) Therefore, $|n\log(1+\frac{a_n}{n})-a| \underset{\Delta\text{-ineq.}}{\leq} |n\log(1+\frac{a_n}{n})-a_n| + |a_n-a| \overset{\text{3), ass.}}{\underset{n\to\infty}{\to}} 0$, so
   $\log((1+\frac{a_n}{n})^n) \to a \underset{\exp(\cdot)\text{ cont.}}{\Rightarrow} (1+\frac{a_n}{n})^n \to e^a$. □

Lemma 7.14 (Expansion of cfs near 0)
If $\mathbb{E}(|X|^m) < \infty$ for $m \in \mathbb{N}$, then $\phi_X(t) = \sum_{k=0}^m \frac{(it)^k}{k!}\mathbb{E}(X^k) + o(|t|^m)$ for $t \to 0$.

Proof. Recall that $h \underset{t\to 0}{=} o(g)$ iff $\frac{|h(t)|}{g(t)} \underset{t\to 0}{\to} 0$. By L. 7.8 1),

$$\left|e^{ix} - \sum_{k=0}^m \frac{(ix)^k}{k!}\right| \le \min\left\{\frac{|x|^{m+1}}{(m+1)!}, \frac{2|x|^m}{m!}\right\}, \quad x \in \mathbb{R}.$$

Using $x = tX$ and taking expectations, we obtain

$$\left|\phi_X(t) - \sum_{k=0}^m \frac{(it)^k \mathbb{E}(X^k)}{k!}\right| \overset{\text{lin.}}{\underset{\Delta}{\le}} \mathbb{E}\left(\left|e^{itX} - \sum_{k=0}^m \frac{(itX)^k}{k!}\right|\right)$$
$$\le |t|^m \mathbb{E}\left(\min\left\{\frac{|t||X|^{m+1}}{(m+1)!}, \frac{2|X|^m}{m!}\right\}\right).$$

Pointwise, $\min\left\{\frac{|t||X|^{m+1}}{(m+1)!}, \frac{2|X|^m}{m!}\right\} \underset{t\to 0}{\to} 0$ and the term is bounded by $\frac{2|X|^m}{m!}$ which is integrable by assumption $\overset{\text{DOM}}{\Rightarrow} \mathbb{E}\left(\min\left\{\frac{|t||X|^{m+1}}{(m+1)!}, \frac{2|X|^m}{m!}\right\}\right) \underset{t\to 0}{\to} \mathbb{E}(0) = 0$, so equals $o(1)$ for $t \to 0$. Furthermore, $|t|^m o(1) = o(|t|^m)$ for $t \to 0$ ✓. □

We now have all ingredients ready to prove the *central limit theorem*.

Theorem 7.15 (Central limit theorem (CLT))

If $(X_n)_{n\in\mathbb{N}}$ is a sequence of iid rvs with $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \mathrm{var}(X_1) < \infty$, then

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \mathrm{N}(0,1).$$

Proof. With $Z_k := (X_k - \mu)/\sigma$ we have left to show that $\sqrt{n}\bar{Z}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \mathrm{N}(0,1)$.

$$\phi_{\sqrt{n}\bar{Z}_n}(t) = \mathbb{E}\bigl(e^{i\frac{t}{\sqrt{n}}\sum_{k=1}^n Z_k}\bigr) \underset{\text{ind.}}{=} \prod_{k=1}^n \mathbb{E}\bigl(e^{i\frac{t}{\sqrt{n}}Z_k}\bigr) \underset{\text{id}}{=} \phi_{Z_1}^n\Bigl(\frac{t}{\sqrt{n}}\Bigr)$$

$$\overset{\text{L. 7.14, } m=2}{\underset{\text{at } t \leftarrow t/\sqrt{n}}{=}} \Bigl(1 + i\frac{t}{\sqrt{n}}\mathbb{E}(Z_1) - \frac{t^2/n}{2}\mathbb{E}(Z_1^2) + o\Bigl(\frac{t^2}{n}\Bigr)\Bigr)^n$$

$$\overset{\mathbb{E}(Z_1)=0}{\underset{\mathbb{E}(Z_1^2)=1}{=}} \Bigl(1 - \frac{\frac{t^2}{2} - o(t^2/n)/(1/n)}{n}\Bigr)^n \overset{\text{L. 7.13}}{\underset{n\to\infty}{\to}} e^{-\frac{t^2}{2}}, \quad t \in \mathbb{R}.$$

$$\underset{\text{T. 7.5 2)}}{\Rightarrow} \sqrt{n}\bar{Z}_n \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} Z \text{ with cf } \phi_Z(t) = e^{-\frac{t^2}{2}} \overset{\text{uniqueness}}{\underset{\text{E. 7.4 1)}}{\Rightarrow}} Z \sim \mathrm{N}(0,1).$$

□

Remark 7.16

1) Note that $\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$.

2) The factor $\sqrt{n}$ is crucial in the CLT, as $\frac{\bar{X}_n-\mu}{\sigma} \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} 0$ by the SLLN.

3) By the CLT, we (approximately) know $\bar{X}_n \underset{\text{n large}}{\overset{\text{approx.}}{\sim}} \mathrm{N}(\mu, \frac{\sigma^2}{n})$ and $S_n \underset{n \text{ large}}{\overset{\text{approx.}}{\sim}} \mathrm{N}(n\mu, n\sigma^2)$ even though we don't know F (only that it has finite second moment).

4) Under iid and finite second moment, $\bar{X}_n \underset{\text{n large}}{\overset{\text{approx.}}{\sim}} \mathrm{N}(\mu, \frac{\sigma^2}{n})$. Compare:
   - For an iid sequences from any distribution with finite second moments, L. 5.32 implies that $\mathbb{E}(\bar{X}_n) = \mu$ and $\mathrm{var}(\bar{X}_n) = \frac{\sigma^2}{n}$ (just the distribution of $\bar{X}_n$ was unclear before).
   - If $X_1, \dots, X_n \overset{\text{ind.}}{\sim} \mathrm{N}(\mu, \sigma^2) \underset{\text{C. 7.12}}{\Rightarrow} \bar{X}_n \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$. The CLT asymptotically obtains the same result, just with the assumption of normality replaced by a second moment.

5) If $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ is iid with $\boldsymbol{\mu} = (\mathbb{E}(X_{1,1}), \dots, \mathbb{E}(X_{1,d}))$ and finite covariances $\Sigma_{j_1,j_2} = \mathrm{cov}(X_{1,j_1}, X_{1,j_2})$, $j_1, j_2 \in \{1, \dots, d\}$, the Cramér–Wold device and CLT imply the *multivariate CLT* $\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}) \underset{n\to\infty}{\overset{\text{d}}{\to}} \mathrm{N}(\boldsymbol{0}, \Sigma)$.

6) If $\mathbb{E}(|X_1|^3) < \infty$, the *Berry–Esseen Theorem* states that there exists a constant

$c \in (1/\sqrt{2\pi}, 0.4748)$ such that

$$\sup_{x\in\mathbb{R}} |F_{\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}}(x) - \Phi(x)| \le \frac{c\mathbb{E}(|\frac{X_1-\mu}{\sigma}|^3)}{\sqrt{n}}, \quad n \in \mathbb{N}.$$

This provides a rate of convergence in the CLT under known third moment. For a proof (for larger c), see Durrett (2019, Theorem 3.4.17).

7) A generalization of the CLT is the *Lindeberg–Feller CLT*: Here $(X_n)_{n\in\mathbb{N}}$ are independent with $\sigma_n^2 = \operatorname{var}(X_n) < \infty$, $\mu_n = \mathbb{E}(X_n)$ and $s_n^2 = \operatorname{var}(\sum_{k=1}^n X_k) = \sum_{k=1}^n \sigma_k^2$. If the *Lindeberg condition*

$$\forall\, \varepsilon > 0, \quad \frac{1}{s_n^2}\sum_{k=1}^n \mathbb{E}\Big((X_k-\mu_k)^2 \mathbb{1}_{\frac{|X_k-\mu_k|}{s_n}>\varepsilon}\Big) \underset{n\to\infty}{\to} 0$$

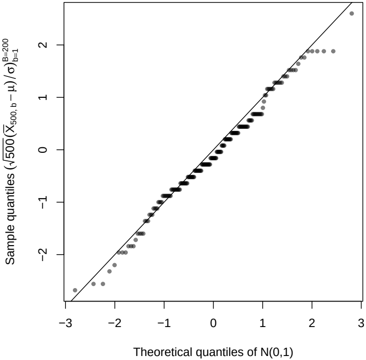
holds, then $\frac{\sum_{k=1}^n (X_k-\mu_k)}{s_n} \underset{n\to\infty}{\overset{\mathrm{d}}{\to}} \mathrm{N}(0,1)$. On can show that the Lindeberg condition holds if the *Lyapunov condition*

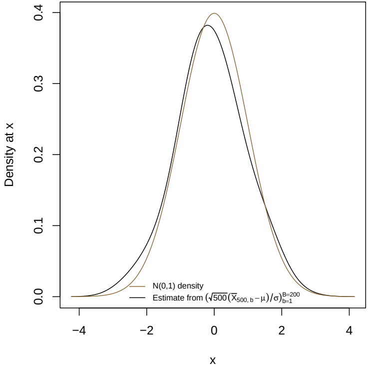
$$\exists\, \delta > 0, \quad \frac{\sum_{k=1}^n \mathbb{E}(|X_k-\mu_k|^{2+\delta})}{s_n^{2+\delta}} \underset{n\to\infty}{\to} 0$$

holds.

Example 7.17 (E. 7.1 continued)

We visually assess the sample $(\sqrt{n}\frac{\bar{X}_{n,b}-\mu}{\sigma})_{b=1}^{B}$ from E. 7.1 against $\mathrm{N}(0,1)$ with a *Q-Q plot* (lhs; $\{(F_0^{-1}(\frac{i-1/2}{n}), x_{(i)}) = (F_0^{-1}(\frac{i-1/2}{n}), \hat{F}_n^{-1}(\frac{i}{n})) : i = 1, \dots, n\}$ for $F_0 = \Phi$, so that if $F_0 \approx F$, the points lie on the line $y = x$) and a density estimate (rhs).





References

Durrett, R. (2019), Probability: Theory and Examples, 5th ed., Duxbury Press.

8 Conditional expectation

8.1 Ordinary conditional probability and implied conditional expectation

- **Idea:** Calculate expectations (incl. probabilities) under additional information.
- **Recall:** On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, if $B \in \mathcal{F} : \mathbb{P}(B) > 0$, $\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, $A \in \mathcal{F}$, is the (ordinary) conditional probability of A given B.
- **Motivation:** For a random vector $\boldsymbol{X}_2 : \Omega \to \mathbb{R}^{d_2}$, we are often interested in $A = \{\boldsymbol{X}_2 \le \boldsymbol{x}_2\} = \{\omega \in \Omega : \boldsymbol{X}_2(\omega) \le \boldsymbol{x}_2\}$. If $\mathbb{P}(B) > 0$, then

$$F_{\boldsymbol{X}_2|B}(\boldsymbol{x}_2) := \mathbb{P}(\boldsymbol{X}_2 \le \boldsymbol{x}_2 \,|\, B) = \frac{\mathbb{P}(\{\boldsymbol{X}_2 \le \boldsymbol{x}_2\} \cap B)}{\mathbb{P}(B)}, \quad \boldsymbol{x}_2 \in \mathbb{R}^d,$$

  is the *(ordinary) conditional df of $\boldsymbol{X}_2$ given B*; see E. 4.17, where $F_{\boldsymbol{X}|I=i}$ appeared. If it exists, the mean of $F_{\boldsymbol{X}_2|B}$ is the *(ordinary) conditional expectation of $\boldsymbol{X}_2$ given B*.
- **Interpretation:** For fixed B, $\boldsymbol{x}_2 \mapsto F_{\boldsymbol{X}_2|B}(\boldsymbol{x}_2)$ is a df depending on B. Oftentimes, $B = \{\boldsymbol{X}_1 = \boldsymbol{x}_1\}$ for $(\boldsymbol{X}_1, \boldsymbol{X}_2) \sim F$.
- **Question:** What if $F_{\boldsymbol{X}_1}$ is continuous, so $\mathbb{P}(B) = 0$?

- If $F_{\boldsymbol{X}_1}$ is discrete, then $B = \{\boldsymbol{X}_1 = \boldsymbol{x}_1\}$ satisfies $\mathbb{P}(B) > 0\ \forall\, \boldsymbol{x}_1 \in \operatorname{supp}(F_{\boldsymbol{X}_1})$ and we can use

$$F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1) := \begin{cases} \mathbb{P}(\boldsymbol{X}_2 \le \boldsymbol{x}_2 \,|\, \boldsymbol{X}_1 = \boldsymbol{x}_1), & \boldsymbol{x}_1 \in \operatorname{supp}(F_{\boldsymbol{X}_1}), \\ \text{e.g. } 0, & \text{otherwise.} \end{cases} \tag{6}$$

- If $F_{\boldsymbol{X}_1}$ has density $f_{\boldsymbol{X}_1}$, then $\mathbb{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1) = 0\ \forall\, \boldsymbol{x}_1 \in \mathbb{R}^{d_1}$ and using (6) is no longer possible.
- **Idea:** Let $(\boldsymbol{X}_1, \boldsymbol{X}_2) \sim F$. For $\boldsymbol{x}_1 \in \mathbb{R}^{d_1} : f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) > 0$, $\boldsymbol{h} = (h_1, \dots, h_{d_1}) > \boldsymbol{0}$,

$$\mathbb{P}(\boldsymbol{X}_2 \le \boldsymbol{x}_2 \,|\, \boldsymbol{X}_1 \in (\boldsymbol{x}_1 - \boldsymbol{h}, \boldsymbol{x}_1]) = \frac{\mathbb{P}(\boldsymbol{X}_1 \in (\boldsymbol{x}_1 - \boldsymbol{h}, \boldsymbol{x}_1], \boldsymbol{X}_2 \le \boldsymbol{x}_2)}{\mathbb{P}(\boldsymbol{X}_1 \in (\boldsymbol{x}_1 - \boldsymbol{h}, \boldsymbol{x}_1])}$$
$$= \frac{\Delta_{\left(\binom{\boldsymbol{x}_1 - \boldsymbol{h}}{-\boldsymbol{\infty}}, \binom{\boldsymbol{x}_1}{\boldsymbol{x}_2}\right]} F}{\Delta_{(\boldsymbol{x}_1 - \boldsymbol{h}, \boldsymbol{x}_1]} F_{\boldsymbol{X}_1}}.$$

- If F has density f (similarly if $\boldsymbol{X}_1$ has a density and $\boldsymbol{X}_2$ is discrete) and $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^d : f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) > 0$, then

$$\mathbb{P}(\boldsymbol{X}_2 \le \boldsymbol{x}_2 \,|\, \boldsymbol{X}_1 \in (\boldsymbol{x}_1 - \boldsymbol{h}, \boldsymbol{x}_1]) = \frac{\int_{-\boldsymbol{\infty}}^{\boldsymbol{x}_2} \int_{\boldsymbol{x}_1 - \boldsymbol{h}}^{\boldsymbol{x}_1} f(\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2)\, \mathrm{d}\tilde{\boldsymbol{x}}_1\, \mathrm{d}\tilde{\boldsymbol{x}}_2}{\int_{\boldsymbol{x}_1 - \boldsymbol{h}}^{\boldsymbol{x}_1} f_{\boldsymbol{X}_1}(\tilde{\boldsymbol{x}}_1)\, \mathrm{d}\tilde{\boldsymbol{x}}_1}$$

$$= \frac{\int_{-\infty}^{x_2} \frac{1}{\prod_{j=1}^{d_1} h_j} \int_{x_1 - h}^{x_1} f(\tilde{x}_1, \tilde{x}_2) \,\mathrm{d}\tilde{x}_1 \,\mathrm{d}\tilde{x}_2}{\frac{1}{\prod_{j=1}^{d_1} h_j} \int_{x_1 - h}^{x_1} f_{X_1}(\tilde{x}_1) \,\mathrm{d}\tilde{x}_1}$$

$$\underset{h \to 0+}{=} \frac{\int_{-\infty}^{x_2} f(x_1, \tilde{x}_2) \,\mathrm{d}\tilde{x}_2}{f_{X_1}(x_1)}$$

by the Lebesgue differentiation theorem. One would thus like to define

$$F_{X_2 | X_1}(x_2 \,|\, x_1) = \lim_{h \to 0+} \mathbb{P}(X_2 \le x_2 \,|\, X_1 \in (x_1 - h, x_1]), \tag{7}$$

with conditional density $f_{X_2|X_1}(x_2 \,|\, x_1) = \frac{\partial}{\partial x_2} F_{X_2|X_1}(x_2 \,|\, x_1)$ being

$$f_{X_2|X_1}(x_2 \,|\, x_1) = \begin{cases} \frac{f(x_1, x_2)}{f_{X_1}(x_1)}, & f_{X_1}(x_1) > 0, \\ \text{e.g. } 0, & \text{otherwise.} \end{cases} \tag{8}$$

- Although this is indeed often used, there are two problems:
  1) A joint density does not always exist.
  2) Even if, 'definitions' such as (7) may lead to ill-definedness (different results depending on how the limit is computed).

Example 8.1 (Borel—Kolmogorov paradox for uniform distr. on semi-circle)

Question: If $(X_1, X_2) \sim \mathrm{U}(D)$ for $D = \left\{\binom{x_1}{x_2} : x_1 \in [-1,1], x_2 \in \left[0, \sqrt{1-x_1^2}\right]\right\}$, what is $\mathbb{P}(X_2 \in [0,1/2] \,|\, X_1 = 0)$?

- **Version 1 (Cartesian coordinates):**
  $f(x_1,x_2) = \frac{1}{1^2\pi/2}\mathbb{1}_D((x_1,x_2)) = \frac{2}{\pi}\mathbb{1}_D((x_1,x_2))$ and $f_{X_1}(x_1) = \frac{2}{\pi}\int_0^{\sqrt{1-x_1^2}}\,\mathrm{d}x_2$ $= \frac{2\sqrt{1-x_1^2}}{\pi}\mathbb{1}_{[-1,1]}(x_1)$, so $f_{X_2|X_1}(x_2\,|\,x_1) \underset{\text{if } x_1 \in (-1,1)}{\overset{(8)}{=}} \frac{f(x_1,x_2)}{f_{X_1}(x_1)} = \frac{1}{\sqrt{1-x_1^2}}\mathbb{1}_D((x_1,x_2))$.
  Hence $f_{X_2|X_1}(x_2\,|\,0) = \mathbb{1}_{[0,1]}(x_2)$, so $\mathbb{P}(X_2 \in [0,1/2] \,|\, X_1 = 0) = \int_0^{1/2} 1\,\mathrm{d}x_2 = \frac{1}{2}$.
- **Version 2 (Polar coordinates):**
  Substituting $\binom{x_1}{x_2} = \binom{r\cos(\theta)}{r\sin(\theta)} =: h(r,\theta)$ with $\det\left(\left(\frac{\partial h}{\partial(r,\theta)}\right)\right) = \begin{vmatrix}\cos(\theta) & -r\sin(\theta)\\ \sin(\theta) & r\cos(\theta)\end{vmatrix} =$ r, we have $1 = \int_D \frac{2}{\pi}\,\mathrm{d}(x_1,x_2) \underset{\text{polar}}{=} \int_0^\pi\int_0^1 \frac{2}{\pi}\,r\,\mathrm{d}r\,\mathrm{d}\theta$, so $f_{R,\Theta}(r,\theta) = \frac{2}{\pi}r$ is a density on $[0,1]\times[0,\pi]$. Thus $f_\Theta(\theta) = \int_0^1 \frac{2}{\pi}r\,\mathrm{d}r = \frac{1}{\pi}$ and $f_{R|\Theta}(r\,|\,\theta) = \frac{f_{R,\Theta}(r,\theta)}{f_\Theta(\theta)} =$ $\frac{2r/\pi}{1/\pi} = 2r$, $r \in [0,1]$, so $\mathbb{P}(R \in [0,1/2] \,|\, \Theta = \pi/2) = \int_0^{1/2} 2r\,\mathrm{d}r = \frac{1}{4}$.
- Hence we can get different limits if we approximate the same “slice” differently.

Instead, we need a measure-theoretic definition of (first conditional expectations and then) conditional probabilities.

8.2 Measure-theoretic conditional expectation and implied conditional probability

Definition 8.2 (Conditional expectation, conditional probability, etc.)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ a σ-algebra on Ω and $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. The *conditional expectation of X given $\mathcal{G}$* (notation: $\mathbb{E}(X \,|\, \mathcal{G})$) is any $Y : \Omega \to \mathbb{R}$ such that

i) $Y \in \mathcal{G}$, i.e. Y is $\mathcal{G}$-measurable ($\Rightarrow$ characteristics of a rv); and
ii) $\mathbb{E}(Y \mathbb{1}_A) = \mathbb{E}(X \mathbb{1}_A) \;\forall\, A \in \mathcal{G}$ ($\Rightarrow$ characteristics of an expectation).

Any such Y is called a *version of* $\mathbb{E}(X \,|\, \mathcal{G})$.

- For $X_j \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, $j = 1, \dots, d$, $\mathbb{E}(\boldsymbol{X} \,|\, \mathcal{G}) := (\mathbb{E}(X_1 \,|\, \mathcal{G}), \dots, \mathbb{E}(X_d \,|\, \mathcal{G}))$.
- If $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$, $\mathbb{E}(X^p \,|\, \mathcal{G})$ is the *conditional pth moment of X given $\mathcal{G}$*. $\mathrm{var}(X \,|\, \mathcal{G}) := \mathbb{E}\big((X - \mathbb{E}(X \,|\, \mathcal{G}))^2 \,\big|\, \mathcal{G}\big)$ is the *conditional variance of X given $\mathcal{G}$*.
- $\mathbb{P}(A \,|\, \mathcal{G}) := \mathbb{E}(\mathbb{1}_A \,|\, \mathcal{G})$, $A \in \mathcal{F}$, is a *version of the cond. prob. of A given $\mathcal{G}$*.
- If $(Z_i)_{i \in I}$ are rvs, then $\mathbb{E}(X \,|\, Z_i, i \in I) := \mathbb{E}(X \,|\, \sigma(Z_i, i \in I))$.

$\mathbb{E}(X \,|\, \mathcal{G})$ is any member of the equivalence class of rvs that satisfy the defining properties i)–ii). $\mathbb{E}(X \,|\, \mathcal{G})$ is often found by a guess and verification of i)–ii).

Lemma 8.3 (Integrability)
Any $Y : \Omega \to \mathbb{R}$ satisfying i)–ii) is indeed integrable.

Proof. With $A := \{Y \geq 0\} = Y^{-1}([0,\infty)) \underset{\text{i)}}{\in} \mathcal{G}$, we have

$$\mathbb{E}(|Y|) \underset{\text{L. 5.8 6)}}{=} \int_A Y \,\mathrm{d}\mathbb{P} + \int_{A^c} (-Y)\,\mathrm{d}\mathbb{P} \underset{A^c \in \mathcal{G}}{\overset{\text{ii),}}{=}} \int_A X \,\mathrm{d}\mathbb{P} + \int_{A^c} (-X)\,\mathrm{d}\mathbb{P}$$
$$\underset{\text{mon.}}{\leq} \int_A |X|\,\mathrm{d}\mathbb{P} + \int_{A^c} |X|\,\mathrm{d}\mathbb{P} = \int_\Omega |X|\,\mathrm{d}\mathbb{P} \underset{\text{L. 5.8 6)}}{=} \mathbb{E}(|X|) \underset{\text{ass.}}{<} \infty. \qquad \square$$

Theorem 8.4 (Existence and uniqueness a.s.)
$\mathbb{E}(X \,|\, \mathcal{G})$ exists and is unique a.s..

Proof.

- Consider existence. Assume first that $X \geq 0$. $\underset{\text{L. 5.8 6)}}{\Rightarrow}$ $\nu(A) := \mathbb{E}(X \mathbb{1}_A)$, $A \in \mathcal{G}$, is a measure on $(\Omega, \mathcal{G})$. By L. 5.11 6), $\nu \ll \mathbb{P}$. Therefore,

$$\int_A X \,\mathrm{d}\mathbb{P} \underset{\text{def.}}{=} \mathbb{E}(X\mathbb{1}_A) \underset{\text{def. } \nu}{=} \nu(A) \underset{\text{RN}}{=} \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mathbb{P}}\,\mathrm{d}\mathbb{P}, \quad A \in \mathcal{G},$$

for the $\mathbb{P}$-a.s. unique integrable, and thus (by definition) $\mathcal{G}$-measurable, RN derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mathbb{P}} : \Omega \to [0,\infty)$. By definition, $\frac{\mathrm{d}\nu}{\mathrm{d}\mathbb{P}}$ is thus a version of $\mathbb{E}(X \,|\, \mathcal{G})$.

- In general, let $X = X^+ - X^-$ and $Y^- := \mathbb{E}(X^- \,|\, \mathcal{G})$, $Y^+ := \mathbb{E}(X^+ \,|\, \mathcal{G})$. Then $Y := Y^+ - Y^- \underset{\text{sum}}{\in} \mathcal{G}$, Y is integrable (as Y^-, Y^+ are by L. 8.3) and
$$\int_A X \,\mathrm{d}\mathbb{P} \underset{\text{lin.}}{=} \int_A X^+ \,\mathrm{d}\mathbb{P} - \int_A X^- \,\mathrm{d}\mathbb{P} \underset{\text{def.}}{=} \int_A Y^+ \,\mathrm{d}\mathbb{P} - \int_A Y^- \,\mathrm{d}\mathbb{P} \underset{\text{def. } Y}{\overset{\text{lin.}}{=}} \int_A Y \,\mathrm{d}\mathbb{P},$$
for all $A \in \mathcal{G}$, so Y is a version of $\mathbb{E}(X \,|\, \mathcal{G})$.
- Consider uniqueness a.s.. If $Y, \tilde{Y} : \Omega \to \mathbb{R}$ satisfy i)–ii), then $\mathbb{E}(Y \mathbb{1}_A) \underset{\text{ii)}}{=} \mathbb{E}(X \mathbb{1}_A) \underset{\text{ii)}}{=} \mathbb{E}(\tilde{Y} \mathbb{1}_A)\ \forall A \in \mathcal{G}$ and, by L. 8.3, $Y, \tilde{Y}$ are integrable $\underset{\text{P. 5.14}}{\Rightarrow} Y = \tilde{Y}$ a.s.. □

Remark 8.5 (Conditional probability)

- Equalities such as $Y = \mathbb{E}(X \,|\, \mathcal{G})$ are understood a.s., so $Y \overset{\text{a.s.}}{=} \mathbb{E}(X \,|\, \mathcal{G})$, so we implicitly work with a representative of the equivalence class of a.s. equal rvs. Another convention: $\mathbb{E}(X \,|\, \mathcal{A}) := \mathbb{E}(X \,|\, \sigma(\mathcal{A}))$, $\mathcal{A} \subseteq \mathcal{F}$, and $\mathbb{E}(X \,|\, Z) := \mathbb{E}(X \,|\, \sigma(Z))$ for a random element Z.
- If we consider $\mathbb{P}(A \,|\, \mathcal{G}) = \mathbb{E}(\mathbb{1}_A \,|\, \mathcal{G})$ for countably-many different $A \in \mathcal{F}$, there is no problem (the countable union of null sets is a null set). But for uncountably-many different $A \in \mathcal{F}$, we need to be careful, e.g. when defining a conditional distribution function of X given $\mathcal{G}$, so $\mathbb{P}(X \leq x \,|\, \mathcal{G})$, $x \in \mathbb{R}$.

- Generally, it is wrong that $A \mapsto \mathbb{P}(A\,|\,\mathcal{G})$ is a probability measure since $\mathbb{P}(A\,|\,\mathcal{G})$ is only defined up to a.s. equality and so this map may not be well-defined (we would need to specify a *particular* version).
- Even then, $A \mapsto \mathbb{P}(A\,|\,\mathcal{G})$ may not be a proper probability measure. It would need to be σ-additive, so $\mathbb{P}(\biguplus_{i=1}^{\infty} A_i\,|\,\mathcal{G}) \underset{(*)}{\overset{\text{a.s.}}{=}} \sum_{i=1}^{\infty} \mathbb{P}(A_i\,|\,\mathcal{G})$ for any countable collection of pairwise disjoint sets $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$. However, in general there can be uncountably-many such $\{\{A_i\}_{i\in\mathbb{N}}\}_{i\in I} \subseteq \mathcal{F}$, each leaving us with a potentially different null set N_i, $i \in I$, where $(*)$ would not hold. Whether $\bigcup_{i\in I} N_i \in \mathcal{F}$ and, if so, whether $\mathbb{P}(\bigcup_{i\in I} N_i) = 0$ is unclear.
- If $(\Omega, \mathcal{F})$, $(\Omega', \mathcal{F}')$ are measurable spaces and $X : \Omega \mapsto \Omega'$ is $(\mathcal{F}, \mathcal{F}')$-measurable, then $\mathbb{P}^* : \mathcal{F} \times \Omega \mapsto [0, 1]$ is a *regular conditional probability measure given* $\mathcal{G}$ if
  i) $\forall\, A \in \mathcal{F}$, $\mathbb{P}^*(A, \cdot)$ is a version of $\mathbb{P}(A\,|\,\mathcal{G})$ (verbose: $\forall\, A \in \mathcal{F}$, the rv $\omega \mapsto \mathbb{P}^*(A, \omega)$ is a version of the rv $\omega \mapsto \mathbb{P}(A\,|\,\mathcal{G})(\omega)$ satisfying i)–ii)); and
  ii) $\forall\, \omega \in \Omega$, $\mathbb{P}^*(\cdot, \omega)$ is a probability measure on $(\Omega, \mathcal{F})$.

  Similarly, $F^*_{\boldsymbol{X}|\mathcal{G}} : \mathbb{R}^d \times \Omega \mapsto [0, 1]$ is a *regular conditional df of $\boldsymbol{X}$ given $\mathcal{G}$* if
  i) $\forall\, \boldsymbol{x} \in \mathbb{R}^d$, $F^*_{\boldsymbol{X}|\mathcal{G}}(\boldsymbol{x}, \cdot)$ is a version of $\mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}\,|\,\mathcal{G})$; and
  ii) $\forall\, \omega \in \Omega$, $F^*_{\boldsymbol{X}|\mathcal{G}}(\cdot, \omega)$ is a df.

- One can show that if $(\Omega', \mathcal{F}')$ is *nice* (so if $\exists$ a one-to-one $\varphi : \Omega' \to \mathbb{R}^d$ so that φ, φ^{-1} are measurable), then regular condititional probability measures and dfs exist. In particular, $(\Omega', \mathcal{F}') = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $d \ge 1$, is nice, which is the case in most applications and which we assume moving forward, so we assume to work with regular versions.

Lemma 8.6 ($\mathcal{G}$ being generated by a countable partition)

Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. If $\mathcal{A} = \{A_n\}_{n\in\mathbb{N}}$ is a partition of Ω, let $\mathcal{G} = \sigma(A_n, n \in \mathbb{N}) \underset{\text{L. 2.11}}{=} \{\biguplus_{i\in I} A_i : A_i \in \mathcal{A}\ \forall\, i \in I, I \subseteq \mathbb{N}\}$. Then

$$\mathbb{E}(X \,|\, \mathcal{G}) = \sum_{n=1}^{\infty} \mathbb{E}_{A_n}(X)\mathbb{1}_{A_n} \quad \text{where } \mathbb{E}_{A_n}(X) := \begin{cases} \frac{\mathbb{E}(X\mathbb{1}_{A_n})}{\mathbb{P}(A_n)}, & \mathbb{P}(A_n) > 0, \\ 0, & \mathbb{P}(A_n) = 0. \end{cases}$$

Proof. Let $Y := \sum_{n=1}^{\infty} \mathbb{E}_{A_n}(X)\mathbb{1}_{A_n} : \Omega \to \mathbb{R}$ be a candiate for a version of $\mathbb{E}(X \,|\, \mathcal{G})$. We verify the defining properties i)–ii) of $\mathbb{E}(X \,|\, \mathcal{G})$.

1) $Y_m := \sum_{n=1}^{m} \mathbb{E}_{A_n}(X)\mathbb{1}_{A_n} + 0 \cdot \mathbb{1}_{\biguplus_{k=m+1}^{\infty} A_k} \overset{Y_m \text{ simple}}{\underset{\text{E. 3.3 3)}}{\in}} \mathcal{G} \underset{\text{L. 3.12 2)}}{\Rightarrow} Y = \lim_{m\to\infty} Y_m \in \mathcal{G}$ $\Rightarrow$ i) ✓.

2) Preparation for ii): $|Y| \underset{\Delta}{\le} \sum_{n=1}^{\infty} |\mathbb{E}_{A_n}(X)|\mathbb{1}_{A_n} \underset{\Delta}{\le} \sum_{n=1}^{\infty} \mathbb{E}_{A_n}(|X|)\mathbb{1}_{A_n}$, so

$$\mathbb{E}(|Y|) \underset{\text{mon.}}{\leq} \mathbb{E}\Big(\sum_{n=1}^{\infty} \mathbb{E}_{A_n}(|X|)\mathbb{1}_{A_n}\Big) \underset{\text{MON}}{=} \lim_{N\to\infty} \mathbb{E}\Big(\sum_{n=1}^{N} \mathbb{E}_{A_n}(|X|)\mathbb{1}_{A_n}\Big)$$

$$\underset{\text{lin.}}{=} \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{E}_{A_n}(|X|)\mathbb{P}(A_n) \underset{\text{def.}}{=} \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{E}(|X|\mathbb{1}_{A_n})$$

$$\underset{\text{lin.}}{=} \lim_{N\to\infty} \mathbb{E}\Big(\sum_{n=1}^{N} |X|\mathbb{1}_{A_n}\Big) = \lim_{N\to\infty} \mathbb{E}(|X|\mathbb{1}_{\biguplus_{n=1}^{N} A_n}) \underset{\text{MON}}{=} \mathbb{E}(|X|) < \infty.$$

Therefore, $\forall\, A \in \mathcal{G}$, $\mathbb{E}(|Y\mathbb{1}_A|) \underset{\text{mon.}}{\leq} \mathbb{E}(|Y|) < \infty$, so $Y\mathbb{1}_A \in L^1(\Omega, \mathcal{F}, \mathbb{P})$.

3) $A \in \mathcal{G} \Rightarrow A = \biguplus_{i\in I} A_i$ for some countable $I \subseteq \mathbb{N}$. Then

$$\mathbb{E}(Y\mathbb{1}_A) \underset{\text{L. 5.12}}{\overset{2)}{=}} \sum_{i\in I} \mathbb{E}(Y\mathbb{1}_{A_i}) \underset{\text{def.}}{=} \sum_{i\in I} \int_{A_i} Y \,\mathrm{d}\mathbb{P} \underset{\text{on } A_i}{\overset{Y = \mathbb{E}_{A_i}(X)}{=}} \sum_{i\in I} \int_{A_i} \mathbb{E}_{A_i}(X) \,\mathrm{d}\mathbb{P}$$

$$\underset{\text{def.}}{=} \sum_{i\in I} \int_{\Omega} \mathbb{E}_{A_i}(X)\mathbb{1}_{A_i} \,\mathrm{d}\mathbb{P} \underset{\text{lin.}}{=} \sum_{i\in I} \mathbb{E}_{A_i}(X) \int_{\Omega} \mathbb{1}_{A_i} \,\mathrm{d}\mathbb{P}$$

$$\underset{\text{L. 5.3 5)}}{=} \sum_{i\in I} \underbrace{\mathbb{E}_{A_i}(X)\mathbb{P}(A_i)}_{\underset{\text{def.}}{=} \begin{cases} \frac{\mathbb{E}(X\mathbb{1}_{A_i})}{\cancel{\mathbb{P}(A_i)}}\cancel{\mathbb{P}(A_i)}, & \mathbb{P}(A_i) > 0 \\ 0 \cdot 0, & \mathbb{P}(A_i) = 0 \end{cases} \underset{\text{L. 5.11 6)}}{=} \mathbb{E}(X\mathbb{1}_{A_i})} = \sum_{i\in I} \mathbb{E}(X\mathbb{1}_{A_i}) \underset{\text{L. 5.12}}{\overset{2)}{=}} \mathbb{E}(X\mathbb{1}_A).$$

□

Interpretation:

- **Conditional expectation.** Consider an experiment with sample space Ω in which an unknown outcome ω has happened. We only know that $\omega \in A_n$ for some known $n \in \mathbb{N}$. Conditional on this information, our best prediction for X is thus the mean of X over A_n relative to the probability of A_n happening, so $\mathbb{E}(X \,|\, \mathcal{G})(\omega) \underset{\omega \in A_n}{=} \frac{\mathbb{E}(X \mathbb{1}_{A_n})}{\mathbb{P}(A_n)} \underset{\text{def.}}{=} \mathbb{E}_{A_n}(X)$. Therefore, $\mathbb{E}(X \,|\, \mathcal{G}) = \sum_{n=1}^{\infty} \mathbb{E}_{A_n}(X) \mathbb{1}_{A_n}$.
- **Conditional probability.** If $X = \mathbb{1}_A$, then

$$\mathbb{P}(A \,|\, \mathcal{G}) = \mathbb{E}(\mathbb{1}_A \,|\, \mathcal{G}) = \sum_{n=1}^{\infty} \frac{\mathbb{E}(\mathbb{1}_A \mathbb{1}_{A_n})}{\mathbb{P}(A_n)} \mathbb{1}_{A_n} = \sum_{n=1}^{\infty} \frac{\mathbb{E}(\mathbb{1}_{A \cap A_n})}{\mathbb{P}(A_n)} \mathbb{1}_{A_n}$$

$$\underset{\text{L. 5.35)}}{=} \sum_{n=1}^{\infty} \frac{\mathbb{P}(A \cap A_n)}{\mathbb{P}(A_n)} \mathbb{1}_{A_n} = \sum_{n=1}^{\infty} \mathbb{P}(A \,|\, A_n) \mathbb{1}_{A_n},$$

i.e. $\mathbb{P}(A \,|\, \mathcal{G})(\omega) \underset{\omega \in A_n}{=} \mathbb{P}(A \,|\, A_n)$ ($= 0$ if $\mathbb{P}(A_n) = 0$). In particular, if $\boldsymbol{X}$ has support $\{\boldsymbol{x}_n\}_{n \in \mathbb{N}}$, then $\mathcal{G} = \sigma(\boldsymbol{X}) = \{\boldsymbol{X}^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^d)\} = \sigma(\boldsymbol{X} = \boldsymbol{x}_n, n \in \mathbb{N}) = \sigma(A_n, n \in \mathbb{N})$, where $A_n = \{\boldsymbol{X} = \boldsymbol{x}_n\}$ are pairwise disjoint, so we obtain $\mathbb{P}(A \,|\, \boldsymbol{X}) = \sum_{n=1}^{\infty} \mathbb{P}(A \,|\, \boldsymbol{X} = \boldsymbol{x}_n) \mathbb{1}_{\{\boldsymbol{X} = \boldsymbol{x}_n\}}$.

Continuity of $T \mapsto I_\alpha(T)$, $T>0$, is clear. For $\alpha>0$, $|I_\alpha(T)-I_\alpha(0)| \underset{T\in[0,\pi/\alpha]}{\overset{I_\alpha(0)=0}{=}} I_\alpha(T) \underset{x\ge 0}{\overset{\sin(x)\le x}{\le}} \int_0^T \alpha\,\mathrm{d}t = \alpha T \underset{T\to 0+}{\to} 0$. Similarly for $\alpha<0$. For $\alpha=0$, $I_\alpha(T)=0$, $T\ge 0$, is clearly continuous.

Now consider $\lim_{T\to\infty} I_\alpha(T)$. The case $\alpha=0$ is trivial. For $\alpha>0$, note that

$$\int_0^{\alpha T}\int_0^\infty |\sin(x)e^{-xz}|\,\mathrm{d}z\,\mathrm{d}x \underset{\frac{1}{x}=\int_0^\infty e^{-xz}\,\mathrm{d}z}{=} \int_0^{\alpha T}\frac{|\sin(x)|}{x}\,\mathrm{d}x \underset{x\ge 0}{\overset{\sin(x)\le x}{\le}} \int_0^{\alpha T}\mathrm{d}x = \alpha T < \infty,$$

so by Tonelli, $\int_0^{\alpha T}\int_0^\infty \sin(x)e^{-xz}\,\mathrm{d}z\,\mathrm{d}x$ exists and Fubini implies that

$$I_\alpha(T) \underset{\text{shown}}{=} \int_0^{\alpha T}\frac{\sin(x)}{x}\,\mathrm{d}x \underset{\frac{1}{x}=\int_0^\infty e^{-xz}\,\mathrm{d}z}{=} \int_0^{\alpha T}\sin(x)\int_0^\infty e^{-xz}\,\mathrm{d}z\,\mathrm{d}x$$

$$\underset{\text{Fubini}}{=} \int_0^\infty \int_0^{\alpha T}\sin(x)e^{-xz}\,\mathrm{d}x\,\mathrm{d}z \underset{\text{def.}}{=} \int_0^\infty h(z)\,\mathrm{d}z.$$

We have

$$h(z) \underset{\text{by parts}}{=} \left[-\cos(x)e^{-xz}\right]_{x=0}^{\alpha T} - \int_0^{\alpha T}(-\cos(x))(-z)e^{-xz}\,\mathrm{d}x$$

$$= 1-\cos(\alpha T)e^{-z\alpha T} - z\int_0^{\alpha T}\cos(x)e^{-xz}\,\mathrm{d}x$$

$$\underset{\text{by parts}}{=} 1-\cos(\alpha T)e^{-z\alpha T} - z\Big(\left[\sin(x)e^{-xz}\right]_{x=0}^{\alpha T} - (-z)h(z)\Big)$$

$$= 1 - \cos(\alpha T)e^{-z\alpha T} - z\sin(\alpha T)e^{-z\alpha T} - z^2 h(z)$$

and thus $h(z) = \frac{1-(\cos(\alpha T)+z\sin(\alpha T))e^{-z\alpha T}}{1+z^2}$. With $\int_0^\infty \frac{1}{1+z^2}\,\mathrm{d}z = \big[\arctan(z)\big]_0^\infty = \frac{\pi}{2}$, we obtain

$$\begin{aligned} I_\alpha(T) &= \int_0^\infty \frac{1-(\cos(\alpha T)+z\sin(\alpha T))e^{-z\alpha T}}{1+z^2}\,\mathrm{d}z \\ &= \frac{\pi}{2} - \int_0^\infty \underbrace{\frac{(\cos(\alpha T)+z\sin(\alpha T))e^{-z\alpha T}}{1+z^2}}_{=:\,g(T,z)}\,\mathrm{d}z. \end{aligned}$$

Since $\lim_{T\to\infty} g(T,z) = 0\ \forall\, z > 0$, $|g(T,z)| \le \frac{M}{1+z^2}\ \forall\, T > 0$ for some $M > 0$ and $\frac{M}{1+z^2}$ is integrable, DOM implies that $\lim_{T\to\infty}\int_0^\infty g(T,z)\,\mathrm{d}z = \int_0^\infty 0\,\mathrm{d}z = 0$, so $\lim_{T\to\infty} I_\alpha(T) = \frac{\pi}{2}$.

Finally, for $\alpha < 0$, $\lim_{T\to\infty} I_\alpha(T) \underset{\text{shown}}{=} -\lim_{T\to\infty} I_{-\alpha}(T) \underset{\text{case }\alpha>0}{=} -\frac{\pi}{2}$.

3) By 2), $T \mapsto I_1(T)$, $T \in [0,\infty)$, is continuous with a finite limit for $T \to \infty$. Hence $\sup_{T\in[0,\infty)} |I_1(T)| < \infty$. Since $\forall\, \alpha \in \mathbb{R}$ and $T \ge 0$, we have $I_\alpha(T) \underset{2)}{=} I_1(\alpha T)$, we obtain that $|I_\alpha(T)| = |I_1(|\alpha| T)| \underset{\tilde{T}=|\alpha|T}{\le} \sup_{\tilde{T}\in[0,\infty)} |I_1(\tilde{T})| \underset{\text{before}}{<} \infty$. □

Proposition 8.7 (Properties of conditional expectation)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ σ-algebras and $X, Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$.

1) $\mathbb{E}(X \,|\, \{\emptyset, \Omega\}) = \mathbb{E}(X)$ (no information)
2) If X is independent of $\mathcal{G}$, then $\mathbb{E}(X \,|\, \mathcal{G}) = \mathbb{E}(X)$ (no relevant information).
3) If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X \,|\, \mathcal{G}) = X$ (full information).
   In particular, $\mathbb{E}(c \,|\, \mathcal{G}) = c \ \forall\, c \in \mathbb{R}$.
4) $\mathbb{E}(aX + bY \,|\, \mathcal{G}) = a\mathbb{E}(X \,|\, \mathcal{G}) + b\mathbb{E}(Y \,|\, \mathcal{G}) \ \forall\, a, b \in \mathbb{R}$ (linearity)
5) If $X \le Y$, then $\mathbb{E}(X \,|\, \mathcal{G}) \le \mathbb{E}(Y \,|\, \mathcal{G})$ (monotonicity).
6) $|\mathbb{E}(X \,|\, \mathcal{G})| \le \mathbb{E}(|X| \,|\, \mathcal{G})$ (triangle ineq.; special case of "conditional Jensen")
7) $\mathbb{E}(\mathbb{E}(X \,|\, \mathcal{H}) \,|\, \mathcal{G}) = \mathbb{E}(X \,|\, \mathcal{H}) = \mathbb{E}(\mathbb{E}(X \,|\, \mathcal{G}) \,|\, \mathcal{H})$ (*tower property* or *smoothing*).
   In particular, $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \,|\, \mathcal{G}))$ (*law of total expectation*).

In line with L. 8.6, P. 8.7 1) and 2), a coarser (finer) $\mathcal{G}$ averages over fewer (more) events and thus more (less) averaging takes place, so $\mathbb{E}(X \,|\, \mathcal{G})$ retains less (more) information of X, e.g. $\mathbb{E}(X \,|\, \mathcal{G}) \underset{\mathcal{G}=\{\emptyset,\Omega\}}{=} \mathbb{E}(X)$ (full averaging, no information about X retained except mean) and $\mathbb{E}(X \,|\, \mathcal{G}) \underset{X \in \mathcal{G}}{=} X$ (no averaging; full information retained). Tower property: coarser σ-algebra (least information) remains (smoothing).

Proof.

1) $\mathbb{E}(X)$ is constant, thus $\{\emptyset, \Omega\}$-measurable (see E. 3.3 1)). Furthermore, $\int_\emptyset \mathbb{E}(X)\,\mathrm{d}\mathbb{P} \underset{\text{L. 5.11 6)}}{=} 0 \underset{\text{L. 5.11 6)}}{=} \int_\emptyset X\,\mathrm{d}\mathbb{P}$ and $\int_\Omega \mathbb{E}(X)\,\mathrm{d}\mathbb{P} = \mathbb{E}(\mathbb{E}(X)) = \mathbb{E}(X) = \int_\Omega X\,\mathrm{d}\mathbb{P}$.

2) $\mathbb{E}(X)$ is constant, thus $\mathcal{G}$-measurable (by E. 3.3 1)). Furthermore,

$$\int_A \mathbb{E}(X)\,\mathrm{d}\mathbb{P} \underset{\text{lin.}}{=} \mathbb{E}(X)\mathbb{P}(A) \underset{\text{L. 5.3 5)}}{=} \mathbb{E}(X)\mathbb{E}(\mathbb{1}_A) \underset{\text{ind.}}{=} \mathbb{E}(X\mathbb{1}_A) \underset{\text{def.}}{=} \int_A X\,\mathrm{d}\mathbb{P},\ \forall\, A \in \mathcal{G}.$$

3) $X \in \mathcal{G}$ by assumption. Furthermore, $\int_A X\,\mathrm{d}\mathbb{P} = \int_A X\,\mathrm{d}\mathbb{P}\ \forall\, A \in \mathcal{G}$, so $\mathbb{E}(X\,|\,\mathcal{G}) = X$. In particular, the rv c is $\mathcal{G}$-measurable, so $\mathbb{E}(c\,|\,\mathcal{G}) = c$.

4) By definition, $\mathbb{E}(X\,|\,\mathcal{G})$, $\mathbb{E}(Y\,|\,\mathcal{G})$ are $\mathcal{G}$-measurable, so $a\mathbb{E}(X\,|\,\mathcal{G}) + b\mathbb{E}(Y\,|\,\mathcal{G})$ is. Furthermore, $\forall\, A \in \mathcal{G}$,

$$\begin{aligned}\int_A a\mathbb{E}(X\,|\,\mathcal{G}) + b\mathbb{E}(Y\,|\,\mathcal{G})\,\mathrm{d}\mathbb{P} &\underset{\text{lin.}}{=} a\int_A \mathbb{E}(X\,|\,\mathcal{G})\,\mathrm{d}\mathbb{P} + b\int_A \mathbb{E}(Y\,|\,\mathcal{G})\,\mathrm{d}\mathbb{P}\\ &\underset{\text{def.}}{=} a\int_A X\,\mathrm{d}\mathbb{P} + b\int_A Y\,\mathrm{d}\mathbb{P} \underset{\text{lin.}}{=} \int_A aX + bY\,\mathrm{d}\mathbb{P},\end{aligned}$$

so $\mathbb{E}(aX + bY\,|\,\mathcal{G}) = a\mathbb{E}(X\,|\,\mathcal{G}) + b\mathbb{E}(Y\,|\,\mathcal{G})$.

5) $\forall\, A \in \mathcal{G}$, $\int_A \mathbb{E}(X\,|\,\mathcal{G})\,\mathrm{d}\mathbb{P} \underset{\text{def.}}{=} \int_A X\,\mathrm{d}\mathbb{P} \overset{\text{ass.}}{\underset{\text{mon.}}{\leq}} \int_A Y\,\mathrm{d}\mathbb{P} \underset{\text{def.}}{=} \int_A \mathbb{E}(Y\,|\,\mathcal{G})\,\mathrm{d}\mathbb{P}$, so

$$\int_A \mathbb{E}(X\,|\,\mathcal{G}) - \mathbb{E}(Y\,|\,\mathcal{G})\,\mathrm{d}\mathbb{P} \underset{(*)}{\leq} 0.$$

$\forall\, \varepsilon > 0$, $A_\varepsilon := \{\mathbb{E}(X\,|\,\mathcal{G}) - \mathbb{E}(Y\,|\,\mathcal{G}) \geq \varepsilon\}$ is the preimage of $[\varepsilon, \infty)$ under a $\mathcal{G}$-measurable function, hence $A_\varepsilon \in \mathcal{G}$. Furthermore,

$$0 \underset{(*)}{\geq} \int_{A_\varepsilon} \mathbb{E}(X\,|\,\mathcal{G}) - \mathbb{E}(Y\,|\,\mathcal{G})\,\mathrm{d}\mathbb{P} \underset{\text{mon.}}{\geq} \int_{A_\varepsilon} \varepsilon\,\mathrm{d}\mathbb{P} \underset{\text{lin.}}{=} \varepsilon\mathbb{P}(A_\varepsilon),$$

so $\mathbb{P}(A_\varepsilon) \overset{!}{\underset{(**)}{=}} 0\ \forall\, \varepsilon > 0$. Therefore, $\mathbb{P}(\mathbb{E}(X\,|\,\mathcal{G}) > \mathbb{E}(Y\,|\,\mathcal{G})) = \mathbb{P}(\bigcup_{n=1}^\infty A_{1/n}) \leq \sum_{n=1}^\infty \mathbb{P}(A_{1/n}) \underset{(**)}{=} 0$, hence $\mathbb{E}(X\,|\,\mathcal{G}) \leq \mathbb{E}(Y\,|\,\mathcal{G})$.

6)
$$\begin{aligned} |\mathbb{E}(X\,|\,\mathcal{G})| &= |\mathbb{E}(X^+ - X^-\,|\,\mathcal{G})| \underset{\text{lin.}}{=} |\mathbb{E}(X^+\,|\,\mathcal{G}) - \mathbb{E}(X^-\,|\,\mathcal{G})| \\ &\underset{\Delta}{\leq} |\mathbb{E}(X^+\,|\,\mathcal{G})| + |\mathbb{E}(X^-\,|\,\mathcal{G})| \overset{X^-,X^+\geq 0}{\underset{5)}{=}} \mathbb{E}(X^+\,|\,\mathcal{G}) + \mathbb{E}(X^-\,|\,\mathcal{G}) \\ &\underset{\text{lin.}}{=} \mathbb{E}(X^+ + X^-\,|\,\mathcal{G}) = \mathbb{E}(|X|\,|\,\mathcal{G}). \end{aligned}$$

7) $\mathbb{E}(X\,|\,\mathcal{H})$ is $\mathcal{H}$-measurable $\underset{\mathcal{H}\subseteq\mathcal{G}}{\Rightarrow}$ $\mathbb{E}(X\,|\,\mathcal{H})$ is $\mathcal{G}$-measurable $\Rightarrow \mathbb{E}(\mathbb{E}(X\,|\,\mathcal{H})\,|\,\mathcal{G}) \underset{3)}{=} \mathbb{E}(X\,|\,\mathcal{H})$, which verifies the first equality. For the second, we have $\forall\, A \in \mathcal{H} \subseteq \mathcal{G}$ that $\int_A \mathbb{E}(\mathbb{E}(X\,|\,\mathcal{G})\,|\,\mathcal{H})\,\mathrm{d}\mathbb{P} \overset{A\in\mathcal{H}}{\underset{\text{def.}}{=}} \int_A \mathbb{E}(X\,|\,\mathcal{G})\,\mathrm{d}\mathbb{P} \overset{A\in\mathcal{G}}{\underset{\text{def.}}{=}} \int_A X\,\mathrm{d}\mathbb{P} \overset{A\in\mathcal{H}}{\underset{\text{def.}}{=}} \int_A \mathbb{E}(X\,|\,\mathcal{H})\,\mathrm{d}\mathbb{P}$.

In particular, $\mathbb{E}(X) \underset{1)}{=} \mathbb{E}(X\,|\,\{\emptyset,\Omega\}) \underset{\text{tower}}{=} \mathbb{E}(\mathbb{E}(X\,|\,\mathcal{G})\,|\,\{\emptyset,\Omega\}) \underset{1)}{=} \mathbb{E}(\mathbb{E}(X\,|\,\mathcal{G}))$.

□

Similar to the unconditional case, we can also define $\|X \,|\, \mathcal{G}\|_p := \mathbb{E}(|X|^p \,|\, \mathcal{G})^{1/p}$, $p \in (0,\infty]$, typically $p \in [1,\infty]$. One can then obtain the following inequalities.

Lemma 8.8 (Inequalities)

1) $\forall\, p,q \in [1,\infty] : \frac{1}{p} + \frac{1}{q} = 1$, we have $\|XY \,|\, \mathcal{G}\|_1 \le \|X \,|\, \mathcal{G}\|_p \cdot \|Y \,|\, \mathcal{G}\|_q$ (*conditional Hölder's inequality*); for $p = q = 2$, we obtain the *conditional Cauchy–Schwarz inequality*.
2) $\|X + Y \,|\, \mathcal{G}\|_p \le \|X \,|\, \mathcal{G}\|_p + \|Y \,|\, \mathcal{G}\|_p \ \forall\, X, Y \in L^p$, $p \in [1,\infty]$ (*conditional Minkowski's inequality*)
3) $\forall\, X, \varphi(X) \in L^1$, φ convex (concave), $\varphi(\mathbb{E}(X \,|\, \mathcal{G})) \le \mathbb{E}(\varphi(X) \,|\, \mathcal{G})$ ($\varphi(\mathbb{E}(X \,|\, \mathcal{G})) \ge \mathbb{E}(\varphi(X) \,|\, \mathcal{G})$) (*conditional Jensen's inequality (cJensen)*)

Proof. Similarly to the proofs of the respective unconditional inequalities. □

Proposition 8.9 (Contraction in L^p and continuity property)

1) If $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$, $p \in [1,\infty)$, then $\|\mathbb{E}(X \,|\, \mathcal{G})\|_p \le \|X\|_p$.
2) If $X_n \underset{n\to\infty}{\overset{L^p}{\to}} X$, then $\mathbb{E}(X_n \,|\, \mathcal{G}) \underset{n\to\infty}{\overset{L^p}{\to}} \mathbb{E}(X \,|\, \mathcal{G})$.

Proof.

1) $\|\mathbb{E}(X\,|\,\mathcal{G})\|_p \underset{\text{def.}}{=} \mathbb{E}(|\mathbb{E}(X\,|\,\mathcal{G})|^p)^{\frac{1}{p}} \underset{\text{cJensen}}{\leq} \mathbb{E}(\mathbb{E}(|X|^p\,|\,\mathcal{G}))^{\frac{1}{p}} \overset{\text{tot.}}{\underset{\text{exp.}}{=}} \mathbb{E}(|X|^p)^{\frac{1}{p}} = \|X\|_p.$

2) $\|\mathbb{E}(X_n\,|\,\mathcal{G}) - \mathbb{E}(X\,|\,\mathcal{G})\|_p \underset{\text{lin.}}{=} \|\mathbb{E}(X_n - X\,|\,\mathcal{G})\|_p \underset{1)}{\leq} \|X_n - X\|_p \overset{\text{ass.}}{\underset{n\to\infty}{\to}} 0.$ □

With slightly stronger assumptions (and here on probability spaces), we get conditional versions of already presented convergence theorems.

Theorem 8.10 (Conditional versions of convergence results)

1) If $(X_n)_{n\in\mathbb{N}} \subseteq L_+$ are integrable and $X_n \nearrow X$ pointwise, then $X \in L_+$ is integrable and $\mathbb{E}(X_n\,|\,\mathcal{G}) \underset{n\to\infty}{\nearrow} \mathbb{E}(X\,|\,\mathcal{G})$ a.s. (*conditional monotone convergence theorem (cMON)*).
2) If $(X_n)_{n\in\mathbb{N}} \subseteq L_+$ are integrable, then $\mathbb{E}(\liminf_{n\to\infty} X_n\,|\,\mathcal{G}) \leq \liminf_{n\to\infty} \mathbb{E}(X_n|\,\mathcal{G})$ a.s. (*conditional Fatou's lemma (cFatou)*).
3) If $X_n \overset{\text{a.s.}}{\underset{n\to\infty}{\to}} X$, $|X_n| \overset{\text{a.s.}}{\leq} Y\ \forall\, n \in \mathbb{N}$ for $Y \in L^1$, then $(X_n) \subseteq L^1$, $X \in L^1$ and $\lim_{n\to\infty} \mathbb{E}(X_n\,|\,\mathcal{G}) \overset{\text{a.s.}}{=} \mathbb{E}(X\,|\,\mathcal{G})$ (*conditional dominated convergence theorem (cDOM)*).

Proof.

1) By monotonicity, $\mathbb{E}(X_n \,|\, \mathcal{G}) \underset{n\to\infty}{\nearrow}$ pointwise $\Rightarrow$ Let $Y := \lim_{n\to\infty} \mathbb{E}(X_n \,|\, \mathcal{G})$ pointwise. By L. 3.12 2), Y is $\mathcal{G}$-measurable. Furthermore, $\forall\, A \in \mathcal{G}$,

$$\int_A Y \,\mathrm{d}\mathbb{P} \underset{\text{def.}}{=} \int_A \lim_{n\to\infty} \mathbb{E}(X_n \,|\, \mathcal{G}) \,\mathrm{d}\mathbb{P} \underset{\text{MON}}{=} \lim_{n\to\infty} \int_A \mathbb{E}(X_n \,|\, \mathcal{G}) \,\mathrm{d}\mathbb{P} \underset{\text{def.}}{=} \lim_{n\to\infty} \int_A X_n \,\mathrm{d}\mathbb{P}$$
$$\underset{\text{MON}}{=} \int_A X \,\mathrm{d}\mathbb{P}.$$

By definition of cond. exp. we thus have $\mathbb{E}(X \,|\, \mathcal{G}) = Y \underset{\text{def.}}{=} \lim_{n\to\infty} \mathbb{E}(X_n \,|\, \mathcal{G})$.

Note: By L. 8.3, $Y \in L^1$ and so is X, hence $\mathbb{E}(X|\mathcal{G})$ is well-defined.

2) Similarly as in L. 5.15, let $Y_n := \inf_{k\ge n} X_k$, $n \in \mathbb{N}$. Then $0 \le Y_n \le X_n$, $n \in \mathbb{N}$, so $(Y_n)_{n\in\mathbb{N}} \subseteq L_+$ are integrable, too. Furthermore, $Y_n \nearrow$ pointwise, so $Y := \lim_{n\to\infty} Y_n$ exists pointwise and we have, pointwise,

$$Y_n \nearrow Y = \lim_{n\to\infty} Y_n = \lim_{n\to\infty} \inf_{k\ge n} X_k \underset{\text{analysis}}{=} \liminf_{n\to\infty} X_n.$$

Therefore,

$$\mathbb{E}(\liminf_{n\to\infty} X_n \,|\, \mathcal{G}) = \mathbb{E}(Y \,|\, \mathcal{G}) \underset{\text{cMON}}{=} \lim_{n\to\infty} \mathbb{E}(Y_n \,|\, \mathcal{G}) = \liminf_{n\to\infty} \mathbb{E}(Y_n \,|\, \mathcal{G})$$
$$\underset{\text{def. } Y_n}{=} \liminf_{n\to\infty} \mathbb{E}(\inf_{k\ge n} X_k \,|\, \mathcal{G}) \underset{\text{mon.}}{\le} \liminf_{n\to\infty} \mathbb{E}(X_n \,|\, \mathcal{G}).$$

Note: We cannot work with $\lim_{n\to\infty}$ as it is unclear whether $\lim_{n\to\infty} \mathbb{E}(X_n \,|\, \mathcal{G})$ exists.

3) The first parts of the statement follow directly from DOM. We have left to show that $\lim_{n\to\infty} \mathbb{E}(X_n \,|\, \mathcal{G}) \overset{\text{a.s.}}{=} \mathbb{E}(X \,|\, \mathcal{G})$. Since

$$|\mathbb{E}(X_n \,|\, \mathcal{G}) - \mathbb{E}(X \,|\, \mathcal{G})| \underset{\text{lin.}}{=} |\mathbb{E}(X_n - X \,|\, \mathcal{G})| \underset{\Delta}{\leq} \mathbb{E}(|X_n - X| \,|\, \mathcal{G})$$
$$\underset{\text{mon.}}{\leq} \mathbb{E}(\sup_{k\geq n} |X_k - X| \,|\, \mathcal{G}),$$

it suffices to show that $\lim\limits_{n\to\infty} \mathbb{E}(Z_n \,|\, \mathcal{G}) \overset{\text{a.s.}}{=} 0$ for $Z_n := \sup_{k\geq n} |X_k - X|$, $n \in \mathbb{N}$.

- By ass., $Z_n \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} 0$. And we have that

$$0 \leq Z_n \underset{\Delta}{\overset{\text{a.s.}}{\leq}} \sup_{n\in\mathbb{N}} |X_n| + |X| \underset{\text{ass.}}{\overset{\text{a.s.}}{\leq}} Y + Y = 2Y \in L^1, \quad \forall\, n \in \mathbb{N}.$$

  By DOM, $\lim_{n\to\infty} \mathbb{E}(Z_n) = \mathbb{E}(0) = 0$.
- $0 \leq \mathbb{E}(Z_n \,|\, \mathcal{G}) \searrow \;\Rightarrow Z := \lim_{n\to\infty} \mathbb{E}(Z_n \,|\, \mathcal{G})$ exists and is a rv by L. 3.12 2).
- Since $0 \leq Z \leq \mathbb{E}(Z_n \,|\, G) \underset{\text{L. 8.3}}{\in} L^1 \;\forall\, n \in \mathbb{N}$, we have that $Z \in L^1$ and thus

$$0 \underset{\text{mon.}}{\leq} \mathbb{E}(Z) \underset{\text{mon.}}{\leq} \mathbb{E}(\mathbb{E}(Z_n \,|\, \mathcal{G})) \underset{\text{exp.}}{\overset{\text{tot.}}{=}} \mathbb{E}(Z_n) \underset{n\to\infty}{\to} 0.$$

  Therefore, $\lim\limits_{n\to\infty} \mathbb{E}(Z_n \,|\, \mathcal{G}) \underset{\text{def.}}{=} Z \overset{\text{a.s.}}{=} 0$. □

Proposition 8.11 (Independence from a σ-algebra)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ σ-algebras and $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. If $\mathcal{H}$ is independent of $\sigma(\sigma(X), \mathcal{G})$, then $\mathbb{E}(X \mid \sigma(\mathcal{G}, \mathcal{H})) = \mathbb{E}(X \mid \mathcal{G})$.

Proof.

- By linearity, we can assume wlog $X \overset{\text{a.s.}}{\geq} 0$. Let Y be a version of $\mathbb{E}(X \mid \mathcal{G})$. Then Y is $\mathcal{G}$-measurable and $Y \underset{\text{mon.}}{\overset{\text{a.s.}}{\geq}} 0$.
- By L. 5.8 6), $\mu(A) := \mathbb{E}(Y \mathbb{1}_A)$ and $\nu(A) := \mathbb{E}(X \mathbb{1}_A)$ are measures on $(\Omega, \mathcal{F})$. Since $\mu(\Omega) = \nu(\Omega) \underset{\text{exp.}}{\overset{\text{tot.}}{=}} \mathbb{E}(X) \underset{\text{ass.}}{<} \infty$, they are finite, thus σ-finite.
- $\forall\, G \in \mathcal{G},\ H \in \mathcal{H}$,

$$\mu(G \cap H) = \mathbb{E}(Y \mathbb{1}_{G \cap H}) = \mathbb{E}(Y \mathbb{1}_G \mathbb{1}_H) \underset{\mathcal{H}\text{ ind. of }\mathcal{G}}{\overset{Y \in \mathcal{G}}{=}} \mathbb{E}(Y \mathbb{1}_G)\mathbb{E}(\mathbb{1}_H) \underset{\text{simple}}{=} \mathbb{E}(Y \mathbb{1}_G)\mathbb{P}(H)$$

$$\underset{G \in \mathcal{G}}{\overset{Y = \mathbb{E}(X \mid \mathcal{G})}{=}} \mathbb{E}(X \mathbb{1}_G)\mathbb{P}(H) \underset{\text{using }\mathcal{H}\text{ ind. of }\sigma(\sigma(X), \mathcal{G})}{\overset{\text{similarly backwards}}{=}} \mathbb{E}(X \mathbb{1}_{G \cap H}) = \nu(G \cap H).$$

  Therefore, μ, ν are σ-finite measures that coincide on the π-system $\{G \cap H : G \in \mathcal{G}, H \in \mathcal{H}\} \underset{\text{P. 2.28}}{\Rightarrow} \mu, \nu$ coincide on $\sigma(\{G \cap H : G \in \mathcal{G}, H \in \mathcal{H}\}) \underset{\text{L. 2.13}}{=} \sigma(\mathcal{G}, \mathcal{H})$.
- We have thus verified that $Y = \mathbb{E}(X \mid \mathcal{G})$ is a version of $\mathbb{E}(X \mid \sigma(\mathcal{G}, \mathcal{H}))$, hence $\mathbb{E}(X \mid \mathcal{G}) = \mathbb{E}(X \mid \sigma(\mathcal{G}, \mathcal{H}))$ a.s. □

- For $\mathcal{G} = \{\emptyset, \Omega\}$ we obtain that if $\mathcal{H}$ is independent of $\sigma(\sigma(X), \mathcal{G}) = \sigma(\sigma(X)) = \sigma(X)$, then $\mathbb{E}(X \,|\, \mathcal{H}) = \mathbb{E}(X \,|\, \sigma(\mathcal{G}, \mathcal{H})) \underset{\text{P. 8.11}}{=} \mathbb{E}(X \,|\, \mathcal{G}) \underset{\text{P. 8.7 1)}}{=} \mathbb{E}(X)$, which we already know from P. 8.7 2).
- We already mentioned that $\mathbb{E}(X \,|\, \mathcal{G})$ is often found by a guess and verification of the defining properties i)–ii). The following result can simplify the verification of the defining properties of conditional expectation.

Lemma 8.12 (Generator)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ a σ-algebra on Ω and $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. If $\mathcal{A} \subseteq \mathcal{F}$ is a π-system such that $\sigma(\mathcal{A}) = \mathcal{G}$ and $\exists\, (A_i)_{i \in \mathbb{N}} \subseteq \mathcal{A}$ with $\bigcup_{i=1}^{\infty} A_i = \Omega$, then any $Y : \Omega \to \mathbb{R}$ such that

i) $Y \in \mathcal{G}$; and

ii) $\mathbb{E}(Y \mathbb{1}_A) = \mathbb{E}(X \mathbb{1}_A)\ \forall\, A \in \mathcal{A}$,

is a version of $\mathbb{E}(X \,|\, \mathcal{G})$.

Proof.

- $\mathbb{E}(|X \mathbb{1}_A|) \underset{\text{mon.}}{\leq} \mathbb{E}(|X|) \underset{\text{ass.}}{<} \infty\ \forall\, A \in \mathcal{F} \Rightarrow X \mathbb{1}_A$ is integrable $\forall\, A \in \mathcal{F}$, in particular $\forall\, A \in \mathcal{A} \subseteq \mathcal{F}$. By ii), $Y \mathbb{1}_A$ is thus integrable $\forall\, A \in \mathcal{A}$.

- By integrability of $X\mathbb{1}_A$, $Y\mathbb{1}_A$ $\forall\, A \in \mathcal{A}$, we have that

$$\mathbb{E}(Y\mathbb{1}_A) = \mathbb{E}(X\mathbb{1}_A) \quad \forall\, A \in \mathcal{A}$$

$$\underset{\text{lin.}}{\Leftrightarrow} \mathbb{E}(Y^+\mathbb{1}_A) - \mathbb{E}(Y^-\mathbb{1}_A) = \mathbb{E}(X^+\mathbb{1}_A) - \mathbb{E}(X^-\mathbb{1}_A) \quad \forall\, A \in \mathcal{A}$$

$$\underset{\text{lin.}}{\Leftrightarrow} \mu(A) := \mathbb{E}((Y^+ + X^-)\mathbb{1}_A) = \mathbb{E}(Y^+\mathbb{1}_A) + \mathbb{E}(X^-\mathbb{1}_A)$$

$$= \mathbb{E}(X^+\mathbb{1}_A) + \mathbb{E}(Y^-\mathbb{1}_A) = \mathbb{E}((X^+ + Y^-)\mathbb{1}_A) =: \nu(A) \quad \forall\, A \in \mathcal{A}.$$

- By L. 5.8 6) and since $Y \underset{\text{i)}}{\in} \mathcal{G}$, we know that μ, ν are measures on $\mathcal{G}$. By $\mu|_{\mathcal{A}} = \nu|_{\mathcal{A}}$ and integrability of X, Y, thus X^+, X^-, Y^+, Y^-, on $\mathcal{A} \subseteq \sigma(\mathcal{A}) = \mathcal{G}$, we obtain $\mu(A_i) \underset{\text{on } \mathcal{A}}{=} \nu(A_i) \overset{\text{integrability}}{\underset{\text{of } X,\, Y \text{ on } \mathcal{A}}{<}} \infty \;\forall\, i \in \mathbb{N}$, where, by ass., $(A_i)_{i\in\mathbb{N}} \subseteq \mathcal{A}$ with $\bigcup_{i=1}^{\infty} A_i = \Omega$. Hence μ, ν are σ-finite measures on $\mathcal{G}$ with $\mu|_{\mathcal{A}} = \nu|_{\mathcal{A}}$.
- We thus obtain that $\mu|_{\mathcal{G}} = \mu|_{\sigma(\mathcal{A})} \underset{\text{P. 2.28}}{=} \nu|_{\sigma(\mathcal{A})} = \nu|_{\mathcal{G}}$, which implies D. 8.2 ii) and thus concludes the proof. □

If $X \in \mathcal{G}$, we already know that $\mathbb{E}(X \,|\, \mathcal{G}) \overset{\text{a.s.}}{\underset{\text{P. 8.7 3)}}{=}} X = X \cdot 1 \overset{\text{a.s.}}{\underset{\text{P. 8.7 3)}}{=}} X\mathbb{E}(1 \,|\, \mathcal{G})$. The following is a generalization.

Proposition 8.13 (Taking out what is known (TOWIK); product rule)
If XY and Y are integrable and $Y \in \mathcal{G}$, then $\mathbb{E}(XY \,|\, \mathcal{G}) = Y\mathbb{E}(X \,|\, \mathcal{G})$.

Proof. $Y\mathbb{E}(X \,|\, \mathcal{G}) \overset{\text{ass.}}{\underset{\text{def.}}{\in}} \mathcal{G}$. We have left to show that $\int_A Y\mathbb{E}(X \,|\, \mathcal{G}) \,\mathrm{d}\mathbb{P} = \int_A XY \,\mathrm{d}\mathbb{P}$ $\forall\, A \in \mathcal{G}$. We apply the standard argument to Y.

1) $Y = \mathbb{1}_B$, $B \in \mathcal{G} \Rightarrow \int_A \mathbb{1}_B \mathbb{E}(X \,|\, \mathcal{G}) \,\mathrm{d}\mathbb{P} = \int_{A\cap B} \mathbb{E}(X \,|\, \mathcal{G}) \,\mathrm{d}\mathbb{P} \overset{A\cap B \in \mathcal{G}}{\underset{\text{def.}}{=}} \int_{A\cap B} X \,\mathrm{d}\mathbb{P}$ $= \int_A \mathbb{1}_B X \,\mathrm{d}\mathbb{P}$. By linearity, the argument holds for all simple Y.

2) $X, Y \geq 0 \underset{\text{L. 5.4}}{\Rightarrow} \exists$ simple $(Y_n)_{n\in\mathbb{N}}$ s.t. $Y_n \nearrow Y$. Then, $\forall\, A \in \mathcal{G}$, we have $Y_n \mathbb{E}(X \,|\, \mathcal{G})\mathbb{1}_A \nearrow Y\mathbb{E}(X \,|\, \mathcal{G})\mathbb{1}_A$ and $XY_n\mathbb{1}_A \nearrow XY\mathbb{1}_A$ pointwise, so that

$$\int_A Y\mathbb{E}(X \,|\, \mathcal{G}) \,\mathrm{d}\mathbb{P} \underset{\text{MON}}{=} \lim_{n\to\infty} \int_A Y_n \mathbb{E}(X \,|\, \mathcal{G}) \,\mathrm{d}\mathbb{P} \overset{\text{simple}}{\underset{1)}{=}} \lim_{n\to\infty} \int_A XY_n \,\mathrm{d}\mathbb{P} \underset{\text{MON}}{=} \int_A XY \,\mathrm{d}\mathbb{P}.$$

3) For general X, Y, use $X = X^+ - X^-$, $Y = Y^+ - Y^-$, linearity and 2). □

Corollary 8.14 (Independence)
If X, Y, XY are integrable and $Y \in \mathcal{F}$ is independent of $\sigma(\sigma(X), \mathcal{G})$, then $\mathbb{E}(XY \,|\, \mathcal{G}) = \mathbb{E}(Y)\mathbb{E}(X \,|\, \mathcal{G})$.

Proof. $\mathbb{E}(XY \,|\, \mathcal{G}) \overset{\text{tower}}{\underset{\mathcal{G}\subseteq\sigma(\sigma(X),\mathcal{G})}{=}} \mathbb{E}(\mathbb{E}(XY \,|\, \sigma(\sigma(X), \mathcal{G})) \,|\, \mathcal{G}) \underset{\text{TOWIK}}{=} \mathbb{E}(X \underbrace{\mathbb{E}(Y \,|\, \sigma(\sigma(X), \mathcal{G}))}_{\overset{\text{ind.}}{\underset{\text{P. 8.7 2)}}{=}} \mathbb{E}(Y)} \,|\, \mathcal{G})$ $\underset{\text{TOWIK}}{=} \mathbb{E}(Y)\mathbb{E}(X \,|\, \mathcal{G})$. □

Theorem 8.15 (Factorisation, Doob-Dynkin lemma)
For $\Omega \neq \emptyset$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\Omega', \mathcal{F}')$ a measurable space and $Z : \Omega \to \Omega'$. Then $Y : \Omega \to \bar{\mathbb{R}}$ is $(\sigma(Z), \mathcal{B}(\bar{\mathbb{R}}))$-measurable iff $\exists$ a $(\mathcal{F}', \mathcal{B}(\bar{\mathbb{R}}))$-measurable $h : \Omega' \to \bar{\mathbb{R}}$ such that $Y = h(Z)$.

Proof.

"$\Leftarrow$": Z is $(\sigma(Z), \mathcal{F}')$-measurable, h is $(\mathcal{F}', \mathcal{B}(\bar{\mathbb{R}}))$-measurable $\underset{\text{P. 3.6}}{\Rightarrow}$ ✓

"$\Rightarrow$": Let Y be $(\sigma(Z), \mathcal{B}(\bar{\mathbb{R}}))$-measurable. We apply the standard argument.

1) If $Y = \sum_{i=1}^n y_i \mathbb{1}_{A_i}$ for $n \in \mathbb{N}$ and $\{A_i\}_{i=1}^n \subseteq \sigma(Z)$ then $A_i = Z^{-1}(A_i')$ for $A_i' \in \mathcal{F}'$, $i = 1, \dots, n$. Therefore, $Y(\omega) = \sum_{i=1}^n y_i \mathbb{1}_{A_i}(\omega) = \sum_{i=1}^n y_i \mathbb{1}_{Z^{-1}(A_i')}(\omega) = \sum_{i=1}^n y_i \mathbb{1}_{A_i'}(Z(\omega)) = h(Z(\omega))$ for the $(\mathcal{F}', \mathcal{B}(\bar{\mathbb{R}}))$-measurable $h(z) = \sum_{i=1}^n y_i \mathbb{1}_{A_i'}(z)$.

2) $Y \geq 0 \underset{\text{L. 5.4}}{\Rightarrow} \exists$ simple $(Y_n)_{n\in\mathbb{N}}$ s.t. $Y_n \nearrow Y \underset{1)}{\Rightarrow} \exists$ $(\mathcal{F}', \mathcal{B}(\bar{\mathbb{R}}))$-measurable $(h_n)_{n\in\mathbb{N}} : Y_n = h_n(Z)$. The claim follows with $h(z) := \sup_{n\in\mathbb{N}} h_n(z) \underset{h_n \nearrow}{=} \lim_{n\to\infty} h_n(z)$, which is $(\mathcal{F}', \mathcal{B}(\bar{\mathbb{R}}))$-measurable by L. 3.12 1) or 2).

3) General $Y \underset{2)}{\Rightarrow} \exists$ $(\mathcal{F}', \mathcal{B}(\bar{\mathbb{R}}))$-measurable h^+, h^- s.t. $Y^+ = h^+(Z)$, $Y^- = h^-(Z)$. So $Y = h(Z)$ for the $(\mathcal{F}', \mathcal{B}(\bar{\mathbb{R}}))$-measurable $h := h^+ - h^-$. □

$\mathbb{E}(X|Z = z) \underset{\text{fact.}}{:=} h(z)$ is the *conditional expectation of X given $Z = z$*; often $\Omega' = \mathbb{R}^d$

Alternatively, and specifically for $\mathbb{E}(X \mid \boldsymbol{Z}) = h(\boldsymbol{Z})$, one can argue as follows.

Theorem 8.16 (Factorisation, Doob-Dynkin lemma for cond. expectations)
For $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and a d-dimensional random vector $\boldsymbol{Z}$, there exists a measurable $h : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}(X \mid \boldsymbol{Z}) = h(\boldsymbol{Z})$.

Proof.

- Wlog assume $X \geq 0$; otherwise consider X^-, X^+. Then $\mathbb{E}(X) = \mathbb{E}(|X|) \underset{\text{ass.}}{<} \infty$.
- By L. 5.8 6), $\nu(B) := \int_{\boldsymbol{Z}^{-1}(B)} X \,\mathrm{d}\mathbb{P}$, $B \in \mathcal{B}(\mathbb{R}^d)$, is a measure. Furthermore, $\nu(\mathbb{R}^d) = \int_\Omega X \,\mathrm{d}\mathbb{P} = \mathbb{E}(X) < \infty$, so ν is a finite measure on $\mathbb{R}^d$. And $\forall\, \mathbb{P}_{\boldsymbol{Z}}$-null sets N, we have $\mathbb{P}(\boldsymbol{Z}^{-1}(N)) = \mathbb{P}_{\boldsymbol{Z}}(N) = 0 \underset{\text{L. 5.11 6)}}{\Rightarrow} \nu(N) = 0 \Rightarrow \nu \ll \mathbb{P}_{\boldsymbol{Z}}$.
- RN $\Rightarrow \exists$ a $\mathbb{P}_{\boldsymbol{Z}}$-a.s. unique integrable $h : \mathbb{R}^d \to [0, \infty)$ such that

$$\int_{\boldsymbol{Z}^{-1}(B)} X \,\mathrm{d}\mathbb{P} \underset{\text{def.}}{=} \nu(B) \underset{\text{RN}}{=} \int_B h(\boldsymbol{z}) \,\mathrm{d}\mathbb{P}_{\boldsymbol{Z}}(\boldsymbol{z}) \underset{\text{T. 5.21}}{=} \int_{\boldsymbol{Z}^{-1}(B)} h(\boldsymbol{Z}) \,\mathrm{d}\mathbb{P} \quad \forall\, B \in \mathcal{B}(\mathbb{R}^d). \quad (9)$$

- As an integrable function, h is measurable by definition, so $\forall\, B \in \mathcal{B}(\mathbb{R})$, we have $(h(\boldsymbol{Z}))^{-1}(B) = \boldsymbol{Z}^{-1}(h^{-1}(B)) \underset{h^{-1}(B) \in \mathcal{B}(\mathbb{R}^d)}{\in} \boldsymbol{Z}^{-1}(\mathcal{B}(\mathbb{R}^d)) = \sigma(\boldsymbol{Z})$ and thus $h(\boldsymbol{Z}) \in \sigma(\boldsymbol{Z})$, so $h(\boldsymbol{Z})$ is $\sigma(\boldsymbol{Z})$-measurable and thus D. 8.2 i) holds.
- By (9), $h(\boldsymbol{Z})$ fulfills D. 8.2 ii), so $\mathbb{E}(X \mid \boldsymbol{Z}) \underset{\text{def.}}{=} \mathbb{E}(X \mid \sigma(\boldsymbol{Z})) = h(\boldsymbol{Z})$ a.s. □

Factorisation is often applied to $X = g(\boldsymbol{X}, \boldsymbol{Z})$ where $\boldsymbol{X}, \boldsymbol{Z}$ are independent.

Theorem 8.17 (Factorisation under independence)
If $\boldsymbol{X}, \boldsymbol{Z}$ are independent $d_{\boldsymbol{x}}$- and $d_{\boldsymbol{z}}$-dimensional random vectors, $g : \mathbb{R}^{d_{\boldsymbol{x}}+d_{\boldsymbol{z}}} \to \mathbb{R}$ a measurable function such that $g(\boldsymbol{X}, \boldsymbol{Z}) \in L_+$ or $g(\boldsymbol{X}, \boldsymbol{Z}) \in L^1$, and $h(\boldsymbol{z}) := \mathbb{E}(g(\boldsymbol{X}, \boldsymbol{z}))$, then $\mathbb{E}(g(\boldsymbol{X}, \boldsymbol{Z}) \,|\, \boldsymbol{Z}) = h(\boldsymbol{Z})$ a.s.

Proof. We check the two defining properties of $\mathbb{E}(g(\boldsymbol{X}, \boldsymbol{Z}) \,|\, \boldsymbol{Z})$ for $h(\boldsymbol{Z})$:

i) By def., integrable functions are measurable, so $(h(\boldsymbol{Z}))^{-1}(B) = \boldsymbol{Z}^{-1}(h^{-1}(B)) \underset{h^{-1}(B)\in\mathcal{B}(\mathbb{R}^{d_{\boldsymbol{z}}})}{\in} \boldsymbol{Z}^{-1}(\mathcal{B}(\mathbb{R}^{d_{\boldsymbol{z}}})) = \sigma(\boldsymbol{Z}) \ \forall\, B \in \mathcal{B}(\mathbb{R})$. Therefore $h(\boldsymbol{Z}) \in \sigma(\boldsymbol{Z})$.

ii) Consider $A \in \sigma(\boldsymbol{Z})$, that is $A = \boldsymbol{Z}^{-1}(B)$ for some $B \in \mathcal{B}(\mathbb{R}^{d_{\boldsymbol{z}}})$. Note that $\mathbb{1}_A(\omega) = \mathbb{1}_B(\boldsymbol{Z}(\omega)) \ \forall\, \omega \in \Omega$. If $g(\boldsymbol{X}, \boldsymbol{Z}) \in L_+$ ($g(\boldsymbol{X}, \boldsymbol{Z}) \in L^1$), apply Tonelli's (Fubini's) theorem to see that

$$\int_A g(\boldsymbol{X}, \boldsymbol{Z}) \,\mathrm{d}\mathbb{P} = \int_\Omega g(\boldsymbol{X}, \boldsymbol{Z}) \mathbb{1}_A \,\mathrm{d}\mathbb{P} = \int_\Omega g(\boldsymbol{X}, \boldsymbol{Z}) \mathbb{1}_B(\boldsymbol{Z}) \,\mathrm{d}\mathbb{P}$$

$$\underset{\text{T. 5.21}}{=} \int_{\mathbb{R}^{d_{\boldsymbol{x}}+d_{\boldsymbol{z}}}} g(\boldsymbol{x}, \boldsymbol{z}) \mathbb{1}_B(\boldsymbol{z}) \,\mathrm{d}F_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{z})$$

$$\underset{\text{ind.}}{\overset{\text{Tonelli/Fubini}}{=}} \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} \int_{\mathbb{R}^{d_{\boldsymbol{x}}}} g(\boldsymbol{x}, \boldsymbol{z}) \mathbb{1}_B(\boldsymbol{z}) \,\mathrm{d}F_{\boldsymbol{X}}(\boldsymbol{x}) \,\mathrm{d}F_{\boldsymbol{Z}}(\boldsymbol{z})$$

$$\underset{\text{lin.}}{=} \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} \mathbb{1}_B(\boldsymbol{z}) \underbrace{\int_{\mathbb{R}^{d_{\boldsymbol{x}}}} g(\boldsymbol{x},\boldsymbol{z}) \,\mathrm{d}F_{\boldsymbol{X}}(\boldsymbol{x})}_{=\mathbb{E}(g(\boldsymbol{X},\boldsymbol{z}))=h(\boldsymbol{z})} \mathrm{d}F_{\boldsymbol{Z}}(\boldsymbol{z})$$

$$= \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} \mathbb{1}_B(\boldsymbol{z}) h(\boldsymbol{z}) \,\mathrm{d}F_{\boldsymbol{Z}}(\boldsymbol{z}) \underset{\text{T. 5.21}}{=} \int_\Omega \mathbb{1}_B(\boldsymbol{Z}) h(\boldsymbol{Z}) \,\mathrm{d}\mathbb{P}$$

$$= \int_\Omega \mathbb{1}_A h(\boldsymbol{Z}) \,\mathrm{d}\mathbb{P} = \int_A h(\boldsymbol{Z}) \,\mathrm{d}\mathbb{P}. \qquad \square$$

- Factorisation under independence allows us to calculate $\mathbb{E}(g(\boldsymbol{X},\boldsymbol{Z}) \,|\, \boldsymbol{Z})$ by first calculating $h(\boldsymbol{z}) = \mathbb{E}(g(\boldsymbol{X},\boldsymbol{z}))$ (treating $\boldsymbol{Z}$ as if $\boldsymbol{Z} = \boldsymbol{z}$) and then returning $h(\boldsymbol{Z})$; see P. 8.21 for an application.
- **Note:** $\mathbb{E}(g(\boldsymbol{X},\boldsymbol{Z}) \,|\, \boldsymbol{Z})$ is a rv, whereas $\mathbb{E}(g(\boldsymbol{X},\boldsymbol{Z}) \,|\, \boldsymbol{Z} = \boldsymbol{z}) = h(\boldsymbol{z})$ is a value.
- In statistics,

$$\mathbb{E}(g(\boldsymbol{X},\boldsymbol{Z}) \,|\, \boldsymbol{Z} = \boldsymbol{z})$$

is known as the *regression function* of $g(\boldsymbol{X},\boldsymbol{Z})$ on $\boldsymbol{Z}$ at $\boldsymbol{z}$. If $g(\boldsymbol{X},\boldsymbol{Z}) \in L^2$, it is the best L^2-approximation of $g(\boldsymbol{X},\boldsymbol{Z})$ given one has observed $\boldsymbol{Z} = \boldsymbol{z}$ (see later).

8.3 Applications

We now present various applications of conditional expectations. We start with (regular) conditional dfs.

Corollary 8.18 (Central conditional distribution formula)

If $(\boldsymbol{X}, \boldsymbol{Z}) \sim F$ for a $(d_{\boldsymbol{x}} + d_{\boldsymbol{z}})$-dimensional df F then

$$F(\boldsymbol{x}, \boldsymbol{z}) = \int_{(-\infty, \boldsymbol{z}]} F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \,|\, \tilde{\boldsymbol{z}}) \,\mathrm{d}F_{\boldsymbol{Z}}(\tilde{\boldsymbol{z}}), \quad (\boldsymbol{x}, \boldsymbol{z}) \in \mathbb{R}^{d_{\boldsymbol{x}} + d_{\boldsymbol{z}}}. \tag{10}$$

Proof. With $h(\tilde{\boldsymbol{z}}) = \mathbb{E}(\mathbb{1}_{\{\boldsymbol{X} \le \boldsymbol{x}\}} \,|\, \boldsymbol{Z} = \tilde{\boldsymbol{z}}) = \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x} \,|\, \boldsymbol{Z} = \tilde{\boldsymbol{z}}) \underset{\text{def.}}{=} F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \,|\, \tilde{\boldsymbol{z}})$, we have

$$\begin{aligned}
F(\boldsymbol{x}, \boldsymbol{z}) &= \mathbb{E}(\mathbb{1}_{\{\boldsymbol{X} \le \boldsymbol{x}, \boldsymbol{Z} \le \boldsymbol{z}\}}) = \mathbb{E}(\mathbb{1}_{\{\boldsymbol{X} \le \boldsymbol{x}\}} \mathbb{1}_{\{\boldsymbol{Z} \le \boldsymbol{z}\}}) \overset{\text{tot.}}{\underset{\text{exp.}}{=}} \mathbb{E}_{\boldsymbol{Z}}(\mathbb{E}_{\boldsymbol{X}}(\mathbb{1}_{\{\boldsymbol{X} \le \boldsymbol{x}\}} \mathbb{1}_{\{\boldsymbol{Z} \le \boldsymbol{z}\}} \,|\, \boldsymbol{Z})) \\
&\underset{\text{TOWIK}}{=} \mathbb{E}_{\boldsymbol{Z}}(\mathbb{1}_{\{\boldsymbol{Z} \le \boldsymbol{z}\}} \mathbb{E}_{\boldsymbol{X}}(\mathbb{1}_{\{\boldsymbol{X} \le \boldsymbol{x}\}} \,|\, \boldsymbol{Z})) \underset{\text{def.}}{=} \mathbb{E}_{\boldsymbol{Z}}(\mathbb{1}_{\{\boldsymbol{Z} \le \boldsymbol{z}\}} h(\boldsymbol{Z})) \\
&\overset{\text{change of}}{\underset{\text{variables}}{=}} \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} \mathbb{1}_{\{\tilde{\boldsymbol{z}} \le \boldsymbol{z}\}} h(\tilde{\boldsymbol{z}}) \,\mathrm{d}F_{\boldsymbol{Z}}(\tilde{\boldsymbol{z}}) = \int_{(-\infty, \boldsymbol{z}]} h(\tilde{\boldsymbol{z}}) \,\mathrm{d}F_{\boldsymbol{Z}}(\tilde{\boldsymbol{z}}) \\
&\overset{\text{as}}{\underset{\text{shown}}{=}} \int_{(-\infty, \boldsymbol{z}]} F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \,|\, \tilde{\boldsymbol{z}}) \,\mathrm{d}F_{\boldsymbol{Z}}(\tilde{\boldsymbol{z}}).
\end{aligned}$$

□

Important special cases:

1) If F has continuous second partial derivatives, then $F(\boldsymbol{x},\boldsymbol{z}) = \int_{(-\infty,\boldsymbol{z}]} F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}\,|\,\tilde{\boldsymbol{z}}) f_{\boldsymbol{Z}}(\tilde{\boldsymbol{z}})\,\mathrm{d}\tilde{\boldsymbol{z}}$, so F has density

$$f(\boldsymbol{x},\boldsymbol{z}) = \frac{\partial^2}{\partial \boldsymbol{z}\partial \boldsymbol{x}} F(\boldsymbol{x},\boldsymbol{z}) \overset{\text{Schwarz}}{\underset{\text{Leibniz}}{=}} \frac{\partial}{\partial \boldsymbol{x}} F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}\,|\,\boldsymbol{z}) f_{\boldsymbol{Z}}(\boldsymbol{z}) \underset{\text{def.}}{=} f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}\,|\,\boldsymbol{z}) f_{\boldsymbol{Z}}(\boldsymbol{z}),$$

for $f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}\,|\,\boldsymbol{z}) := \frac{\partial}{\partial \boldsymbol{x}} F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}\,|\,\boldsymbol{z})$. For $f_{\boldsymbol{Z}}(\boldsymbol{z}) > 0$, we can thus recover the classical

$$f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}\,|\,\boldsymbol{z}) = \frac{f(\boldsymbol{x},\boldsymbol{z})}{f_{\boldsymbol{Z}}(\boldsymbol{z})} \tag{11}$$

and

$$F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}\,|\,\boldsymbol{z}) \underset{\text{fund. thm.}}{=} \int_{(-\infty,\boldsymbol{x}]} \frac{\partial}{\partial \tilde{\boldsymbol{x}}} F_{\boldsymbol{X}|\boldsymbol{Z}}(\tilde{\boldsymbol{x}}\,|\,\boldsymbol{z})\,\mathrm{d}\tilde{\boldsymbol{x}} \overset{\text{see}}{\underset{\text{above}}{=}} \int_{(-\infty,\boldsymbol{x}]} f_{\boldsymbol{X}|\boldsymbol{Z}}(\tilde{\boldsymbol{x}}\,|\,\boldsymbol{z})\,\mathrm{d}\tilde{\boldsymbol{x}}. \tag{12}$$

- Similarly for pmfs f, where (11) holds for pmfs and (12) in terms of sums.
- (11) resembles "$\mathbb{P}(A\,|\,B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$", but in general $f_{\boldsymbol{Z}}(\boldsymbol{z}) \neq \mathbb{P}(\boldsymbol{Z}=\boldsymbol{z}) \overset{\text{abs.}}{\underset{\text{cont.}}{=}} 0$.
- $f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}\,|\,\boldsymbol{z}) \overset{(11)}{\underset{\boldsymbol{X},\,\boldsymbol{Z}\text{ ind.}}{=}} \frac{f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{Z}}(\boldsymbol{z})}{f_{\boldsymbol{Z}}(\boldsymbol{z})} = f_{\boldsymbol{X}}(\boldsymbol{x})$, which aligns with intuition.

2) If $\boldsymbol{X}$ has pmf $f_{\boldsymbol{X}}$ and $F_{\boldsymbol{Z}}$ has density $f_{\boldsymbol{Z}}$, then, $\forall\, \boldsymbol{x} \in \operatorname{supp}(f_{\boldsymbol{X}}) = \{\boldsymbol{x} \in \mathbb{R}^{d_{\boldsymbol{x}}} : f_{\boldsymbol{X}}(\boldsymbol{x}) > 0\}$,

$$\begin{aligned}\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} \le \boldsymbol{z}) &= F(\boldsymbol{x}, \boldsymbol{z}) - F(\boldsymbol{x}-, \boldsymbol{z}) \\ &\underset{\text{lin.}}{\overset{\text{C. 8.18}}{=}} \int_{(-\infty, \boldsymbol{z}]} \big(F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \,|\, \tilde{\boldsymbol{z}}) - F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}- \,|\, \tilde{\boldsymbol{z}})\big) f_{\boldsymbol{Z}}(\tilde{\boldsymbol{z}}) \,\mathrm{d}\tilde{\boldsymbol{z}} \\ &\underset{\text{def.}}{=} \int_{(-\infty, \boldsymbol{z}]} \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Z} = \tilde{\boldsymbol{z}})\, f_{\boldsymbol{Z}}(\tilde{\boldsymbol{z}}) \,\mathrm{d}\tilde{\boldsymbol{z}}. \end{aligned} \tag{13}$$

Defining $f(\boldsymbol{x}, \boldsymbol{z}) := \frac{\partial}{\partial \boldsymbol{z}} \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} \le \boldsymbol{z})$ and $f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \,|\, \boldsymbol{z}) := \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Z} = \boldsymbol{z})$, we (again) obtain that

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{z}) &\underset{\text{def.}}{=} \frac{\partial}{\partial \boldsymbol{z}} \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} \le \boldsymbol{z}) \underset{\text{Leibniz}}{\overset{(13)}{=}} \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Z} = \boldsymbol{z}) f_{\boldsymbol{Z}}(\boldsymbol{z}) \\ &\underset{\text{def.}}{=} f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \,|\, \boldsymbol{z}) f_{\boldsymbol{Z}}(\boldsymbol{z}), \end{aligned} \tag{14}$$

which generalizes (11) for conditional densities to discrete $\boldsymbol{X}$ and absolutely continuous $\boldsymbol{Z}$. This case is needed in Bayesian statistics when the parameter is continuously distributed but the distribution considered is discrete (e.g. consider uncountable mixtures of discrete dfs).

3) By taking the limit $\boldsymbol{z} \to \boldsymbol{\infty}$, we obtain from (10) the margin

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \lim_{\boldsymbol{z}\to\boldsymbol{\infty}} F(\boldsymbol{x}, \boldsymbol{z}) = \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \,|\, \boldsymbol{z}) \,\mathrm{d}F_{\boldsymbol{Z}}(\boldsymbol{z}), \quad \boldsymbol{x} \in \mathbb{R}^{d_{\boldsymbol{x}}}. \tag{15}$$

- If $F_{\boldsymbol{X}|\boldsymbol{Z}}$ has integrable derivative, then $F_{\boldsymbol{X}}$ has density

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}} F_{\boldsymbol{X}}(\boldsymbol{x}) \underset{\text{Leibniz}}{=} \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \,|\, \boldsymbol{z}) \,\mathrm{d}F_{\boldsymbol{Z}}(\boldsymbol{z}), \quad \boldsymbol{x} \in \mathbb{R}^{d_{\boldsymbol{x}}}. \tag{16}$$

  Furthermore, if $F_{\boldsymbol{Z}}$ has density $f_{\boldsymbol{Z}}$, then

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \,|\, \boldsymbol{z}) f_{\boldsymbol{Z}}(\boldsymbol{z}) \,\mathrm{d}\boldsymbol{z}, \quad \boldsymbol{x} \in \mathbb{R}^{d_{\boldsymbol{x}}}. \tag{17}$$

  Similarly for pmfs $f_{\boldsymbol{X}}$.
- If $F_{\boldsymbol{Z}}$ has density $f_{\boldsymbol{Z}}$, then

$$F_{\boldsymbol{X}}(\boldsymbol{x}) \underset{(15)}{=} \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \,|\, \boldsymbol{z}) f_{\boldsymbol{Z}}(\boldsymbol{z}) \,\mathrm{d}\boldsymbol{z}, \quad \boldsymbol{x} \in \mathbb{R}^{d_{\boldsymbol{x}}}. \tag{18}$$

- If $F_{\boldsymbol{Z}}$ has pmf $f_{\boldsymbol{Z}}$, then $F_{\boldsymbol{X}}(\boldsymbol{x}) \underset{(15)}{=} \sum_{\boldsymbol{z}\in\operatorname{supp}(f_{\boldsymbol{Z}})} F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \,|\, \boldsymbol{z}) f_{\boldsymbol{Z}}(\boldsymbol{z})$, $\boldsymbol{x} \in \mathbb{R}^{d_{\boldsymbol{x}}}$, so we can recover the law of total probability

$$\mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}) = \sum_{\boldsymbol{z}\in\operatorname{supp}(f_{\boldsymbol{Z}})} \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x} \,|\, \boldsymbol{Z} = \boldsymbol{z}) \mathbb{P}(\boldsymbol{Z} = \boldsymbol{z}).$$

Example 8.19

- Applied to $\mathcal{G} = \sigma(\boldsymbol{Z})$, the law of total expectation says that for a rv X with $\mathbb{E}(|X|) < \infty$, one has

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid \boldsymbol{Z})).$$

- With the formulas we just derived, we can also prove this in case $(X, \boldsymbol{Z})$ has a density (or pmf) $f_{X,\boldsymbol{Z}}$ via

$$\begin{aligned}
\mathbb{E}(X) &= \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x \underset{(17)}{=} \int_{\mathbb{R}} x \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} f_{X|\boldsymbol{Z}}(x \mid \boldsymbol{z}) f_{\boldsymbol{Z}}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}x \\
&\underset{\text{Fubini}}{=} \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} \int_{\mathbb{R}} x f_{X|\boldsymbol{Z}}(x \mid \boldsymbol{z})\,\mathrm{d}x\, f_{\boldsymbol{Z}}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \\
&\underset{\text{def.}}{=} \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} \mathbb{E}(X \mid \boldsymbol{Z} = \boldsymbol{z}) f_{\boldsymbol{Z}}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \\
&= \mathbb{E}(\mathbb{E}(X \mid \boldsymbol{X} = \boldsymbol{z})|_{\boldsymbol{z}=\boldsymbol{Z}}) \\
&\underset{\text{fact.}}{=} \mathbb{E}(\mathbb{E}(X \mid \boldsymbol{Z})).
\end{aligned}$$

Proposition 8.20 (Law of total variance)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ a σ-algebra. If $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, then $\mathrm{var}(X) = \mathbb{E}(\mathrm{var}(X \,|\, \mathcal{G})) + \mathrm{var}(\mathbb{E}(X \,|\, \mathcal{G}))$ (*law of total variance*).

Proof. We have

$$\begin{aligned}
\mathrm{var}(X \,|\, \mathcal{G}) &\underset{\text{def.}}{=} \mathbb{E}\big((X - \mathbb{E}(X \,|\, \mathcal{G}))^2 \,\big|\, \mathcal{G}\big) \overset{\text{multiply}}{\underset{\text{out}}{=}} \mathbb{E}\big(X^2 - 2X\mathbb{E}(X \,|\, \mathcal{G}) + (\mathbb{E}(X \,|\, \mathcal{G}))^2 \,\big|\, \mathcal{G}\big) \\
&\underset{\text{lin.}}{=} \mathbb{E}(X^2 \,|\, \mathcal{G}) - 2\mathbb{E}(X\mathbb{E}(X \,|\, \mathcal{G}) \,|\, \mathcal{G}) + \mathbb{E}\big((\mathbb{E}(X \,|\, \mathcal{G}))^2 \,\big|\, \mathcal{G}\big) \\
&\underset{\text{2x TOWIK}}{=} \mathbb{E}(X^2 \,|\, \mathcal{G}) - 2\mathbb{E}(X \,|\, \mathcal{G})\mathbb{E}(X \,|\, \mathcal{G}) + (\mathbb{E}(X \,|\, \mathcal{G}))^2 \cdot 1 \\
&= \mathbb{E}(X^2 \,|\, \mathcal{G}) - (\mathbb{E}(X \,|\, \mathcal{G}))^2
\end{aligned}$$

and thus

$$\begin{aligned}
\mathrm{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 &\overset{\text{2x tot.}}{\underset{\text{exp.}}{=}} \mathbb{E}(\mathbb{E}(X^2 \,|\, \mathcal{G})) - \big(\mathbb{E}(\mathbb{E}(X \,|\, \mathcal{G}))\big)^2 \\
&\overset{\text{just}}{\underset{\text{shown}}{=}} \mathbb{E}\big(\mathrm{var}(X \,|\, \mathcal{G}) + (\mathbb{E}(X \,|\, \mathcal{G}))^2\big) - \big(\mathbb{E}(\mathbb{E}(X \,|\, \mathcal{G}))\big)^2 \\
&\underset{\text{lin.}}{=} \mathbb{E}(\mathrm{var}(X \,|\, \mathcal{G})) + \mathbb{E}(\mathbb{E}(X \,|\, \mathcal{G})^2) - \big(\mathbb{E}(\mathbb{E}(X \,|\, \mathcal{G}))\big)^2 \\
&= \mathbb{E}(\mathrm{var}(X \,|\, \mathcal{G})) + \mathrm{var}(\mathbb{E}(X \,|\, \mathcal{G}))
\end{aligned}$$

□

For $\mathcal{G} = \sigma(\boldsymbol{Z})$, $\mathrm{var}(X) = \mathbb{E}(\mathrm{var}(X \,|\, \boldsymbol{Z}))$ is generally wrong as it does not take into account the variance of $\boldsymbol{Z}$ itself.

Proposition 8.21 (Random sums)

Let $N, X_1, X_2, \dots$ be independent, $N \in \mathbb{N}_0$ a.s., $(X_i)_{i\in\mathbb{N}}$ iid and $S = \sum_{i=1}^N X_i$.

1) If $N, X_1 \in L^1$, then $\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(X_1)$.

2) If $N, X_1 \in L^2$, then $\operatorname{var}(S) = \mathbb{E}(N)\operatorname{var}(X_1) + \operatorname{var}(N)(\mathbb{E}(X_1))^2$.

Proof.

1) i) If $X_1 \overset{\text{a.s.}}{\geq} 0$, apply T. 8.17 with $g(\boldsymbol{X}, N) = \sum_{i=1}^N X_i \in L_+$. Then $h(n) = \mathbb{E}(g(\boldsymbol{X}, n)) = \mathbb{E}(\sum_{i=1}^n X_i) \underset{\text{lin.}}{=} \sum_{i=1}^n \mathbb{E}(X_i)$, so that, a.s., $\mathbb{E}(S \mid N) = \mathbb{E}(\sum_{i=1}^N X_i \mid N) = \mathbb{E}(g(\boldsymbol{X}, N) \mid N) \underset{\text{T. 8.17}}{=} h(N) \underset{\text{plug-in}}{=} \sum_{i=1}^N \mathbb{E}(X_i) \underset{\text{id}}{=} N\mathbb{E}(X_1)$.

Therefore, $\mathbb{E}(S) \overset{\text{tot.}}{\underset{\text{exp.}}{=}} \mathbb{E}(\mathbb{E}(S \mid N)) \overset{\text{as}}{\underset{\text{shown}}{=}} \mathbb{E}(N\mathbb{E}(X_1)) \underset{\text{lin.}}{=} \mathbb{E}(N)\mathbb{E}(X_1)$.

ii) General X_1: $S = \sum_{i=1}^N (X_i^+ - X_i^-) = \sum_{i=1}^N X_i^+ - \sum_{i=1}^N X_i^- =: S^+ - S^-$
$\Rightarrow \mathbb{E}(S) \underset{\text{lin.}}{=} \mathbb{E}(S^+) - \mathbb{E}(S^-) \underset{\text{i)}}{=} \mathbb{E}(N)\mathbb{E}(X_1^+) - \mathbb{E}(N)\mathbb{E}(X_1^-) \underset{\text{lin.}}{=} \mathbb{E}(N)\mathbb{E}(X_1)$.

2) $\operatorname{var}(S) \overset{\text{tot.}}{\underset{\text{var.}}{=}} \mathbb{E}(\operatorname{var}(S \mid N)) + \operatorname{var}(\mathbb{E}(S \mid N)) \overset{\text{ind.}}{\underset{\text{T. 8.17}}{=}} \mathbb{E}(\sum_{i=1}^N \operatorname{var}(X_i \mid N)) + \operatorname{var}(\mathbb{E}(S \mid N))$
$\overset{\text{ind.}}{\underset{\text{P. 8.7 2)}}{=}} \mathbb{E}(\sum_{i=1}^N \operatorname{var}(X_i)) + \operatorname{var}(\mathbb{E}(S \mid N)) \underset{\text{id}}{=} \mathbb{E}(N\operatorname{var}(X_1)) + \operatorname{var}(\mathbb{E}(S \mid N)) \underset{\text{1)}}{=}$
$\mathbb{E}(N)\operatorname{var}(X_1) + \operatorname{var}(N\mathbb{E}(X_1)) = \mathbb{E}(N)\operatorname{var}(X_1) + \operatorname{var}(N)(\mathbb{E}(X_1))^2$. □

Question: What is the best L^2-approximation to a rv X by a $\mathcal{G}$-measurable rv?

Recall: Since

$$\begin{aligned}\mathbb{E}((X-c)^2) &= \mathbb{E}\Big((X - \mathbb{E}(X) + \mathbb{E}(X) - c)^2\Big)\\ &\underset{\text{lin.}}{\overset{\text{multiply out}}{=}} \operatorname{var}(X) + 2\mathbb{E}((X-\mathbb{E}(X))(\mathbb{E}(X)-c)) + (\mathbb{E}(X)-c)^2\\ &\underset{\text{lin.}}{=} \operatorname{var}(X) + 0 + (\mathbb{E}(X)-c)^2,\end{aligned}$$

we know that $\mathbb{E}(X)$ is the best L^2-approximation to X by a constant.

Answer: If $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, it is $\mathbb{E}(X\,|\,\mathcal{G})$ ($\Rightarrow$ interpretation in this case).

Proposition 8.22 (Conditional expectation as best L^2-approximation)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ a σ-algebra and $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then $\mathbb{E}(X\,|\,\mathcal{G})$ is the $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})$ which minimizes $\mathbb{E}((X-Y)^2)$, that is

$$\mathbb{E}((X-Y)^2) = \inf_{Z \in L^2(\Omega,\mathcal{G},\mathbb{P})} \mathbb{E}((X-Z)^2) \quad \text{iff} \quad Y \overset{\text{a.s.}}{=} \mathbb{E}(X\,|\,\mathcal{G}).$$

Proof. Multiplying out and using linearity,

$$\mathbb{E}((X-Y)^2) = \mathbb{E}((X - \mathbb{E}(X\,|\,\mathcal{G}) + \mathbb{E}(X\,|\,\mathcal{G}) - Y)^2)$$

$$= \mathbb{E}\big((X-\mathbb{E}(X\,|\,\mathcal{G}))^2\big)+2\mathbb{E}\big((X-\mathbb{E}(X\,|\,\mathcal{G}))(\mathbb{E}(X\,|\,\mathcal{G})-Y)\big)+\mathbb{E}\big((\mathbb{E}(X\,|\,\mathcal{G})-Y)^2\big).$$

The first summand does not depend on Y, the third summand is minimal iff $Y \overset{\text{a.s.}}{=} \mathbb{E}(X\,|\,\mathcal{G})$, and any such Y is $\mathcal{G}$-measurable and in L^2 by P. 8.9 1) since $X \in L^2$. We are thus done if we can show that the second summand is 0:

$$\begin{aligned}
&\mathbb{E}\big((X-\mathbb{E}(X\,|\,\mathcal{G}))(\mathbb{E}(X\,|\,\mathcal{G})-Y)\big)\\
\overset{\text{tot.}}{\underset{\text{exp.}}{=}}\;& \mathbb{E}\Big(\mathbb{E}\big((X-\mathbb{E}(X\,|\,\mathcal{G}))(\mathbb{E}(X\,|\,\mathcal{G})-Y)\,|\,\mathcal{G}\big)\Big)\\
\underset{\text{TOWIK}}{=}\;& \mathbb{E}\Big((\mathbb{E}(X\,|\,\mathcal{G})-Y)\mathbb{E}\big((X-\mathbb{E}(X\,|\,\mathcal{G}))\,|\,\mathcal{G}\big)\Big)\\
\underset{\text{lin.}}{=}\;& \mathbb{E}\Big((\mathbb{E}(X\,|\,\mathcal{G})-Y)\big(\mathbb{E}(X\,|\,\mathcal{G})-\mathbb{E}(\mathbb{E}(X\,|\,\mathcal{G})\,|\,\mathcal{G})\big)\Big)\\
\overset{\text{tower or}}{\underset{\text{TOWIK}}{=}}\;& \mathbb{E}\Big((\mathbb{E}(X\,|\,\mathcal{G})-Y)\big(\mathbb{E}(X\,|\,\mathcal{G})-\mathbb{E}(X\,|\,\mathcal{G})\big)\Big)=0.
\end{aligned}$$ □

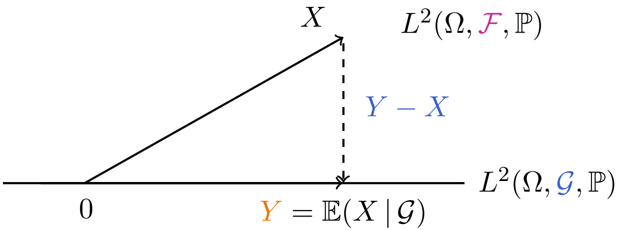
Remark 8.23 (Interpretation via orthogonal projections)

- $L^2(\Omega,\mathcal{F},\mathbb{P})$ is a Hilbert space with inner product $\langle X,Y\rangle := \mathbb{E}(XY)$ and $L^2(\Omega,\mathcal{G},\mathbb{P})$ is a closed subspace.
- The point $Y \in L^2(\Omega,\mathcal{G},\mathbb{P})$ closest to $X \in L^2(\Omega,\mathcal{F},\mathbb{P})$ is known as *orthogonal*

projection of X onto $L^2(\Omega, \mathcal{G}, \mathbb{P})$ and is given by

$$\underset{Y \in L^2(\Omega,\mathcal{G},\mathbb{P})}{\operatorname{arginf}} \mathbb{E}((X-Y)^2) \underset{\text{T. 8.22}}{=} \mathbb{E}(X \mid \mathcal{G}).$$

- Sketch:



Indeed,

$$\langle Y, Y - X\rangle = \mathbb{E}(\mathbb{E}(X \mid \mathcal{G}) \cdot (\mathbb{E}(X \mid \mathcal{G}) - X)) \overset{\text{multiply}}{\underset{\text{out}}{=}} \mathbb{E}(\mathbb{E}(X \mid \mathcal{G})^2 - X\mathbb{E}(X \mid \mathcal{G}))$$

$$\underset{\text{lin.}}{=} \mathbb{E}(\mathbb{E}(X \mid \mathcal{G})^2) - \mathbb{E}(X\mathbb{E}(X \mid \mathcal{G}))$$

$$\overset{\text{tot.}}{\underset{\text{exp.}}{=}} \mathbb{E}(\mathbb{E}(X \mid \mathcal{G})^2) - \mathbb{E}\big(\mathbb{E}(X\mathbb{E}(X \mid \mathcal{G}) \mid \mathcal{G})\big)$$

$$\underset{\text{TOWIK}}{=} \mathbb{E}(\mathbb{E}(X \mid \mathcal{G})^2) - \mathbb{E}(\mathbb{E}(X \mid \mathcal{G}) \cdot \mathbb{E}(X \mid \mathcal{G})) = 0.$$