Advanced probability

November 7, 2025

Contents

Chapter 1

Introduction

1.1 Probability space

Definition 1.1.1 (Probability space)**.** Considering an experiment, it's a triplet $(\Omega, \mathscr{F}, \mathbb{P})$, used to describe the experiment it in mathematical way, composed by a set called sample space Ω, a σ-algebra $\mathscr{A}$ on it (or, same σ-field $\mathscr{F}$) and a probability function $\mathbb{P}$.

1.1.1 Sample space, events

Definition 1.1.2 (Sample space, Ω)**.** The (non-null) set of possible outcomes of an experiment, $\Omega = \{\omega_1, \omega_2, \ldots\}$, of which *only one will occur.*

Remark 1. The assumption is that before executing the experiment we can know all the possible outcomes.

Example 1.1.1 (Coin toss)**.** Here $\Omega = \{h, t\}$ while h is one possible outcome. We could be interested in the events outcome is head $\{h\}$ (singleton), outcome is either head or tail, outcome is not a head etc.

Example 1.1.2 (Two dice throwing)**.** $\Omega = \{(1,1), (2,1), \ldots, (6,6)\}$. The event $E = \text{first is one} = \{(1,1), \ldots, (1,6))\}$

Example 1.1.3 (Arrival order)**.** In arrival order of a race with 7 numbered horses $\Omega = \{7!\ \text{permutations of (1,2,3,4,5,6,7)}\}$.

Example 1.1.4 (Number of cars counted at a crossroad during a minute)**.** $\Omega = \{0, 1, 2, \ldots\}$

Example 1.1.5 (Bulb lifetime)**.** Will be a positive real number so $\Omega = \{x \in \mathbb{R}^+ | x \geq 0\}$.

Example 1.1.6 (Multivariate Rigo's examples)**.** Ω could be $\mathbb{R}^n$ or $\mathbb{C}$ (set of complex number, $(\mathbb{I}, \mathbb{R})$ (but this latter definition belongs to theoretical eperiments)

Definition 1.1.3 (Sample space cardinality)**.** Sample spaces of experiments can be *finite* (eg 1.1.1, 1.1.2) 1.1.3) *countable* (in bijection with $\mathbb{N}$, eg 1.1.4) or *non countable* (bijection with $\mathbb{R}$, eg 1.1.5)

Definition 1.1.4 (Outcome, ω)**.** One possible result of the experiment: $\omega \in \Omega$.

Definition 1.1.5 (Event (E or A))**.** Any subset of the sample space Ω.

Definition 1.1.6 (Occurred event)**.** E occurred if it contains the result of the experiment.

Remark 2. Since an event is any subset of Ω the following are valid.

Definition 1.1.7 (True event (Ω))**.** Always occurs, since at least an element of the Ω occurs during the event.

Definition 1.1.8 (Impossible event ($\emptyset$))**.** Never occurs.

Definition 1.1.9 (Singleton event (eventi elementari), $\{\omega\}$)**.** Events composed by a single experiment outcome.

Remark 3 (Plotting). With Venn diagram Ω is given by a rectangle, while events are represented by circles.

1.1.1.1 Events algebra

Remark 4. Rules that applies to create new events; inherits from set theory being the events a set.

Definition 1.1.10 (Union $A \cup B$)**.** Event that occurs if occurs one of A or B.

Remark 5. The outcomes composing the event are given by union of the outcomes of starting events.

Remark 6. Union can be extended to a numerable infinite number of events

$$E_1 \cup E_2 \cup \ldots \cup E_n \cup \ldots = \bigcup_{i=1}^{\infty} E_i \tag{1.1}$$

and verifies if at least one of E_i happens.

Definition 1.1.11 (Intersection $A \cap B$ (A, B or AB))**.** Event that occurs if occur both A and B.

Remark 7. The outcome composing the event are given by intersection of the outcomes of starting events.

Remark 8. Similarly intersection event can be extended to a numerable infinite set of events

$$E_1 \cap E_2 \cap \ldots \cap E_n \cap \ldots = \bigcap_{i=1}^{\infty} E_i \tag{1.2}$$

Definition 1.1.12 (Complement/negation event)**.** The negation of the event A, typed $\overline{A}$ o A^c, is the event that happens if A does not: $A^c = \Omega \setminus A$.

Definition 1.1.13 (Difference $A \setminus B$)**.** Events that occurs when A occurs but not B: $A \setminus B = A \cap \overline{B}$.

Remark 9. The outcome composing the event are given by the set difference $A \setminus B$ outcomes of starting events.

Definition 1.1.14 (Symmetric difference $A \bigtriangleup B$ (xor))**.** Events that occur if A or B occurs, but not both

Remark 10. The outcome composing the event are given by $(A \cup B) \setminus (A \cap B)$.

Property	Union	Intersection
Idempotenza	$A \cup A = A$	$A \cap A = A$
Elemento neutro	$A \cup \emptyset = A$	$A \cap \Omega = A$
Commutativa	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Associativa	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
Distributiva	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Table 1.1: Proprietà di unione ed intersezione

Operation properties

Important remark 1. Operation properties are the same as set properties and summarized in tab 1.1; same for DeMorgan Laws.

Proposition 1.1.1 (DeMorgan laws)**.** *With two events*

$$\overline{A \cap B} = \overline{A} \cup \overline{B} \tag{1.3}$$

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \tag{1.4}$$

while in the general form

$$\overline{\bigcap_i E_i} = \bigcup_i \overline{E_i} \tag{1.5}$$

$$\overline{\bigcup_i E_i} = \bigcap_i \overline{E_i} \tag{1.6}$$

1.1.1.2 Relationship between events

Definition 1.1.15 (Inclusion, $A \subseteq B$)**.** Event A is included in B, $A \subseteq B$ if each time A happens, B happens as well.

Example 1.1.7. $E_1 = \{1, 2\}$ ("dice below 3") is included in $E_2 = \{1, 2, 3\}$ ("dice below 4")

Definition 1.1.16 (Monotone increasing sequence of events)**.** A sequence of events $E_1, E_2, \ldots$ where $E_1 \subseteq E_2 \subseteq \ldots$.

Definition 1.1.17 (Monotone decreasing sequence of events)**.** A sequence of events $E_1, E_2, \ldots$ where $E_1 \supseteq E_2 \supseteq \ldots$

Definition 1.1.18 (Incompatibility/disjointness, $A \cap B = \emptyset$)**.** A and B are incompatible (or disjoint) if they can't verify together, that is, $A \cap B = \emptyset$.

Example 1.1.8. If $A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ (two dice sum to 7) and $B = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$ (sum to 6) are incompatible because $A \cap B = \emptyset$.

Remark 11. In Venn diagrams, two disjoint events are represented by non overlapping areas.

Definition 1.1.19 (Pairwise disjointness/incompatibility/exclusiveness)**.** Given a collection of events E_i, $1 \le i \le \infty$, they are pairwise disjoint if

$$E_i \cap E_j = \emptyset \quad \forall i \neq j$$

Important remark 2. The same can be defined for 3-folded incompatibility or n-folded. Clearly pairwise disjointness implies higher level disjointness (eg 3-folded, etc); viceversa does not happens.

Definition 1.1.20 (Jointly exhaustive events (eventi necessari), $A \cup B = \Omega$)**.** A and B are jointly exhaustive if at least one event occurs, that is $A \cup B = \Omega$.

Remark 12. Same applies for a collection: $E_i, 1 \leq i \leq \infty$ is jointly exhaustive if at least one event occurs $\bigcup_{i=1}^{\infty} E_i = \Omega$

Definition 1.1.21 (Ω partition)**.** It's a set of events $\{E_i\}_{i \in I}, E_i \subseteq \Omega$ which are both disjoint and jointly exhaustive:

$$E_i \cap E_j = \emptyset \quad i \neq j, \qquad \bigcup_{i=1}^{\infty} E_i = \Omega$$

Remark 13. If the set of events E_i is finite, countable or uncountable (eg idem the set of index I), the partition of Ω will respectively be called finite, countable or uncountable.

Remark 14. On Venn diagrams it's a set of non overlapping shapes that sum up to Ω.

Example 1.1.9. Suppose $\Omega = \mathbb{R}$, collection of all $\{x\}$ with $x \in \mathbb{R}$ is a partition (not finite nor countable, it's uncountable).

1.1.2 σ-algebra $\mathscr{A}$ (or σ-field $\mathscr{F}$)

Definition 1.1.22 (σ-algebra $\mathscr{A}$ (or σ-field $\mathscr{F}$))**.** Set of all the possible events of interest, $\mathscr{A} \subseteq \mathcal{P}(\Omega)$ having the following properties

1. $\Omega \in \mathscr{A}$
2. $\mathscr{A}$ is closed under complements: $A \in \mathscr{A} \implies A^c \in \mathscr{A}$
3. $\mathscr{A}$ is closed under *finite* or *countable* unions (and intersection as well): if $E_1, E_2, \ldots \in \mathscr{A}$ is a finite or countable set of events then $\bigcup_{i=1}^{\infty} E_i \in \mathscr{A}$

Lemma 1.1.2. *Thus we have that* $\emptyset \in \mathscr{A}$ *and* $\mathscr{A}$ *is closed under finite or countable intersection as well:*

$$\begin{aligned} &\emptyset = \Omega^c \in \mathscr{A} \\ &E_1, E_2, \ldots \in \mathscr{A} \implies \bigcap_{i=1}^{+\infty} E_i = \left(\bigcup_{i=1}^{+\infty} E_i^c \right)^c \in \mathscr{A} \end{aligned}$$

the last by applying proprieties 2, 3 of the definition and DeMorgan's laws.

Important remark 3 (The idea). Events are subset of Ω but it's not needed all the subsets of Ω, elements of $\mathcal{P}(\Omega)$, to be events (for technical complex reasons). It suffices for us to think of the collection of events as a subcollection $\mathscr{A} \subseteq \mathcal{P}(\Omega)$ of the power set of the sample space, having certain reasonable/minimal properties. The idea is that:

- $\mathscr{A}$ can be thought as the set of all possible events that are relevant regarding the considered experiment (probabilistic meaning of $\mathscr{A}$)

- if I make some operations of interest between events (unions, intersections, complement), I can be confident of being inside the σ-algebra.
- if the set of possible events $\mathscr{E}$ of our interest is not a σ-algebra, then we set $\mathscr{A} = \sigma(\mathscr{E})$ as the minimum σ-algebra containing $\mathscr{E}$, and "work" with this one.

Example 1.1.10. $\mathscr{A} = \{\emptyset, \Omega\}$ is the least possible (più piccola) σ-algebra

Example 1.1.11. $\mathscr{A} = \{\emptyset, \Omega, A, A^c\}$ is the least possible σ-algebra including A.

Example 1.1.12 (Power set (insieme delle parti) as $\mathscr{A}$)**.** $\mathscr{A} = \mathcal{P}(\Omega)$ is the most possible sigma field; no other $\mathscr{A}$ can be bigger (in terms of cardinality). If:

- Ω is finite, it can be $\mathscr{A} = \mathcal{P}(\Omega)$.
- Ω is countable (eg $\mathbb{N}$), its power set can be a σ-algebra (see here).
- Ω is *non countable* (eg $\Omega = \mathbb{R}$), its power set is a too large collection for probabilities to be assigned reasonably (eg all being non negative and singleton events probabilities summing up to 1) to all its members

Important remark 4. In case of $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^n$ we consider a particular case of σ-field/algebra called Borel σ-field/algebra

Definition 1.1.23 (Intervals of $\mathbb{R}$)**.** The intervals of $\mathbb{R}$ are (a, b), $[a, b]$, $(a, b]$, $[a, b)$ $(-\infty, b]$, $(-\infty, b)$, (a, ∞), $[a, \infty)$, and $\mathbb{R}$ as well.

Definition 1.1.24 (Borel σ-field on $\mathbb{R}$)**.** The borel sigma-field on $\mathbb{R}$, denoted by $\beta(\mathbb{R})$, is the least possible sigma-field including all the $\mathbb{R}$ intervals.

Remark 15. Some remarks:

- if $\Omega = \mathbb{R}$ and $\mathscr{E}$ is a set of intervals of $\mathbb{R}$ but *not* a σ-algebra (eg like borel), by definition it could happen that $(-1, 5) \cup [7, 8] \notin \mathscr{E}$; so the property/definition of borel seem reasonable/desiderable;
- $\beta(\mathbb{R})$ includes all sets which can be obtained, starting from intervals, by a countable numbers of unions, intersections and complements;
- note that $\exists A \subset \mathbb{R}$ such as that $A \notin \beta(\mathbb{R})$; in other terms $\beta(\mathbb{R})$ is *not* the power set of $\mathbb{R}$.

Example 1.1.13 (singleton events and $\beta(\mathbb{R})$)**.** Singleton events are contained in $\beta(\mathbb{R})$ since can be written as intersection between intervals $x = (x-1, x] \cap [x, x+1) \in \beta(\mathbb{R})\ \forall x \in \mathbb{R}$.

Definition 1.1.25 (Borel σ-field on $\mathbb{R}^n$)**.** In the same way, if $\Omega = \mathbb{R}^n$, $\beta(\mathbb{R}^n)$ equals to the least σ-field on $\mathbb{R}^n$ including all sets of the form $I_1 \times I_2 \times \ldots \times I_n$, where each I_i is an interval of $\mathbb{R}$.

Example 1.1.14. Graphically think as set of rectangles in the space, eg if $n = 2$ a set of rectangles $I_1 \times I_2$

1.1.3 Probability measure $\mathbb{P}$

Remark 16. In our construction the third element is the probability function $\mathbb{P}$, defined according to three Kolmogorov axioms that specifies basic features of any probability function.

Definition 1.1.26 (Measure)**.** A measure, generally speaking, is a function:

1. assigning a positive number to each set
2. for which measure of union of disjoint set is sum of measure of the sets.

Definition 1.1.27 (Probability function, $\mathbb{P}$)**.** It's a measure characterized[1] by $\mathbb{P}(\Omega) = 1$, so it's a function $\mathbb{P} : \mathscr{A} \to [0, 1]$ such that:

$$\mathbb{P}(A) \geq 0, \quad \forall A \in \mathscr{A} \tag{1.7}$$

$$\mathbb{P}(\Omega) = 1 \tag{1.8}$$

$$A_i \cap A_j = \emptyset, \forall i \neq j \implies \mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i) \tag{1.9}$$

For the latter one (called σ-additivity), set $\{A_1, A_2, \ldots\}$ is a *finite* or *countable* set of incompatible events.

Example 1.1.15. A coin, possibly biased is tossed once. We have $\Omega = \{h, t\}$, $\mathscr{A} = \{\emptyset, \{h\}, \{t\}, \Omega\}$ and a *possible* probability measure (it fullfill the requirements) $\mathbb{P} : \mathscr{A} \to [0, 1]$ is given by

$$\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(\{h\}) = p, \quad \mathbb{P}(\{t\}) = 1 - p, \quad \mathbb{P}(\Omega = \{h, t\}) = 1$$

where p is a fixed real number in the interval $[0, 1]$. If $p = \frac{1}{2}$ then we say the coin is *fair* or unbiased.

Definition 1.1.28 (Null event)**.** Events A such as $\mathbb{P}(A) = 0$.

Definition 1.1.29 (Almost sure event)**.** Event A such as $\mathbb{P}(A) = 1$.

Important remark 5 (Null vs impossible events, true vs almost surely events). Null events should not be confused with the impossible event $\emptyset$: null events are happening all around us, even though they have zero probability (eg what's the chance that a dart strikes any given point of the target at which it's thrown). That is: the impossible event is null, but null events need not to be impossible. Specular considerations for Ω with events A such as $\mathbb{P}(A) = 1$.

1.2 Probability

1.2.1 Immediate or useful general results

Remark 17. Let's see some properties following directly from the definition; in what follows we consider generic events $A, B \subseteq \Omega$.

[1] For a measure (in general), it may be that $P(\Omega) = 0$ or $P(\Omega) = +\infty$ as well; but not for probability, for which $\mathbb{P}(\Omega) = 1$.

Proposition 1.2.1.

$$\boxed{\mathbb{P}\left(\overline{A}\right) = 1 - \mathbb{P}\left(A\right)} \tag{1.10}$$

Proof.

$$\begin{aligned} \Omega &= A \cup \overline{A} \\ \mathbb{P}\left(\Omega\right) &= \mathbb{P}\left(A \cup \overline{A}\right) \\ 1 &= \mathbb{P}\left(A\right) + \mathbb{P}\left(\overline{A}\right) \end{aligned}$$

□

Example 1.2.1. If the probability of having head with coin is $\frac{3}{8}$ then probability of tail have to be $\frac{5}{8}$.

Proposition 1.2.2.

$$\boxed{\mathbb{P}\left(\emptyset\right) = 0} \tag{1.11}$$

Proof. Setting $A = \Omega$ in 1.10,

$$\begin{aligned} \mathbb{P}\left(\overline{\Omega}\right) &= 1 - \mathbb{P}\left(\Omega\right) \\ \mathbb{P}\left(\emptyset\right) &= 1 - 1 \end{aligned}$$

□

Proposition 1.2.3.

$$\boxed{A \subseteq B \implies \mathbb{P}\left(A\right) \leq \mathbb{P}\left(B\right)} \tag{1.12}$$

Proof. If $A \subseteq B$, B can be written as union of two incompatible events A and $(B \setminus A)$; applying third axiom

$$\begin{aligned} B &= A \cup (B \setminus A) \\ \mathbb{P}\left(B\right) &= \mathbb{P}\left(A\right) + \mathbb{P}\left(B \setminus A\right) \end{aligned}$$

since $\mathbb{P}\left(B \setminus A\right) \geq 0$ by axioms, then $\mathbb{P}\left(B\right) \geq \mathbb{P}\left(A\right)$, □

Proposition 1.2.4 (Probability that A occurs but not B)**.**

$$\boxed{\mathbb{P}\left(A \setminus B\right) = \mathbb{P}\left(A \cap \overline{B}\right) = \mathbb{P}\left(A\right) - \mathbb{P}\left(A \cap B\right)} \tag{1.13}$$

Proof. Looking at A as union of incompatible events (think using Venn diagram):

$$\begin{aligned} A &= (A \cap B) \cup (A \cap \overline{B}) \\ \mathbb{P}\left(A\right) &= \mathbb{P}\left(A \cap B\right) + \mathbb{P}\left(A \cap \overline{B}\right) \end{aligned}$$

then we conclude as in proposition. □

Proposition 1.2.5 (Probability of union)**.**

$$\boxed{\mathbb{P}\left(A \cup B\right) = \mathbb{P}\left(A\right) + \mathbb{P}\left(B\right) - \mathbb{P}\left(A \cap B\right)} \tag{1.14}$$

Proof. Writing $A \cup B$ as union of two incompatible events,we apply axioms and 1.13:

$$\begin{aligned} A \cup B &= A \cup (B \cap \overline{A}) \\ \mathbb{P}\left(A \cup B\right) &= \mathbb{P}\left(A\right) + \mathbb{P}\left(B \cap \overline{A}\right) \\ \mathbb{P}\left(A \cup B\right) &= \mathbb{P}\left(A\right) + \mathbb{P}\left(B\right) - \mathbb{P}\left(A \cap B\right) \end{aligned}$$

□

Proposition 1.2.6 (Inclusion/exclusion formula)**.** *Considering a finite union of events, probability of their union is calculated according to the following:*

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) &= \sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1 < \ldots < i_r} \mathbb{P}\left(E_{i1} \cap E_{i2} \cap \ldots \cap E_{ir}\right) & (1.15) \\ &= \sum_{i} \mathbb{P}\left(E_i\right) - \sum_{i<j} \mathbb{P}\left(E_i \cap E_j\right) + \sum_{i<j<k} \mathbb{P}\left(E_i \cap E_j \cap E_k\right) - \ldots \\ &\ldots + (-1)^{n+1}\, \mathbb{P}\left(E_1 \cap \ldots \cap E_n\right) & (1.16) \end{aligned}$$

Proof. Can be proved by induction, as we'll see in 3.3.3. □

Example 1.2.2. In case of three events, E, F, G:

$$\begin{aligned} \mathbb{P}\left(E \cup F \cup G\right) = \mathbb{P}\left(E\right) + \mathbb{P}\left(F\right) + \mathbb{P}\left(G\right) - \mathbb{P}\left(E \cap F\right) \ldots \\ - \mathbb{P}\left(E \cap G\right) - \mathbb{P}\left(F \cap G\right) + \mathbb{P}\left(E \cap G \cap F\right) \end{aligned}$$

Proposition 1.2.7 (Boole inequality (on union))**.**

$$\mathbb{P}\left(E_1 \cup E_2 \cup \ldots \cup E_n\right) \leq \sum_{i=1}^{n} \mathbb{P}\left(E_i\right) \tag{1.17}$$

Proof. Done in the following section 3.3.3. □

Proposition 1.2.8 (Bonferroni inequality (on intersection))**.**

$$\mathbb{P}\left(E_1 \cap E_2 \cap \ldots \cap E_n\right) \geq 1 - \sum_{i=1}^{n} \mathbb{P}\left(\overline{E_i}\right) \tag{1.18}$$

Proof. In section 3.3.3. □

Proposition 1.2.9. *If $A_1, A_2, \ldots$ is an increasing sequence of events, so that $A_1 \subseteq A_2 \subseteq \ldots$ and we set A as the limit of the* union*:*

$$A = \bigcup_{i=1}^{+\infty} A_i = \lim_{i \to +\infty} A_i$$

then it follows that

$$\mathbb{P}\left(A\right) = \lim_{i \to +\infty} \mathbb{P}\left(A_i\right) \tag{1.19}$$

Proposition 1.2.10. *Similarly if $B_1, B_2, \dots$ is decreasing sequence of events $B_1 \supseteq B_2 \supseteq \dots$ and we set as B the limit of the* intersection*:*

$$B = \bigcap_{i=1}^{+\infty} B_i = \lim_{i \to +\infty} B_i$$

then

$$\mathbb{P}(B) = \lim_{i \to +\infty} \mathbb{P}(B_i) \tag{1.20}$$

Proof. We prove only the first; we have that A can be seen as an union of a disjoint family of events

$$A = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \dots$$

Thus by definition of the probability function its probability is a sum of the disjoint events (again think with Venn, these are enclosing circles)

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}(A_1) + \sum_{i=1}^{+\infty} \mathbb{P}(A_{i+1} \setminus A_i) \\
&= \mathbb{P}(A_1) + \lim_{n \to +\infty} \sum_{i=1}^{n-1} [\mathbb{P}(A_{i+1}) - \mathbb{P}(A_i)] \\
&= \lim_{n \to +\infty} \mathbb{P}(A_n)
\end{aligned}$$

The last passage involve semplification/elision. For the second results on B, take complements and use the first part. ☐

1.2.2 Finite equiprobable Ω and probability evaluation

Remark 18. In previous section we never evaluated a probability. In this one we show how it's done for the particular case where Ω is finite with every $\omega \in \Omega$ having the same probability of occurring.
It's a reasonable assumption in several cases (eg balanced dice, coins etc)

Proposition 1.2.11 (Probability of singleton a event)**.** *If Ω is* finite*, $\Omega = \{1, 2, \dots, n\}$, and $\mathbb{P}(1) = \mathbb{P}(2) = \dots = \mathbb{P}(n) = p$, being the singleton events disjoint and the probability of their union summing to 1 ($p \cdot n = 1$), we'll have*

$$p = \frac{1}{n}$$

Proposition 1.2.12 (Probability of general event)**.** *Given a generic event E, its probability will be*

$$\mathbb{P}(E) = \frac{\text{\# of outcomes composing } E}{\text{\# possible outcomes}} = \frac{|E|}{|\Omega|}$$

Remark 19. In words, number of favorable outcome of event E out of possible outcomes of Ω. Often, count of numerator/denominator uses combinatorics.

Remark 20. Suppose a partition $E_1, E_2, \dots$ of Ω is *finite* or *countable* and we want to assign the same probability to all E_i. Is it possible?

Proposition 1.2.13. *It's possible to assign to element/events of a* finite *partition of Ω the same probability; if the partition is* countable *this is no more possible.*

Proof. If the partition is *finite* in n events E_i, it suffices to assign $\mathbb{P}(E_i) = \frac{1}{n}$, so that $\mathbb{P}(\Omega) = \mathbb{P}(\cup_{i=1}^{n} E_i) = 1$.
If the partition is countable this is impossible: let's prove it by absurd/contradiction. Suppose be $\mathbb{P}(E_i) = c \geq 0, \forall i$. Then

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathbb{P}(E_i) = \sum_{i=1}^{\infty} c = \begin{cases} 0 & \text{if } c = 0 \\ +\infty & \text{if } c > 0 \end{cases}$$

Therefore we have a contraddiction: 1 can't be equal to 0 or $+\infty$ $\square$

Example 1.2.3 (Concordanza estrazione trial)**.** We have an urn with n numbered balls from 1 to n, we draw without replacement. Let's define C_i = "concordance at trial i" as the selected ball at draw i is numbered i. We are interested in evaluating $\mathbb{P}(E)$ where E = no concordance in n draws.
By applying the previous properties:

$$\begin{aligned}\mathbb{P}(E) &= 1 - \mathbb{P}(\text{at least one concordance}) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} C_i\right) \\ &= 1 - \left\{\sum_{i} \mathbb{P}(C_i) - \sum_{i<j} \mathbb{P}(C_i \cap C_j) + \sum_{i<j<k} \mathbb{P}(C_i \cap C_j \cap C_k) \ldots + (-1)^{n+1} \mathbb{P}(C_1 \cap \ldots \cap C_n)\right\}\end{aligned}$$

Now

$$\mathbb{P}(C_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$$

we have n slots, the sequences of balls can be $n!$, while the sequence where i ball is at the i-th place are $(n-1)!$ (fix i in its place and then permute the remaining balls). Furthermore for similar reasons

$$\begin{aligned}&\mathbb{P}(C_i \cap C_j) = \frac{(n-2)!}{n!} \\ &\mathbb{P}(C_i \cap C_j \cap C_k) = \frac{(n-3)!}{n!} \\ &\ldots \\ &\mathbb{P}(C_1 \cap \ldots \cap C_n) = \frac{1}{n!}\end{aligned}$$

Therefore

$$\mathbb{P}(E) = 1 - \left\{n \cdot \frac{1}{n} - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} \ldots + (-1)^{n+1}\frac{1}{n!}\right\}$$

1.2.3 Conditional probability

1.2.3.1 Introduction/definition

Remark 21 (Idea). Often is needed to compute probability of an event in case another happened; or it's easier to compute a probability of event A conditioning on information of another event B.

Definition 1.2.1 (Conditioned probability of A given B)**.** If $\mathbb{P}(B) > 0$ it's defined as

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}} \tag{1.21}$$

Important remark 6. $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$; denominators are different.

Remark 22. Limit/extreme cases:

$$\begin{aligned} A \cap B = \emptyset &\implies \mathbb{P}(A|B) = 0 \\ A \subseteq B &\implies \mathbb{P}(A|B) = 1 \end{aligned}$$

1.2.3.2 Probability of intersection

Proposition 1.2.14 (For two events, $\mathbb{P}(A \cap B)$)**.** *If* $\mathbb{P}(B) > 0$*:*

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(B)\,\mathbb{P}(A|B)} \tag{1.22}$$

Symmetrically, if $\mathbb{P}(A) > 0$*:*

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B|A)} \tag{1.23}$$

Proof. Algebric manipulation of 1.21. □

Proposition 1.2.15 (n events (*product rule*))**.** *Given* $E_1, \ldots, E_n \in \mathscr{A}$ *if* $\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_{n-1}) > 0$*, then:*

$$\mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right) = \mathbb{P}(E_1) \cdot \mathbb{P}(E_2|E_1) \cdot \mathbb{P}(E_3|E_1 \cap E_2) \cdot \ldots \cdot \mathbb{P}(E_n|E_1 \cap E_2 \cap \ldots \cap E_{n-1})$$

Proof. To verify it we apply recursively the definition 1.23 to the second member:

$$\mathbb{P}(E_1) \cdot \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_1)} \cdot \frac{\mathbb{P}(E_1 \cap E_2 \cap E_3)}{\mathbb{P}(E_1 \cap E_2)} \cdot \ldots \cdot \frac{\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_n)}{\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_{n-1})} \tag{1.24}$$

and after simplifying it remains $\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_n) = \mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right)$.
Note that denominators in 1.24 are strictly positive thanks to the hypothesis $\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_{n-1}) > 0$: since intersection on $n-1$ events is not null, even the intersection of less events will be. □

Remark 23. In practice we can handle/manipulate events as we prefer, eg:

$$\begin{aligned} \mathbb{P}(E_1 \cap E_2 \cap E_3) &= \mathbb{P}(E_1) \cdot \mathbb{P}(E_2|E_1) \cdot \mathbb{P}(E_3|E_1 \cap E_2) \\ &= \mathbb{P}(E_3) \cdot \mathbb{P}(E_2|E_3) \cdot \mathbb{P}(E_1|E_3 \cap E_2) \end{aligned}$$

1.2.3.3 Law of total probability

Remark 24 (Conditioning for problem solving). Sometimes is difficult to calculate $\mathbb{P}(E)$; this can become easier if we can condition on C (and $\overline{C}$), and summing up applying the previous formula. It's common practice to condition on hypothesis/hypothetical situation or, in sequential experiment, conditioning on previous steps.

Definition 1.2.2 (LTP with a single event (and its complement))**.** If E and C are two events we have (with E of interest for probability evaluation) then:

$$\mathbb{P}(E) = \mathbb{P}(C)\,\mathbb{P}(E|C) + \mathbb{P}\left(\overline{C}\right)\mathbb{P}\left(E|\overline{C}\right) \tag{1.25}$$

Proof. We can split E in disjoint union as follows:

$$E = (E \cap C) \cup (E \cap \overline{C})$$

Being disjoint:

$$\begin{aligned}
\mathbb{P}(E) &= \mathbb{P}\left((E \cap C) \cup (E \cap \overline{C})\right) \\
&= \mathbb{P}(E \cap C) + \mathbb{P}\left(E \cap \overline{C}\right) \\
&= \mathbb{P}(C)\,\mathbb{P}(E|C) + \mathbb{P}\left(\overline{C}\right)\mathbb{P}\left(E|\overline{C}\right)
\end{aligned}$$

□

Example 1.2.4. Domani potrebbe o piovere o nevicare, ma i due eventi non si possono verificare contemporaneamente. La probabilità che piova è 2/5 , mentre la probabilità che nevichi è 3/5 . Se pioverà, la probabilità che io faccia tardi a lezione è di 1/5 , mentre la probabilità corrispondente nel caso in cui nevichi è di 3/5. Calcolare la probabilità che io sia in ritardo.
Si ha P = piove, $N = P^c$ = nevica, R = ritardo; avendo a che fare con una partizione

$$\mathbb{P}(R) = \mathbb{P}(P)\,\mathbb{P}(R|P) + \mathbb{P}(N)\,\mathbb{P}(R|N) = \frac{2}{5}\frac{1}{5} + \frac{3}{5}\frac{3}{5} = \frac{11}{25}$$

Theorem 1.2.16 (LTP with a partition)**.** *If* $C_1, C_2, \dots$ *is a* finite *or* countable *partition of* Ω*, the probability of a generic event* E *can be written as (*disintegrability*):*

$$\boxed{\mathbb{P}(E) = \sum_i \mathbb{P}(C_i)\,\mathbb{P}(E|C_i)} \tag{1.26}$$

Important remark 7. Looking at the formula, here it's not a problem if $\mathbb{P}(C_i) = 0$ (which is at the denominator of $\mathbb{P}(E|C_i)$, which would be undefined); undefined multiplied by zero is not considered in the sum.

Proof. If $C_1, C_2, \dots, C_n$ is a partition of Ω, we can split E in disjoint pieces by intersection with C_i

$$E = \Omega \cap E = \left(\bigcup_{i=1}^{n} C_i\right) \cap E = (C_1 \cap E) \cup (C_2 \cap E) \cup \ldots \cup (C_n \cap E)$$

Being $(C_i \cap A)$ disjoint probability is the sum:

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(C_i \cap E) = \sum_{i=1}^{n} \mathbb{P}(C_i)\,\mathbb{P}(E|C_i) \tag{1.27}$$

and in the last passage we substitued 1.23. $\square$

Example 1.2.5 (Esempio Rigo)**.** Having an urn with n_w white and n_b black balls, we draw without replacement. We are interested in $\mathbb{P}(W_2)$ where $W_2 =$ "white ball at second draw": it is not trivial without formula, since we don't know the result of the first trial. We however can calculate it conditioning on first draw results.
Let's set $W_1 =$ "white at first draw" and $B_1 =$ "black at first draw"; since $\{W_1, B_1\}$ is a finite partition of the sample space of the first trial, we can apply the law of total probabilities:

$$\mathbb{P}(W_2) = \mathbb{P}(W_1)\,\mathbb{P}(W_2|W_1) + \mathbb{P}(B_1)\,\mathbb{P}(W_2|B_1)$$

Given that we have $n = n_w + n_b$ balls and we draw without replacement

$$\mathbb{P}(W_1) = \frac{n_w}{n},\ \mathbb{P}(B_1) = \frac{n_b}{n},\ \mathbb{P}(W_2|W_1) = \frac{n_w - 1}{n-1},\ \mathbb{P}(W_2|B_1) = \frac{n_w}{n-1},$$

Therefore, overall

$$\mathbb{P}(W_2) = \frac{n_w}{n} \cdot \frac{n_w - 1}{n-1} + \frac{n_b}{n} \cdot \frac{n_w}{n-1} = \ldots = \frac{n_w}{n}$$

This is a counterintuitive result, since it's the same as drawing *with* replacement.
In general if $W_j =$ white at draw j, $\mathbb{P}(W_j)$ is still $\frac{n_w}{n}$. In this case we have to condition on the partition of the first $j-1$ trials.
For example, considering "$W_3 =$ white at draw 3" the first two draws will have $\Omega = \{ww, wb, bw, bb\}$, so

$$\begin{aligned}\mathbb{P}(W_3) &= \mathbb{P}(ww)\,\mathbb{P}(W_3|ww) + \mathbb{P}(wb)\,\mathbb{P}(W_3|wb) + \mathbb{P}(bw)\,\mathbb{P}(W_3|bw) + \mathbb{P}(bb)\,\mathbb{P}(W_3|bb)\\ &= \ldots = \frac{n_w}{n}\end{aligned}$$

Eg in this case $\mathbb{P}(W_3|ww) = \frac{n_w - 2}{n-2}$

1.2.3.4 Bayes formula

Theorem 1.2.17 (Bayes formula)**.** *If A, B are two events, with $P(B) > 0$ then*

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B|A)}{\mathbb{P}(B)}} \tag{1.28}$$

Proof. Substitute 1.23 in 1.21. $\square$

Remark 25 (Decision making and knowledge update)*.* When performing a test to verify an hypothesis, bayes formula is used like this: let H be "my hypothesis is true", and T "positive test"; then:

$$\mathbb{P}(H|T) = \frac{\mathbb{P}(H) \cdot \mathbb{P}(T|H)}{\mathbb{P}(T)}$$

in this case $\mathbb{P}(H)$ is called *a priori probability* $\mathbb{P}(T|H)$ *likelihood* and $\mathbb{P}(H|T)$ *posterior probability* (the denominator is merely a normalizing constant).

Remark 26 (Bayes in diagnostic: PPV and NPV). If D is "being diseased" and T è "being positive to diagnostic test", $\mathbb{P}(D|T)$ (applying bayes formula) is Positive predictive value while $\mathbb{P}\left(\overline{D}|\overline{T}\right)$ is negative predictive value..

Corollary 1.2.18. *Let E be a generic event and $C_1, C_2, \dots$ a* finite *or* countable *partition of Ω; the conditional probability of C_i given E is:*

$$\boxed{\mathbb{P}(C_i|E) = \frac{\mathbb{P}(C_i)\,\mathbb{P}(E|C_i)}{\sum_i \mathbb{P}(C_i)\,\mathbb{P}(E|C_i)}}$$

Proof. We started from $\mathbb{P}(C_i|E)$ defined using Bayes law and then substituted the denominator using the law of total probability:

$$\mathbb{P}(C_i|E) = \frac{\mathbb{P}(C_i)\,\mathbb{P}(E|C_i)}{\mathbb{P}(E)} = \frac{\mathbb{P}(C_i)\,\mathbb{P}(E|C_i)}{\sum_i \mathbb{P}(C_i)\,\mathbb{P}(E|C_i)}$$

□

Remark 27. For example in Bayesian statistics $C_1, C_2, \dots$ are the possible values of a random parameter while E is the observed sample.

Remark 28 (Interpretation). E can be thought as an occurred event/effect that is dued to only one of n causes C_i (disjoint, exaustive: that is one and only one of them surely happened) each one of the cause has probability $\mathbb{P}(C_i)$ to happen.
The theorem allows us to evaluate $\mathbb{P}(C_i|E)$, that is probability that having observed E, this has been caused by C_i. In the process we use prior probability $\mathbb{P}(C_i)$ and likelihood $\mathbb{P}(E|C_i)$ at numerator (denominator is a normalizing constant):

- when prior probability is not known, if the partition is *finite* (see 1.2.13), one can assign a common probability $\mathbb{P}(C_i) = 1/n, \forall i$;
- likelihood is generally easier to know/evaluate;
- we conclude C_i as the most reasonable cause if its $\mathbb{P}(C_i|E)$ is higher than the others;
- the final result depends only on the numerator, being the denominator a normalizing constant common for all C_i (and making posteriors $\mathbb{P}(C_i|E)$ to sum up to 1). For this reason we can write

  $$\mathbb{P}(C_i|E) \propto \mathbb{P}(C_i)\,\mathbb{P}(E|C_i)$$

  that is posterior probability is proportional to the prior time likelihood

Important remark 8. It's often useful the simpler version of (where the partition of Ω composed by two events, only one of which is of interest, the other is the complement) reported here:

$$\mathbb{P}(H|T) = \frac{\mathbb{P}(H)\cdot\mathbb{P}(T|H)}{\mathbb{P}(H)\cdot\mathbb{P}(T|H) + \mathbb{P}\left(\overline{H}\right)\cdot\mathbb{P}\left(T|\overline{H}\right)} \tag{1.29}$$

Example 1.2.6 (Moneta bilanciata)**.** Abbiamo una moneta bilanciata e una sbilanciata che cade su testa con probabilità $3/4$. Si sceglie una moneta a caso e la si lancia tre volte; restituisce testa tutte e tre le volte. Quale è la probabilità che la moneta scelta sia quella bilanciata?
Se H è l'evento "testa tre volte" e B è l'evento "scelta la moneta bilanciata"; siamo interessati alla probabilità $\mathbb{P}(B|H)$. Ci risulta tuttavia più semplice trovare $\mathbb{P}(H|B)$ e $\mathbb{P}\left(H|\overline{B}\right)$ dato che aiuta sapere quale moneta consideriamo per calcolare la probabilità di tre teste. Questo suggerisce l'utilizzo del teorema di Bayes e della legge delle probabilità totali. Si ha

$$\begin{aligned}\mathbb{P}(B|H) &= \frac{\mathbb{P}(B)\cdot\mathbb{P}(H|B)}{\mathbb{P}(B)\cdot\mathbb{P}(H|B)+\mathbb{P}\left(\overline{B}\right)\cdot\mathbb{P}\left(H|\overline{B}\right)}\\ &= \frac{(1/2)\cdot(1/2)^3}{(1/2)\cdot(1/2)^3+(1/2)\cdot(3/4)^3}\\ &\approx 0.23\end{aligned}$$

Example 1.2.7 (Test di una malattia rara)**.** Un paziente è testato per una malattia che colpisce l'1% della popolazione. Sia D l'evento che "il paziente ha la malattia" e T il test è positivo (ossia suggerisce che il paziente abbia la malattia). Il paziente sottoposto al test risulta effettivamente positivo. Supponendo che il test sia accurato al 95%, ossia che $\mathbb{P}(T|D) = 0.95$ (la sensitività) ma anche che $\mathbb{P}\left(\overline{T}|\overline{D}\right) = 0.95$ (la specificità), qual è la probabilità che il paziente abbia effettivamente la malattia data la positività del test?
Applicando la formula di Bayes:

$$\begin{aligned}\mathbb{P}(D|T) &= \frac{\mathbb{P}(D)\,\mathbb{P}(T|D)}{\mathbb{P}(T)}\\ &= \frac{0.01\cdot 0.95}{\mathbb{P}(T)}\end{aligned}$$

$\mathbb{P}(T)$ non è così facile da ottenere (necessiterebbe di provare il test su tutta la popolazione), ma il teorema delle probabilità totali ci viene in soccorso:

$$\begin{aligned}\mathbb{P}(D|T) &= \frac{0.01\cdot 0.95}{\mathbb{P}(D)\,\mathbb{P}(T|D)+\mathbb{P}\left(\overline{D}\right)\mathbb{P}\left(T|\overline{D}\right)}\\ &= \frac{0.01\cdot 0.95}{0.01\cdot 0.95+0.99\cdot 0.05}\\ &\approx 0.16\end{aligned}$$

Pertanto vi è il 16% di probabilità che il paziente sia malato, anche se il test è positivo e lo strumento è affidabile: il fatto è che la malattia è estremamente rara e potrebbe essere un falso positivo, ossia un errore del test applicato (nella maggioranza dei casi) ad individui negativi.

1.3 Indipendent events

Definition 1.3.1 (Independence of a collection (even infinite) of events)**.** In general, a collection (even infinite) of events $\mathcal{E} = \{E_1, E_2, \dots\} \subset \mathscr{A}$ is composed

NB: Per rigo potrebbe essere unesercizio verificare indipendenza by independent events if, for *every finite* subset of the collection $\{E_1, \ldots, E_n\} \subset \mathcal{E}$, we have that

$$\mathbb{P}(E_1 \cap \ldots \cap E_n) = \mathbb{P}(E_1) \cdot \ldots \cdot \mathbb{P}(E_n)$$

Remark 29. Things become easier when events are independent but in reality this is rarely happening.

1.3.1 Two events

Example 1.3.1 (2 independent events, $A \perp\!\!\!\perp B$)**.** Applying definition 1.3.1 to a collection $\mathcal{E} = \{A, B\}$ of two events we say that A, B are independent if:

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)} \tag{1.30}$$

Example 1.3.2. Tossing a fair coin two times we have $\Omega = \{ht, hh, th, tt\}$ each outcome with probability $1/4$. Defining H_i = "i-th toss is a hed", we have $H_1 = \{ht, hh\}$, $H_2 = \{th, hh\}$; each has probability $\frac{1}{2}$. We have that $H_1 \cap H_2 = \{hh\}$ and since that

$$\mathbb{P}(H_1 \cap H_2) = \frac{1}{4} = \mathbb{P}(H_1) \cdot \mathbb{P}(H_2) = \frac{1}{2} \cdot \frac{1}{2}$$

the two events are independent: $H_1 \perp\!\!\!\perp H_2$. It makes sense since the result of the first outcome does not affect the next.

Important remark 9 (Conditional probability of independent events)*.* If A and B are independent and at the same time we have that $\mathbb{P}(B) > 0$, then we can redefine conditional probability as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A) \tag{1.31}$$

Thus undere these conditions $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Remark 30. Think independence in this latter way ($\mathbb{P}(A|B) = \mathbb{P}(A)$) may be clearer (knowing that B occurs or not it's the same, we don't need it), but we can't define independence $\mathbb{P}(A|B) = \mathbb{P}(A)$ out of the box because we're assuming $\mathbb{P}(B) > 0$

Proposition 1.3.1. *If* $\mathbb{P}(B) = 0 \vee \mathbb{P}(B) = 1$*, then* A *is independent of* B*,* $\forall A$*.*

Proof.

$$\begin{aligned}
\mathbb{P}(B) = 0 &\implies \mathbb{P}(A \cap B) = 0 = 0 \cdot \mathbb{P}(A) = \mathbb{P}(B) \cdot \mathbb{P}(A) \\
\mathbb{P}(B) = 1 &\implies \mathbb{P}(A \cap B) = \mathbb{P}(A) = 1 \cdot \mathbb{P}(A) = \mathbb{P}(B) \cdot \mathbb{P}(A)
\end{aligned}$$

□

Important remark 10. The previous results applies even if the two events seems to be somewhat connected. Eg suppose $\mathbb{P}(B) = 0$ and $B \subseteq A$. According to intuition these seems not to be independent because if B happens A happens as well. However logic and math definition/point of view can be different in practice.

Proposition 1.3.2 (Independence and complements)**.** *If A and B are independent then the following couples of events are independent as well: A and $\overline{B}$, $\overline{A}$ and B, $\overline{A}$* e *$\overline{B}$.*

Proof. Showing the first; suppose A,B are independent so $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. We want to prove

$$\mathbb{P}\left(A \cap \overline{B}\right) = \mathbb{P}(A)\,\mathbb{P}\left(\overline{B}\right)$$

We split $A = (A \cap B) \cup (A \cap \overline{B})$ in a disjoint union and sum its component probability:

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}\left(A \cap \overline{B}\right)$$

therefore

$$\begin{aligned} \mathbb{P}\left(A \cap \overline{B}\right) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) - \mathbb{P}(A)\,\mathbb{P}(B) \\ &= \mathbb{P}(A)\,[1 - \mathbb{P}(B)] \\ &= \mathbb{P}(A)\,\mathbb{P}\left(\overline{B}\right) \end{aligned}$$

Regarding $\overline{A}$ e B independence (and $\overline{A}$ e $\overline{B}$) it suffices to swap roles by negation/complement. □

1.3.2 n events

Example 1.3.3 (Independence of n events (finite set))**.** Again, applying definition 1.3.1 to a finite set of n events $A_1, \ldots, A_n \subset \Omega$, we have independence if for any subgroup of m events, $1 < m \leq n$, we have:

$$\mathbb{P}\left(\bigcap_{i=1}^{m} A_i\right) = \prod_{i=1}^{m} \mathbb{P}(A_i) \tag{1.32}$$

Example 1.3.4 (Independence and pairwise independence of 3 events)**.** E, F, G are independent if:

$$\begin{aligned} \mathbb{P}(E \cap F) &= \mathbb{P}(E)\,\mathbb{P}(F) \\ \mathbb{P}(E \cap G) &= \mathbb{P}(E)\,\mathbb{P}(G) \\ \mathbb{P}(F \cap G) &= \mathbb{P}(F)\,\mathbb{P}(G) \\ \mathbb{P}(E \cap F \cap G) &= \mathbb{P}(E)\,\mathbb{P}(F)\,\mathbb{P}(G) \end{aligned}$$

E, F, G are *pairwise* independent if the first three equation above holds.

Important remark 11. Generally speaking, n-wise independence implies $(n-1)$-wise of its components but viceversa does not hold; eg having *pairwise independence* of the above three events is not enough to prove their *independence*.

Important remark 12. In general given *any* collection $\mathcal{E} = \{E_1, E_2, \ldots\} \subset \mathscr{A}$ of events, it may be that $\mathbb{P}(E_i \cap E_j) = \mathbb{P}(E_i) \cdot \mathbb{P}(E_j)$, $\forall i \neq j$, but $\mathcal{E}$ is not independent. An example follows with three events.

Example 1.3.5. Throwing two coins ha $\Omega = \{tt, tc, ct, cc\}$. Following events are pairwise independent but not independent:

NB: Altro esempio, volendo, rigo lez 2023-09-21.

- $A =$ "first tail" $= \{th, tt\}$
- $B =$ "second tail" $= \{ht, tt\}$
- $C =$ "same result" $= \{hh, tt\}$

Infatti

$$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{2}{4} = \frac{1}{2}$$
$$\mathbb{P}(A \cap B) = \mathbb{P}(\{tt\}) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(B)$$
$$\mathbb{P}(A \cap C) = \mathbb{P}(\{tt\}) = \frac{1}{4} = \mathbb{P}(B)\mathbb{P}(C)$$
$$\mathbb{P}(B \cap C) = \mathbb{P}(\{tt\}) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(C)$$
$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\{tt\}) = \frac{1}{4} \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{8}$$

Point is: knowing what happened with A and B gives us complete information on C.

Important remark 13 (Independence and complements)*.* Similar to the two events case, for three events, if E, F, G are indipendent, then E is indipendent from any event formed by union/intersection/complement of F e G.

Example 1.3.6. E is independent from $F \cup G$ being:

$$\begin{aligned}
\mathbb{P}(E \cap (F \cup G)) &= \mathbb{P}((E \cap F) + (E \cap G)) \\
&= \mathbb{P}(E \cap F) + \mathbb{P}(E \cap G) - \mathbb{P}(E \cap F \cap G) \\
&= \mathbb{P}(E)\mathbb{P}(F) + \mathbb{P}(E)\mathbb{P}(G) - \mathbb{P}(E)\mathbb{P}(F \cap G) \\
&= \mathbb{P}(E)\left[\mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(F \cap G)\right] \\
&= \mathbb{P}(E)\mathbb{P}(F \cup G)
\end{aligned}$$

1.4 Esercizi rigo

Example 1.4.1 (Es rigo)**.** Stai viaggiando su un treno con un amico. Nessuno di voi ha il biglietto e il controllore vi ha beccato. Il controllore è autorizzato a infliggervi una punizione molto particolare. Vi porge una scatola contenente 9 cioccolatini identici, 3 dei quali avvelenati. Vi costringe a sceglierne uno a testa, a turno, e mangiarlo immediatamente.

1. Se scegli prima del tuo amico, qual è la probabilità che tu sopravviva?
2. Se scegli per primo e sopravvivi, qual è la probabilità che anche il tuo amico sopravviva?
3. Se scegli per primo e muori, qual è la probabilità che il tuo amico sopravviva?
4. E' nel tuo interesse far scegliere prima al tuo amico?

5. Se scegli per primo, qual è la probabilità che tu sopravviva, tenendo conto del fatto che il tuo amico resti in vita?

Se A="primo cioccolatino scelto è non avvelenato", e B="secondo scelto non avvelenato"

1. $\mathbb{P}(A) = 6/9$
2. $\mathbb{P}(B|A) = 5/8$
3. $\mathbb{P}(B|A^c) = 6/8$
4. $\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(A^c)\mathbb{P}(B|A^c) = \frac{6}{9}\frac{5}{8} + \frac{6}{8}\frac{3}{9} = \frac{6}{9}$ quindi non vi è vantaggio nello scegliere dopo il tuo amico
5. $\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)} = \ldots = \frac{5}{8}$; notiamo che $\mathbb{P}(A|B) = \mathbb{P}(B|A)$ in accordo con l'osservazione precedente, ossia che l'ordine della scelta non influenzi le probabilità di sopravvivenza

Example 1.4.2 (Rs rigo)**.** Un dado a sei facce non truccato viene lanciato due volte.

1. Scrivere lo spazio di probabilità dell'esperimento.
2. Supponiamo che B sia l'evento corrispondente al fatto che il risultato del primo lancio sia un numero non maggiore di 3, e supponiamo anche che C sia l'evento corrispondente al fatto che la somma dei due numeri ottenuti nei due lanci sia uguale a 6. Determinare le probabilità di B e C, e le probabilità condizionali di C dato B, e di B dato C.

Lo spazio di probabilità in questo esperimento è la tripla $(\Omega, \mathscr{A}, \mathbb{P})$, dove:

- $\Omega = \{(1,1), \ldots (6,6)\}$
- $\mathscr{A} = \mathcal{P}(\Omega)$
- ciascun punto in Ω ha uguale probabilità di successo, ossia $\mathbb{P}((i,j)) = 1/36$

Per il secondo punto:

- $B=$ primo lancio $\leq 3 = \{(1,1), \ldots, (1,6), (2,1) \ldots, (2,6), (3,1) \ldots, (3,6)\}$ pertanto $\mathbb{P}(B) = \frac{18}{36}$
- $C =$ somma $= 6 = \{(1,5), (5,1), (2,4), (4,2), (3,3)\}$, $\mathbb{P}(C) = \frac{5}{36}$
- si ha che $C \cap B = \{(1,5), (2,4), (3,3)\}$ quindi $\mathbb{P}(C|B) = \frac{\mathbb{P}(C \cap B)}{\mathbb{P}(B)} = \frac{3/36}{18/36} = \frac{1}{6}$
- $\mathbb{P}(B|C) = \frac{3/36}{5/36} = \frac{3}{5}$

Example 1.4.3. Una scatola contiene n palline di cui k bianche e $n-k$ nere, dove $1 \leq k \leq n-1$. faccio due estrazioni senza reinserimento. Calcolare la

probabilità che la prima sia bianca dato che la seconda estratta è nera.
Si ha

$$\begin{aligned}\mathbb{P}\left(\text{1b}|\text{2n}\right) &= \frac{\mathbb{P}\left(\text{1b e 2n}\right)}{\mathbb{P}\left(\text{2n}\right)} = \frac{\mathbb{P}\left(1b\right)\cdot\mathbb{P}\left(2n|1b\right)}{\mathbb{P}\left(1b\right)\cdot\mathbb{P}\left(2n|1b\right)+\mathbb{P}\left(1n\right)\cdot\mathbb{P}\left(2n|1n\right)+} \\ &= \frac{\frac{k}{n}\frac{n-k}{n-1}}{\frac{k}{n}\frac{n-k}{n-1}+\frac{n-k}{n}\frac{n-k-1}{n-1}} = \frac{k(n-k)}{k(n-k)+(n-k)(n-k-1)} \\ &= \frac{k}{k+n-k-1} = \frac{k}{n-1}\end{aligned}$$

Example 1.4.4. Considerati 3 lanci di una moneta:

1. costruire lo spazio di probabilità che descrive il numero di teste
2. stabilire se gli eventi $A = \{\text{ottengo almeno una testa}\}$ $B = \{\text{ottengo almeno una croce}\}$ sono indipendenti
3. calcolare $\mathbb{P}\left(A \cup B^c\right)$ e $\mathbb{P}\left(A|B^c\right)$

Si ha che

1. $(\Omega, \mathscr{P}(\Omega), \mathbb{P})$ definito a partire da $\Omega = \{ttt, ttc, tct, ctt, tcc, ctc, cct, ccc\}$ e $(X(\Omega), \mathscr{P}(X(\Omega)), \nu)$ $X(\Omega) = \{0, 1, 2, 3\}$ e $\nu(E) = \mathbb{P}\left(X^{-1}(E)\right)$ con, ad esempio:
$$\begin{aligned}\nu(t0) &= \mathbb{P}\left(\{ccc\}\right) = \frac{1}{8} \\ \nu(t1) &= \mathbb{P}\left(\{ttc, tct, ctt\}\right) = \frac{3}{8} \\ \nu(t2) &= \mathbb{P}\left(\{tcc, ctc, cct\}\right) = \frac{3}{8} \\ \nu(t3) &= \mathbb{P}\left(\{ttt\}\right) = \frac{1}{8}\end{aligned}$$
2. i due eventi sono indipendenti se
$$\mathbb{P}\left(A \wedge B\right) = \mathbb{P}\left(A\right)\cdot\mathbb{P}\left(B\right)$$
si ha che
$$\mathbb{P}\left(A \wedge B\right) = \mathbb{P}\left(\text{almeno una testa e almeno una croce}\right) = \mathbb{P}\left(\{ttc, tct, ctt, tcc, ctc, cct\}\right) = \frac{6}{8} = \frac{3}{4}$$
$$\begin{aligned}\mathbb{P}\left(A\right) &= 1 - \mathbb{P}\left(\{ccc\}\right) = \frac{7}{8} \\ \mathbb{P}\left(B\right) &= 1 - \mathbb{P}\left(\{ttt\}\right) = \frac{7}{8} \\ \frac{3}{4} &\neq \frac{7}{8}\frac{7}{8} = \frac{49}{64}\end{aligned}$$
ergo i due eventi non sono indipendenti
3. si ha che $B^c = \{ttt\}$ e $A \cap B^c = \{ttt\}$
$$\mathbb{P}\left(A \cup B^c\right) = \mathbb{P}\left(A\right) + \mathbb{P}\left(B^c\right) - \mathbb{P}\left(A \cup B^c\right) = \frac{7}{8} + \frac{1}{8} - \mathbb{P}\left(\{ttt\}\right) = \frac{7}{8} + \frac{1}{8} - \frac{1}{8}$$
$$\begin{aligned}&\mathbb{P}\left(A|B^c\right) \\ &= \frac{\mathbb{P}\left(A \cap B^c\right)}{\mathbb{P}\left(B^c\right)} = \frac{1/8}{1/8} = 1\end{aligned}$$

Example 1.4.5. Si consideri $\Omega = \{1,2,3,4,5,6,7,8,9,10\}$ con $\mathbb{P}(\{i\}) = \frac{i}{10}$ $\forall i \in \Omega$:

1. stabilire se gli eventi $A = \{\text{multipli di } 2\}$ e $B = \{\text{multipli di } 3\}$ sono indipendenti

2. dato $C = \{< 6\}$ calcolare $\mathbb{P}(A|C)$ e $\mathbb{P}(B|C)$

Si ha

1. $$\begin{aligned} \mathbb{P}(A) &= \frac{2}{55} + \frac{4}{55} + \frac{6}{55} + \frac{8}{55} + \frac{10}{55} = \frac{30}{55} \\ \mathbb{P}(B) &= \frac{3}{55} + \frac{6}{55} + \frac{9}{55} = \frac{18}{55} \\ \mathbb{P}(A \cap B) = \mathbb{P}(6) &= \frac{6}{55} \neq \mathbb{P}(A) \cdot \mathbb{P}(B) \end{aligned}$$

   quindi gli eventi non sono indipendenti

2. $$\begin{aligned} \mathbb{P}(C) &= \frac{1+2+3+4+5}{55} = \frac{15}{55} \\ \mathbb{P}(A|C) &= \frac{\mathbb{P}(A \cap C)}{\mathbb{P}(C)} = \frac{\frac{2+4}{55}}{\frac{15}{55}} = \frac{6}{15} = \frac{2}{5} \\ \mathbb{P}(B|C) &= \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)} = \frac{\frac{3}{55}}{\frac{15}{55}} = \frac{1}{5} \end{aligned}$$

Example 1.4.6. Una scatola contiene due palline bianche e una nera. Estraggo una pallina a caso: se bianca lancio un dado e registro il risultato ottenuto, se è nera lancio due dadi e registro il minore dei due. Calcolare la probabilità di ottenere 2 al termine dell'esperimento. Si ha

$$\begin{aligned} \mathbb{P}(2) &= \mathbb{P}(2|\text{bianca}) \cdot \mathbb{P}(\text{bianca}) + \mathbb{P}(2|\text{nera}) \cdot \mathbb{P}(\text{nera}) \\ &= \frac{2}{3} \cdot \mathbb{P}(\{2\}) + \frac{1}{3} \cdot \mathbb{P}(\{(2,2),(2,3),(2,4),(2,5),(2,6),(3,2),(4,2),(5,2),(6,2)\}) \\ &= \frac{2}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{9}{36} = \frac{1}{9} + \frac{1}{12} \\ &= \frac{7}{36} \end{aligned}$$

Example 1.4.7 (Esercizio esame rigo)**.** Da un'urna contenente 5 palline bianche e 4 nere effettuiamo estrazioni senza reinserimento. Si determini la probabilità di ottenere una pallina bianca alla terza prova.

$$\begin{aligned}\mathbb{P}(3b) &= \mathbb{P}(3b|1b \cap 2b) \cdot \mathbb{P}(1b \cap 2b) + \mathbb{P}(3b|1n \cap 2b) \cdot \mathbb{P}(1n \cap 2b) + \ldots \\ &\ldots + \mathbb{P}(3b|1b \cap 2n) \cdot \mathbb{P}(1b \cap 2n) + \mathbb{P}(3b|1n \cap 2n) \cdot \mathbb{P}(1n \cap 2n)\end{aligned}$$

$$\mathbb{P}(1b \cap 2b) = \frac{5}{9}\frac{4}{8}$$

$$\mathbb{P}(1n \cap 2b) = \frac{4}{9}\frac{5}{8}$$

$$\mathbb{P}(1b \cap 2n) = \frac{5}{9}\frac{4}{8}$$

$$\mathbb{P}(1n \cap 2n) = \frac{4}{9}\frac{3}{8}$$

$$\mathbb{P}(3b) = \frac{5}{9}\frac{4}{8}\frac{3}{7} + \frac{4}{9}\frac{5}{8}\frac{4}{7} + \frac{5}{9}\frac{4}{8}\frac{4}{7} + \frac{4}{9}\frac{3}{8}\frac{5}{7} = \ldots = \frac{5}{9}$$

Chapter 2

Random variables

2.1 Intro

2.1.1 Random variables linking probability spaces

Remark 31. A probability space $(\Omega, \mathscr{A}, \mathbb{P})$ is a particular measurable space.

Definition 2.1.1 (Measurable space)**.** A pair $(S, \mathscr{B})$, composed by a set S and a σ-field $\mathscr{B}$ defined on it.

Definition 2.1.2 (Random variable X)**.** A random variable is a *measurable* function $X : \Omega \to S$ which creates a mapping between a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ and a measurable space $(S, \mathscr{B})$ by connecting the first two sets.

Definition 2.1.3 (Measurability)**.** Being X *measurable* means that

$$\forall E \in \mathscr{B}, \exists X^{-1}(E) = \{\omega \in \Omega : X(\omega) \in E\} \in \mathscr{A}, \tag{2.1}$$

In words if I take any event of $\mathscr{B}$, there's a corresponding event in $\mathscr{A}$ that does produce it through X. $X^{-1}(E)$ is called inverse image of the event E.

Remark 32. In practice in this course the measurable spaces $(S, \mathscr{B})$ of interest will be:

- $(\mathbb{R}, \beta(\mathbb{R}))$: X is called real or univariate random variable, and so is a function of type $X : \Omega \to \mathbb{R}$
- $(\mathbb{R}^n, \beta(\mathbb{R}^n))$: X is called n-variate random variable or n-dimensional random vector, a function of type $X : \Omega \to \mathbb{R}^n$

Remark 33 (Interpretation). The interpretation of rv is the following: one makes the experiment and see the resulting outcome $\omega \in \Omega$. Then after observing ω, $X(\omega)$ make a measurement on the outcome.

Remark 34. While the random variable is a *deterministic* mapping, the random part comes from the experiment.

Definition 2.1.4 (Rv support)**.** It's the image $X(\Omega)$, the set of possible mappings, denoted by $R_X = \{x_1, x_2, \dots\}$

Example 2.1.1 (Two coin throws)**.** Two coin throws can generate the following $\Omega = \{tt, th, ht, hh\}$. On this one we can define $X =$ "sum of heads as follows"

$$X(tt) = 2;\ X(th) = 1;\ X(ht) = 1;\ X(hh) = 0;$$

Finally we have that the support is $R_X = \{0, 1, 2\}$.

Definition 2.1.5 (Probability distribution of X (and second probability space))**.** Given a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, a measurable space $(S, \mathscr{B})$, and a random variable $X : \Omega \to S$ connecting the twos, we can define a further probability space $(S, \mathscr{B}, \nu)$, where the added probability function $\nu : \mathscr{B} \to [0, 1]$ is defined, using $\mathbb{P}$, in the following way:

$$\nu(E) = \mathbb{P}\left(X^{-1}(E)\right) = \mathbb{P}\left(\omega \in \Omega : X(\omega) \in E\right) = \mathbb{P}\left(X \in E\right), \quad \forall E \in \mathscr{B} \tag{2.2}$$

ν is called *probability distribution* of X.

Example 2.1.2. If the experiment is to draw one person from a class, $\Omega = \{everyone\}$,while the random variable X could be height, so if Luca is extracted ($\omega =$ Luca), then X(Luca) $= 1.78$.
Distribution function ν of X is:

$$\nu(E) = \mathbb{P}\left(X \in E\right) = \mathbb{P}\left(\text{quelli di noi la cui altezza cade in } E\right)$$

Eg, if $E = (190, 195]$ and only Paolo and Francesca have an height such as that, then

$$\nu(B) = \mathbb{P}\left(\text{Paolo}\right) + \mathbb{P}\left(\text{Francesca}\right)$$

Important remark 14 (Motivation for measurability request)*.* A possible motivation for requiring measurability of X, as we did, is the need to define its distribution ν. Suppose we don't require X to be measurable; thus can be that:

$$\exists E \in \mathscr{B} : X^{-1}(E) \notin \mathscr{A}$$

there's an event of $\mathscr{B}$ with no corresponding event in $\mathscr{A}$.
In that case $X^{-1}(E)$ does not belong to the domain of $\mathbb{P}$ and thus we cannot define/write $\nu(E) = \mathbb{P}\left(X^{-1}(E)\right) = \mathbb{P}\left(X \in E\right)$.
Therefore the need to define ν forces us to require X to be measurable.

Important remark 15 (Notation)*.* If we say:

- $X \sim \nu$ means that ν is the probability distribution of the rv X; for istance considering a real random variable $X : \Omega \to \mathbb{R}$, if we say $X \sim \mathrm{N}\left(0, 1\right)$ we are stating that probability distribution of X is standard normal;
- $X \sim Y$ means that X and Y have the same distribution (whatever it is).

2.1.2 Discrete and continuous rvs

Remark 35*.* Queste sotto sono definizioni utili per fissare i concetti (le definizioni Rigo style son sotto credo)

Definition 2.1.6 (Discrete rv)**.** Rv which cardinality of support is finite or numerable (1-to-1 with $\mathbb{N}$.)

Example 2.1.3. Head count in two coin throwing is discrete since Card $(R_X) = |\{0, 1, 2\}| = 3$.

Definition 2.1.7 (Continuous rv)**.** Rv which cardinality of support is not numerable (1-to-1 with $\mathbb{R}$).

Example 2.1.4. Numbers of minutes T of bulb lifetime is continue because $R_T = \{t \in \mathbb{R} : t > 0\}$

2.2 Distribution (and other) functions

Remark 36. In order to study random variables, an important concept is distribution function (which is the unifying one for continuous and discrete random variables); here we summarize/prove some results.

Important remark 16 (Jargon). When it's said distribution function we mean the cumulative distribution function.

Definition 2.2.1 (Distribution function)**.** If X is a real valued rv, its distribution function $F : \mathbb{R} \to \mathbb{R}$ is defined as

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \in (-\infty, x]) = \nu((-\infty, x]), \quad \forall x \in \mathbb{R}$$

Remark 37. For any distribution function F, exists *one and only one* probability measure ν on $\beta(\mathbb{R})$ such that $F(x) = \nu((-\infty, x])$, $\forall x \in \mathbb{R}$ and viceversa. **NB**: considerazioni dalla triennale
Distribution functions are in bijection with probability measure on $\beta(\mathbb{R})$; thanks to this, in order to assign a ν on $\beta(\mathbb{R})$ it is enough to assign a distribution function F (once chosen F to it corresponds one and only one ν). And in practical terms, choosing a F is easier than assigning a ν

Proposition 2.2.1 (Fundamental/characterizing properties)**.** *The properties* characterizing *distribution functions are*

1. $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$,
2. *F is not decreasing: if $y > x$ then $F(y) \geq F(x)$;*
3. *F is right continuous $F(x) = \lim_{y\to x^+} F(y)$, $\forall x \in \mathbb{R}$*

Important remark 17. Any function $F : \mathbb{R} \to \mathbb{R}$ which satisfies the three properties is a distribution function, that is, there exists a random variable X such that $F(x) = \mathbb{P}(X \leq x)$, $\forall x \in \mathbb{R}$.

Proposition 2.2.2. *Supposing we want to evaluate the probability of a certain point* $\mathbb{P}(X = x) = \nu(\{x\}) = \nu((-\infty, x] \setminus (-\infty, x)) = \nu((-\infty, x]) - \nu((-\infty, x))$. *The formula is*

$$\mathbb{P}(X = x) = F(x) - F(x^-) \qquad \textit{(jump of } F \textit{ at } x\textit{)} \tag{2.3}$$

where $F(x^-) = \lim_{y\to x^-} F(y)$ *meaning limit with* $y \to x$ *from the left.*

Proof. To prove this, recall (props 1.2.9 and 1.2.10) that for any probability measure $\mathbb{P}$

- if $A_1 \subseteq A_2 \subseteq \ldots$ is a increasing sequence of events, $\mathbb{P}(\cup_n A_n) = \lim_n \mathbb{P}(A_n)$

- if $A_1 \supseteq A_2 \supseteq \dots$ is a decreasing sequence of events, $\mathbb{P}(\cap_n A_n) = \lim_n \mathbb{P}(A_n)$

Now suppose we want to evaluate

$$\mathbb{P}(X < x) = \mathbb{P}\left(\bigcup_{n=1}^{+\infty} \left\{X \leq x - \frac{1}{n}\right\}\right)$$

where we go nearer and nearer to x as n increases. These events are an increasing sequence of events, so

$$\begin{aligned}\mathbb{P}(X < x) &= \mathbb{P}\left(\bigcup_{n=1}^{+\infty} \left\{X \leq x - \frac{1}{n}\right\}\right) = \lim_{n\to+\infty} \mathbb{P}\left(X \leq x - \frac{1}{n}\right) = \lim_{n\to+\infty} F\left(X \leq x - \frac{1}{n}\right)\\ &= F(x^-)\end{aligned}$$

Finally in order to evaluate $\mathbb{P}(X = x)$ we have:

$$\mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = F(x) - F(x^-)$$

□

Important remark 18. As a consequences of 2.3, considering the set:

$$\{x \in \mathbb{R} : \mathbb{P}(X = x) > 0\} = \{x \in \mathbb{R} : \nu(\{x\}) > 0\} = \left\{x \in \mathbb{R} : F(x) > F(x^-)\right\}$$

- this set is *empty*, if the function is continuous: in other words the distribution function is *continuous* if and only if the jump is 0 at each point or in other words

$$F \text{ is continuous} \iff \mathbb{P}(X = x) = 0, \forall x \in R$$

- its cardinality can *at most be countable* (for a calculus result): it can be countable (eg for Poisson, negative binomial etc) or can be finite as well. But can't be uncountable.

2.2.1 Types of RVs

Important remark 19 (RV types). Real random variables can be *discrete*, *singular continuous* (we can ignore it) or *absolutely continuous*. The following result is theoretically important.

Proposition 2.2.3. *If ν is any probability measure on $\beta(\mathbb{R})$, the exists a* unique *triplets (a, b, c) such that:*

- $a, b, c \geq 0$
- $a + b + c = 1$
- $\nu = a\nu_1 + b\nu_2 + c\nu_3$

where ν_1 is discrete probability measure, ν_2 is singular continuous probability measure, ν_3 is absolutely continuous probability measure.

Proof. We skip it. □

Important remark 20. Thanks to the above thm

- if we are able to describe a discrete probability measure, a singular continuous probability measure and an absolute continuous probability measure, we are able to describe ANY probability measure on $\beta(\mathbb{R})$.
- any ν can be written as this mix of this three kind of rv. Clearly, eg

$$\begin{aligned} a = 1, b = c = 0 &\implies \nu = \nu_1 \text{ is discrete} \\ c = 1, a = b = 0 &\implies \nu = \nu_3 \text{ is absolutely continuous} \end{aligned}$$

This is the reason to focus on the three types, of which *only discrete and absolutely continuous are of interest for practical applications.*

Important remark 21. In this course we speak indifferently like:

$$X \text{ is discrete} \iff \nu \text{ is discrete} \iff F \text{ is discrete}$$

Similarly for singular and absolutely continuous rv

2.2.2 Discrete rvs

Definition 2.2.2 (Discrete rv)**.** X is discrete if and only if $\exists B \subset \mathbb{R}$, with B finite or countable such that $\mathbb{P}(X \in B) = 1$.

Example 2.2.1 (Examples of discrete rvs)**.** Some are:

- the degenerate rv, δ_a, where $B = \{a\}$ and thus $P(X \in \{a\}) = 1$; its distribution function is defined as

$$F(x) = \mathbb{P}(X \leq x) = \begin{cases} 1 & x \geq a \\ 0 & x < a \end{cases}$$

- binomial, then $B = \{0, 1, \ldots, n\}$;
- Poisson, $B = \{0, 1, \ldots\}$.

2.2.3 Singular continuous rvs

Remark 38. As we have said probability is a measure. In general

Definition 2.2.3. A measure m is a function that, considered a single set X and a *finite* or *numerable* set of incompatible events $X_1, X_2, \ldots$

$$m(X) \geq 0, \quad \forall X \tag{2.4}$$

$$X_i \cap X_j = \emptyset, \forall i \neq j \implies m\left(\left\{\bigcup_i X_i\right\}\right) = \sum_i m(X_i) \tag{2.5}$$

Important remark 22. The *Lebesgue measure* in $\mathbb{R}$ is the only measure on $\beta(\mathbb{R})$ that has this property, applied to an interval:

$$m(a, b] = b - a, \qquad \forall a < b \tag{2.6}$$

where m is the Lebesgue measure of the interval. Regarding the measure a point, countable and uncountable sets (the real line) Lebesgue measure

$$
\begin{aligned}
&m(\{x\}) = 0, \qquad \forall x \in \mathbb{R} \\
&m(X) = \sum_{x \in X} m(\{x\}) = \sum_{x \in X} 0 = 0 \qquad \forall X \subset \mathbb{R} : X \text{ is countable} \\
&m(\mathbb{R}) = +\infty
\end{aligned}
$$

Definition 2.2.4 (Singular continuous rvs)**.** X is a singular continuous random variable if both

1. the distribution function F is continuous
2. its first derivative is null ($F'(x) = 0$) *almost everywhere* with respect to the Lebesgue measure m (written concisely as "m.a.e."):

$$m(\{x \in \mathbb{R} : F'(x) \neq 0\}) = 0$$

Important remark 23. Note that

- first derivative $\neq 0$ when it doesn't exists (eg left and right limit are different) or exists but is not 0;
- for this kind of rv, distribution may not be differentiable or with derivative 0 at every point
- however these $F'(x) \neq 0$ points are a finite or at most countable set of points.

Remark 39. For *discrete* RVs actually is the same: $F'(x) = 0$ mae (think step F functions) given that:

$$m(\{x \in \mathbb{R} : F'(x) \neq 0\}) = m(\{\text{jump points of } F\ \}) = 0$$

as the set {jump points of F } is finite or countable.
However, if X is discrete, F is certainly discontinuous.

Remark 40. These variables

- seems to be a somewhat hybrid between discrete and absolutely continuous rv (since have characteristic from both the distribution), that is $F'(x) = 0$ mae from the discrete RV, and continuous F from absolutely continuous;
- are not usually used for describing real phenomena, and we will not consider them in what follows.

2.2.4 Absolutely continuous rvs

Example 2.2.2. eg exponential, beta, uniform, normal ...

Definition 2.2.5 (Absolutely continuous rv)**.** X is absolutely continuous if and only if exists a function $f : \mathbb{R} \to \mathbb{R}$, called density, such that:

1. $f \geq 0$ (density is non negative)

2. f is integrable
3. the distribution function at point x can be written as (Lebesgue[1]) integral of density function f

$$F(x) = \int_{-\infty}^{x} f(t)\ \mathrm{d}t, \qquad \forall x \in \mathbb{R}$$

Important remark 24 (Probability of an event). With absolutely continuous random variable the probability of an event $E \in \beta(\mathbb{R})$ is

$$\mathbb{P}\left(X \in E\right) = \int f(t)\mathbb{1}_E(t)\ \mathrm{d}t = \int_E f(t)\ \mathrm{d}t, \quad \forall E \in \beta(\mathbb{R})$$

where we denoted $\mathbb{1}_E(t)$ as the indicator function of the set E, that is

$$\mathbb{1}_E(t) = \begin{cases} 1, & t \in E \\ 0, & t \notin E \end{cases}$$

Important remark 25. Some properties for these RVs:

- $F' = f$ m.a.e:
$$m(\{x \in \mathbb{R} : f(x) \neq F'(x)\}) = 0$$
that is supposing we collect all the points where density doesn't equal the derivative of the distribution function, then they can differ at most in a countable set of $x \in \mathbb{R}$
- from the previous point, if f_1 and f_2 are both densities of the same RV X, can we say $f_1 = f_2$, that density is *unique*?
Since f_1 and f_2 are densities, $f_1 = F'$ mae and $f_2 = F'$ mae, so we have $f_1 = F' = f_2$ m.a.e that is
$$m(\{x \in \mathbb{R} : f_1(x) \neq f_2(x)\}) = 0$$
so the density f is *almost everywhere unique* (can be different but at most in a countable set of points).

Example 2.2.3. Regarding the last property, consider a standard normal $X \sim \mathrm{N}\left(0, 1\right)$ which is absolutely continuous, having density

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

Now we define a new density which is different from the standard normal in a countable set $\mathbb{Q}$ of points[2]:

$$g(x) = \begin{cases} f(x) & \text{if } x \notin \mathbb{Q} \\ 1 + \sin\left(\log|x| + 3\right), & \text{if } x \in \mathbb{Q} \end{cases}$$

[1]In generale l'integrale di Lebesgue è una generalizzazione dell'integrale di Riemann e coincide con quest'ultimo sotto condizioni abbastanza generali

[2]$\mathbb{Q}$ has two properties: it's a *countable* set and it's *dense*, that is $\forall a, b \in \mathbb{Q}, \exists q \in \mathbb{Q}$ such that $a < q < b$

We can say that:

$$m(\{f \neq g\}) \leq \underbrace{m(\mathbb{Q}) = 0}_{\text{being countable}}$$

Therefore the function f agrees with g m.a.e.
Thus f and g are *both* densities for X standard normal.

Remark 41. Another important property of absolutely continuous rvs is the following characterization

Theorem 2.2.4 (Absolutely continuous RV characterization)**.** *X is absolutely continuous if and only if, for every set (event) with lebesgue measure 0, this set has probability 0*

$$X \text{ is absolutely continuous} \iff \begin{cases} \mathbb{P}\left(X \in A\right) = 0 \\ \forall A \in \beta(\mathbb{R}), \textit{such that } m(A) = 0 \end{cases}$$

Remark 42. Quindi non solo punti singoli hanno probabilità nulla ma anche un insieme finito o al più numerabile la ha.

2.3 OLD: Functions of random variables

2.3.1 Discrete rvs: PMF, CDF

Definition 2.3.1 (Probability mass function)**.** Given a rv $X : \Omega \to \mathbb{R}$, PMF is a function $p : \mathbb{R} \to \mathbb{R}$ taking the outcome of the rv and giving its probability

$$p_X(x) = \mathbb{P}\left(X = x\right) = \begin{cases} \mathbb{P}\left(X(\omega) = x\right) & \text{se } x \in X(\Omega) \\ 0 & \text{se } x \in \mathbb{R} \setminus X(\Omega) \end{cases} \tag{2.7}$$

Proposition 2.3.1 (Valid PMF)**.** *If X is a discrete rv with support $X(\Omega) = \{x_1, x_2, \ldots\}$, a valid PMF p_X satisfies:*

$$p_X(x) \geq 0, \quad \forall x \in \mathbb{R} \tag{2.8}$$

$$\sum_{x \in \mathbb{R}} p_X(x) = 1 \tag{2.9}$$

Proof. Il primo criterio deve esser valido dato che la probabilità è non negativa. Il secondo deve essere valido dato che gli eventi $X = x_1, X = x_2, \ldots$ sono disgiunti e X dovrà assumere pur qualche valore:

$$\begin{aligned} \sum_{x \in \mathbb{R}} p_X(x) = \sum_{x \in X(\Omega)} p_X(x) = \sum_j \mathbb{P}\left(X = x_j\right) = \mathbb{P}\left(\bigcup_j \{X = x_j\}\right) \\ = \mathbb{P}\left(X = x_1 \text{ or } X = x_2 \ldots\right) = 1 \end{aligned}$$

□

Example 2.3.1. In two coins throwing 2.1.1

$$p_X(X = 0) = 1/4$$
$$p_X(X = 1) = 1/2$$
$$p_X(X = 2) = 1/4$$

and $p_X(x) = 0$ for $x \notin \{0, 1, 2\}$.

Definition 2.3.2 ((Cumulative) distribution function (CDF))**.** Given a discrete rv X its defined as:

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_j \in X(\Omega): x_j \leq x} p_X(x_j) \tag{2.10}$$

Remark 43 (Function shape)*.* If X is discrete, $F_X(x)$ has starway shape with finite or numerable steps on values of the support $x_1, x_2, \ldots$: the step height is $p_X(x_1), p_X(x_2), \ldots$.

Proposition 2.3.2 (Valid CDF)**.** *If X is a discrete rv with support $X(\Omega) = \{x_1, x_2, \ldots\}$, a valid CDF F_X must satisfy*

$$x_1 \leq x_2 \implies F_X(x_1) \leq F_X(x_2) \tag{2.11}$$

$$\lim_{x \to x_j^+} F_X(x) = F_X(x_j) \quad \textit{(right continuous)} \tag{2.12}$$

$$\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to +\infty} F_X(x) = 1 \tag{2.13}$$

Proof. La prima è giustificata dal fatto che dato che, dato che l'evento $\{X \leq x_1\}$ si verifica sempre quando si verifica $\{X \leq x_2\}$ allora $\mathbb{P}(X \leq x_1) \leq \mathbb{P}(X \leq x_2)$. La continuità da destra deriva dall'aver definito $F_X(x_0)$ come $\mathbb{P}(X \leq x_0)$ (coerentemente con la letteratura internazionale prevalente); altri autori definiscono $F_X(x_0) = \mathbb{P}(X < x_0)$, il che implica la continuità da sinistra.
Per la terza, dato che $F_X(x_{min}) = 0$ con $x_{min} = min(x_1, x_2, \ldots)$ e $-\infty < x_{min}$ allora per la prima proprietà si ha che $F(-\infty) \leq 0$, ma non potendo una probabilità esser negativa, sarà nulla, dunque si conclude che $\lim_{x \to -\infty} F_X(x) = 0$. Altresì sfruttando sempre il fatto che $\{X = x_j\}$ sono eventi indipendenti

$$\lim_{x \to +\infty} F_X(x) = \sum_{x_j \in X(\Omega)} p_X(x_j) = 1$$

□

Example 2.3.2. Dato l'esperimento lancio di due dati, l'evento X somma degli esiti ha PMF e CMF riportate in figura 2.1. Ad esempio $\mathbb{P}(X = 2) = \mathbb{P}(\{1, 1\}) = (\frac{1}{6})^2 = 1/36 \approx 0.02778$. I "salti" nella CDF sono di entità pari alla PMF

2.3.2 Continuous rvs: PDF, CDF

Remark 44*.* PDF is the equivalent of PMF, CDF the same.

Definition 2.3.3 ((Probability) density function (PDF))**.** If X is a continuous rv density is a $f : \mathbb{R} \to \mathbb{R}$, $f_X(x)$ such as, considered $X \in B \subseteq \mathbb{R}$:

$$\mathbb{P}(X \in B) = \int_{x \in B} f_X(x) \, \mathrm{d}x \tag{2.14}$$

Eg, if $a, b \in \mathbb{R}$, $a < b$:

$$\mathbb{P}(X \in [a, b]) = \int_a^b f_X(x) \, \mathrm{d}x \tag{2.15}$$

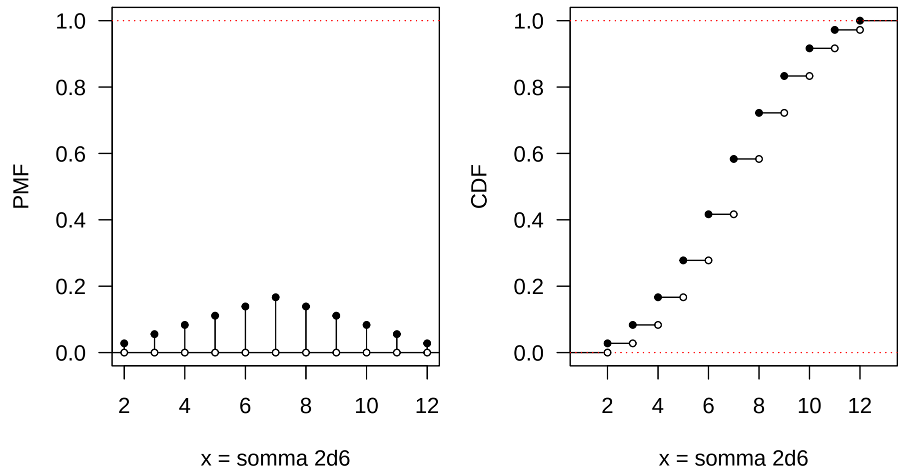


Figure 2.1: Somma del lancio di due d6

Proposition 2.3.3 (Valid PDF)**.** *Must satisfy*

$$f_X(x) \geq 0 \tag{2.16}$$

$$\int_{-\infty}^{\infty} f_X(t)\, \mathrm{d}t = 1 \tag{2.17}$$

Proof. Il primo criterio è necessario perché la probabilità è non negativa: se $f_X(x_0)$ fosse negativa, allora potremmo integrare su un piccolo intorno di x_0 e ottenere una probabilità negativa.
Il secondo criterio è necessario dato che la X, variabile quantitativa, deve avere un esito che sta in $\mathbb{R}$. □

Remark 45. Differently from the discrete case (where PMF can't be more than 1) pdf can be more than 1, as long as integral sums on $\mathbb{R}$ sums up to 1.

Definition 2.3.4 ((Cumulative) distribution function (CDF))**.** If X is a continuous rv, it's the function $F : \mathbb{R} \to \mathbb{R}$ defined as:

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\, \mathrm{d}t \tag{2.18}$$

Proposition 2.3.4 (Valid CDF)**.** *It must satisfy*

$$x_1 \leq x_2 \implies F_X(x_1) \leq F_X(x_2) \tag{2.19}$$

$$\lim_{x \to x_0^+} F_X(x) = F_X(x_0) \quad \textit{(continuità da destra)} \tag{2.20}$$

$$\lim_{x \to -\infty} F_X(x) = 0 \qquad \lim_{x \to +\infty} F_X(x) = 1 \tag{2.21}$$

Example 2.3.3 (Esempio crash course)**.** Let's check if

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x} & x \geq 0 \end{cases}$$

is a distribution function. We have

1. $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = \lim_{x\to+\infty} 1 - e^{-x} = 1$, so check for the first

2. for $y > x$ we must show that $F(y) \geq F(x)$ to ensure non decreasing nature. Let's check the sign of $F(y) - F(x)$ (since if $F(y) - F(x) \geq 0$ then $F(y) \geq F(x)$): we have

$$1 - e^{-y} - 1 + e^{-x} = e^{-x} - e^{-y} \overset{(1)}{\geq} 0$$

with (1) since $e^{-y} < e^{-x}$ given that $y < x$

3. because $F(x)$ is continuous, it is also right continuous

So yes, $F(x)$ is a CDF ($X \sim \text{Exp}(1)$).

Remark 46 (Probability calculation with CDF). If we know CDF we can evaluate probability of an interval $a \leq X \leq b$, $a, b \in \mathbb{R}$ as follows:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a)$$

Remark 47 (Probability of a single value). A differenza delle variabili discrete, nel caso continuo si ha che:

$$\mathbb{P}(X = a) = \int_a^a f_X(x)\,\mathrm{d}x = F_X(a) - F_X(a) = 0$$

Intuitively, if there are infinite outcomes probability of each of them is null.

Remark 48 (Irrilevance of extremes of integration). For the same reason $a, b \in \mathbb{R}$, $a < b$:

$$\mathbb{P}(X \in [a, b]) = \mathbb{P}(X \in (a, b]) = \mathbb{P}(X \in [a, b)) = \mathbb{P}(X \in (a, b)) = \int_a^b f_X(x)\,\mathrm{d}x$$

Example 2.3.4 (Logistic rv)**.** Logistic random variable, plotted in figure 2.2, is defined by:

$$F(x) = \frac{e^x}{1 + e^x}; \quad f(x) = \frac{e^x}{(1 + e^x)^2}$$

```
flogis <- function(x) exp(x)/(1 + exp(x))^2
Flogis <- function(x) exp(x)/(1 + exp(x))
par(mfrow = c(1, 2), mar = c(5,4,1,1))
plot_fun(flogis, from = -4, to = +4, ylim = c(0, 1),
         cartesian_plane = FALSE,
         ylab = 'PDF', las = 1)
abline(h = c(0), col = 'red', lty = 'dotted')
plot_fun(Flogis, from = -4, to = +4, ylim = c(0, 1),
         cartesian_plane = FALSE,
         ylab = 'CDF', las = 1)
abline(h = c(0,1), col = 'red', lty = 'dotted')
```

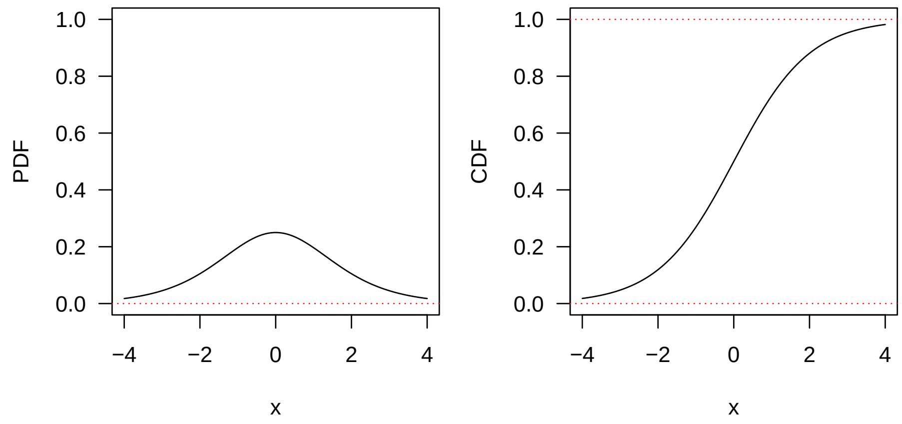


Figure 2.2: Logistic distribution

2.3.3 Other useful rv functions

2.3.3.1 Support indicator

Remark 49. Nel seguito servirà essere compatti/sicuri sul fatto che, al di fuori del supporto R_X della vc X, la probabilità/densità sia nulla. Per farlo si moltiplicherà la PMF/PDF per la funzione indicatrice applicata al supporto della variabile casuale.

Definition 2.3.5 (Funzione indicatrice del supporto di una vc)**.** Definita come:

$$\mathbb{1}_{R_X}(x) = \begin{cases} 1 & \text{se } x \in R_X \\ 0 & \text{se } x \notin R_X \end{cases}$$

2.3.3.2 Survival and hazard function

Remark 50. If rv T has non negative support (eg lifetime), then two function are useful (survival for both discrete and continuous rvs, hazard for continuous)

Definition 2.3.6 (Survival function)**.** Given a rv T such as $\mathbb{P}(T \geq 0) = 1$, it's defined as complement to 1 of cumulative distribution function

$$S(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t) = 1 - F_T(t) \tag{2.22}$$

Definition 2.3.7 (Funzione di azzardo (o rischio))**.** Given a continuous rv T such as $\mathbb{P}(T \geq 0) = 1$, hazard function is defined as

$$H(t) = \frac{f_T(t)}{1 - F_T(t)} = -\frac{d}{\mathrm{d}t} \log(1 - F_T(t)) = -\frac{d}{\mathrm{d}t} \log(S(t)) \tag{2.23}$$

Remark 51. Hazard function can be interpreted as the probability that T stops at t given that it arrived to t

Remark 52. Relationship between Hazard, survival, density and distribution function can be retrieved by the equation. Eg integrating both members between tra $-\infty$ and x we have

$$\begin{aligned} H(t) &= -\frac{d}{\mathrm{d}t}\log\left(S(t)\right) \\ \int_{-\infty}^{x} H(t)\ \mathrm{d}t &= \int_{-\infty}^{x} -\frac{d}{\mathrm{d}t}\log\left(S(t)\right) \\ \int_{-\infty}^{x} H(t)\ \mathrm{d}t &= -\log\left(S(t)\right) \end{aligned}$$

Therefore:

$$\begin{aligned} \log\left(S(t)\right) &= -\int_{-\infty}^{x} H(t)\ \mathrm{d}t \\ S(t) &= \exp\left(-\int_{-\infty}^{x} H(t)\ \mathrm{d}t\right) \end{aligned} \tag{2.24}$$

While for what concerns $F_T(t)$ e $f_T(t)$ we have:

$$F_T(t) = 1 - \exp\left(-\int_{-\infty}^{x} H(t)\ \mathrm{d}t\right) \tag{2.25}$$

$$f_T(t) = H(t)\cdot\exp\left(-\int_{-\infty}^{x} H(t)\ \mathrm{d}t\right) \tag{2.26}$$

Btw, in the lower limit of integration we could have write 0 instead of $-\infty$.

2.4 Transformation

Definition 2.4.1 (Trasform of rv $g(X)$)**.** Considered an experiment with sample space Ω, a random variable X on it and a function $g : \mathbb{R} \to \mathbb{R}$, then $Y = g(X)$ is the random variable mapping $\omega \to g(X(\omega))$, $\forall\omega \in \Omega$ and having support $R_{g(X)} = \{g(X(\omega_1)), g(X(\omega_2)), \ldots\}$. We're interested in finding the distribution of Y, knowing the distribution of X

Remark 53. The logic behind is that, if X is a rv and g is a "well behaved" function (mainly *strictly increasing* or *strictly decreasing*), then $g(X)$ is also a rv. Our main aim is determine density function of $g(X)$.

Remark 54. More generally let X be a n-variate random vector and $Y = g(X)$ where $g : \mathbb{R}^n \to \mathbb{R}^m$ is (borel) measurable. Given the distribution of X, we're interested in finding the distribution of Y.
This problem may be easy but also extremely difficult. Here we discuss a couple of simple cases where $m = n = 1$ (in blue in the continuous area).

2.4.1 Discrete rv transform

Remark 55. In the discrete case finding PMF of $g(X)$ is usually easy, the following are some example.

X	$\mathbb{P}(X = x)$	$Y = 2X$	$\mathbb{P}(Y = y)$	$Z = X^2$	$\mathbb{P}(Z = z)$
-1	0.33	-2	0.33	1	0.66
0	0.33	0	0.33	0	0.33
1	0.33	2	0.33		

Table 2.1: PMF of discrete rv transform, an example

Remark 56. Given a discrete rv X with known PMF, how to get PMF of $Y = g(X)$? If:

- g è *injective*, $X(s_1) \neq X(s_2) \implies g(X(s_1)) \neq g(X(s_2))$, then PMF Y will be the same of X:

$$\mathbb{P}(Y = g(x)) = \mathbb{P}(g(X) = g(x)) = \mathbb{P}(X = x)$$

- otherwise there could be cases where $X(s_1) \neq X(s_2)$ but $\implies g(X(s_1)) = g(X(s_2))$: here we have to sum probability of different x that with g ends in the same y.

The following result is general and is ok for both cases

Proposition 2.4.1 (PMF of $g(X)$)**.** *Let X be a discrete rv and $g : \mathbb{R} \to \mathbb{R}$. Then support of $g(X)$ is the set of y such as that $g(x) = y$ for at least one $x \in R_X$ and PMF of $g(X)$ is*

$$\mathbb{P}(g(X) = y) = \sum_{x:g(x)=y} \mathbb{P}(X = x), \qquad \forall y \in R_{g(X)} \tag{2.27}$$

Example 2.4.1. In table 2.1 an example with X, $Y = 2X$ ($g(x) = 2 \cdot x$, injective) e $Z = X^2$ ($g(x) = x^2$ not injective).

Remark 57. It's a common error to apply g to the PMF (it could take probability over 1): g have to be applied to domain/support of PMF.

Example 2.4.2 (Transformation of a bernoulli)**.** Let $X \sim \text{Bern}(p)$ and we're interested in $g(X) = e^X$. What is the dist of $g(X)$. We have that

$$X = \begin{cases} 1 & \text{with prob } p \\ 0 & \text{with prob } 1 - p \end{cases}, \qquad g(X) = \begin{cases} e^1 = e & \text{with prob } p \\ e^0 = 1 & \text{with prob } 1 - p \end{cases}$$

Therefore

$$\mathbb{P}(g(X) = e) = \mathbb{P}\left(X = g^{-1}(e)\right) = \mathbb{P}(X = 1) = p$$

2.4.2 Continuous rvs transform (linear case)

Definition 2.4.2 (Scale-location transform for continuous rv)**.** Let X be a continuous rv; $Y = \sigma X + \mu$ with $\sigma, \mu \in \mathbb{R}$ is a random variable obtained using a (linear) transform of both position and scale.

Remark 58. Here σ set the scale (if positive spread Y compared to X) while μ the location (if positive moves Y distribution toward right compared to X).

Remark 59. In order to go back to X we standardize Y, aka apply the transformation $X = \dfrac{Y-\mu}{\sigma}$.

Proposition 2.4.2. *Y has the same family of distribution as X.*

Proof. It has been obtained by a linear, injective transformation. □

Remark 60. If this kind of transformation is applied to a discrete rv we have a distribution no more of the same family, considered that support changes (eg linear transform of a binomial does not give a binomial, defined on support $0, 1, \ldots$).

2.4.3 Continuous rvs (monotonic) transform

Proposition 2.4.3 ($g = F$ con F funzione di ripartizione di X)**.** *Let X be a real r.v. with distribution function F (this means that $F(x) = \mathbb{P}(X \le x)$, $\forall x \in \mathbb{R}$). If F is continuous the $Y = F(X) \sim \text{Unif}(0,1)$ that is*

$$\mathbb{P}(Y \le y) = \begin{cases} 0 & y < 0 \\ y & 0 \le y < 1 \\ 1 & y \ge 1 \end{cases}$$

So $Y = F(X) \sim \text{Unif}(0,1)$

Proof. To prove it, for the sake of simplicity assume that F is not only continuous, but also strictly *increasing* (this is not actually needed for thm to hold). In this case $\forall y \in (0,1)$ one obtain

$$\mathbb{P}(Y \le y) = \mathbb{P}(F(X) \le y) = \mathbb{P}\left(X \le F^{-1}(y)\right) = F\left(F^{-1}(y)\right) = y$$

ma questa è proprio la funzione di ripartizione di una uniforme in $(0,1)$. □

Proposition 2.4.4. *Let X be absolutely continuous and suppose that $P(X \in I) = 1$ where I is the interval where a function $g : I \to R$ is defined. Suppose also that g is everywhere differentiable $g' \neq 0$. Then $Y = g(X)$ is still absolutely continuous with density*

$$h_Y(y) = f\left(g^{-1}(y)\right)\left|g^{-1'}(y)\right| \cdot \mathbb{1}_{g(I)}(y), \quad \forall y \in g(I)$$

where f denotes the density of X

Example 2.4.3 ($-\log \text{Unif}(0,1)$)**.** Sia X uniforme in (0,1), voglio la legge di $Y = -\log(X)$. **NB**: dalla triennale
Basta porre $I = (0,1)$, $g(x) = -\log(x)$, da cui $g'(x) = -\frac{1}{x} \neq 0 \forall x \in (0,1)$ e $\mathbb{P}(X \in (0,1)) = 1$. Quindi posso concludere che Y è assolutamente continua con densità

$$\begin{aligned} h(y) &= f(g^{-1}(y)) \cdot \left|(g^{-1}(y))'\right| \cdot \mathbb{1}_{g(I)}(y) \\ &= f(e^{-y}) \cdot \left|(e^{-y})'\right| \cdot \mathbb{1}_{(0,+\infty)}(y) \\ &= f(e^{-y}) \cdot e^{-y}\mathbb{1}_{(0,+\infty)}(y) \\ &= 1 \cdot e^{-y}\mathbb{1}_{(0,+\infty)}(y) \end{aligned}$$

dove:

- $y = -\log x \iff x = e^{-y}$ ovvero $g^{-1}(y) = e^{-y}$
- $g(I) = \{g(x) : x \in I\} = \{-\log x : x \in (0, 1)\} = (0, +\infty)$
- $(e^{-y})' = -e^{-y}$ ma poi prendendone il valore assoluto il meno va via
- essendo X uniforme si ha che

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{altrimenti} \end{cases}$$

  e quindi poiché $e^{-y} \in (0, 1), \forall y \in (0, +\infty)$ si ha $f(e^{-y}) = 1$

In definitiva, poiché $h(y) = \mathbb{1}_{(0,+\infty)}(e)^{-y}$ è la densita di una esponenziale con $\lambda = 1$ abbiamo dimostrato che

$$-\log \mathrm{Unif}\,(0, 1) \sim \mathrm{Exp}\,(1)$$

Example 2.4.4. Per definizione Y ha legge lognormale se $Y > 0$ e $\log Y \sim \mathrm{N}\left(\mu, \sigma^2\right)$. Al fine di avere una legge esplicita per Y possiamo considerare che deve esere $Y \sim e^X$ con $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$. Per ottenerla dunque basta applicare il teorema precedente con $X \sim N$ e $g(x) = \exp(x)$.
Posso dunque concludere che $Y = g(X)$ ha legge assolutamente continua con densità

$$\begin{aligned} h(y) &= f(g^{-1}(y)) \cdot \left|\left(g^{-1}(y)\right)'\right| \cdot \mathbb{1}_{g(I)}(y) \\ &= f(\log(y)) \cdot |(\log(y))'| \cdot \mathbb{1}_{(0,+\infty)}(y) \\ &= \frac{f(\log(y))}{y} \mathbb{1}_{(0,+\infty)}(y) \\ &= \frac{\exp\left[-\frac{1}{2}\left(\frac{\log y - \mu}{\sigma}\right)^2\right]}{y\sqrt{2\pi\sigma^2}} \mathbb{1}_{(0,+\infty)}(y) \end{aligned}$$

Proposition 2.4.5. *If X is a continuous random variable, g a monotonic function (strictly increasing or decreasing), the density function of the random variable $g(X)$, $f_{g(X)}$, is obtained as:*

$$f_{g(X)}(x) = f_X(g^{-1}(x)) \cdot \left|\frac{\partial g^{-1}(x)}{\partial x}\right| \tag{2.28}$$

Proof. For the continuous case we have that, in order to obtain $f_{g(X)}(x)$ we need to differentiate $F_{g(X)}(x)$

$$F_{g(X)}(x) = \mathbb{P}\left(g(X) \leq x\right)$$

Now

- if the function g is *decreasing* we have

$$\begin{aligned} F_{g(X)}(x) &= \mathbb{P}\left(g(X) \leq x\right) = \mathbb{P}\left(X \geq g^{-1}(x)\right) = 1 - \mathbb{P}\left(X < g^{-1}(x)\right) \\ &= 1 - F_X(g^{-1}(x)) \end{aligned}$$

- viceversa if g is *increasing*

$$F_{g(X)}(x) = \mathbb{P}(g(X) \leq x) = \mathbb{P}\left(X \leq g^{-1}(x)\right) = F_X(g^{-1}(x))$$

In any case after that we have that

$$\begin{aligned}
f_{g(X)}(x) = \frac{\partial}{\partial x} F_{g(X)}(x) &= \begin{cases} \frac{\partial\left(1-F_X(g^{-1}(x))\right)}{\partial x} & \text{if increasing} \\ \frac{\partial\left(F_X(g^{-1}(x))\right)}{\partial x} & \text{if decreasing} \end{cases} \\
&= \begin{cases} -f_X(g^{-1}(x)) \cdot \frac{\partial}{\partial x} g^{-1}(x) \\ f_X(g^{-1}(x)) \cdot \frac{\partial}{\partial x} g^{-1}(x) \end{cases}
\end{aligned}$$

The two cases can be combined in the single formula (not clear how to me for the moment) which is the thorem ☐

Example 2.4.5 (Esercizio Berk Tan)**.** Let $X \sim \text{Unif}(0,1)$ and be $g(x) = e^x$; then what is the pdf of $Y = g(X)$? We have that $g^{-1}(Y) = \log Y$, so

$$\frac{\partial}{\partial y}(g^{-1}(y)) = \frac{1}{y}$$

Applying the formula

$$f_Y(y) = \mathbb{1}_{[0,1]}(\log y)\frac{1}{y}$$

and expressing $\mathbb{1}_{[0,1]}(\log y)$ in terms of y we have

$$\begin{aligned} 0 \leq \log y \leq 1 \\ 1 \leq y \leq e \end{aligned}$$

so finally

$$f_Y(y) = \mathbb{1}_{[1,e]}(y)\frac{1}{y} = \begin{cases} \frac{1}{y} & \text{if } y \in [1,e] \\ 0 & \text{elsewhere} \end{cases}$$

Example 2.4.6 (Esame vecchio viroli)**.** Let X have the probability density function given by

$$f_X(x) = \frac{x}{2}$$

with $X \in [0,2]$. Find the density function of $Y = 6X - 3$.
Qua il dominio diventa palesemente $Y \in [-3,9]$, per quanto riguarda la funzione si ha che

$$\begin{aligned}
f_Y(y) &= \left|\frac{\partial}{\partial y} g^{-1}(y)\right| f_X(g^{-1}(y)) \\
g(X) &= 6X - 3 \quad g^{-1}(Y) = \frac{Y+3}{6} \\
f_Y(y) &= \frac{1}{6}\left(\frac{Y+3}{6 \cdot 2}\right) = \frac{1}{6}\left(\frac{Y+3}{12}\right)
\end{aligned}$$

the answer is $f_Y(y) = \frac{3+y}{12}\frac{1}{6}$.
Si può verificare che $\int_{-3}^{9} f_Y(y) = 1$ mediante sympy. Qui non c'è il problema di resprimere le variabili indicatrici (perché non è una uniforme 0,1 e la densità non ne fa uso).

Example 2.4.7 (Assignment 1 Viroli, Exercise 2)**.** Let $X \sim \text{Unif}\,(0,1)$. Find the PDF of $X^{1/\alpha}$ with $\alpha > 0$.
Let $X \sim \text{Unif}\,(0,1)$ and $Y = X^{\frac{1}{\alpha}}$, with $\alpha > 0$. Let's obtain $f_Y(y)$ by applying:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{\partial g^{-1}(y)}{\partial y}\right| \tag{2.29}$$

Being $X \sim \text{Unif}\,(0,1)$ we have that $f_X(x) = \mathbb{1}_{[0,1]}(x)$. Given the transformation $y = x^{1/\alpha}$, its inverse is

$$y = x^{1/\alpha} \iff y^\alpha = x$$

so $g^{-1}(Y) = Y^\alpha$; doing the derivative with respect to y we obtain:

$$\frac{\partial}{\partial y} g^{-1}(y) = \alpha y^{\alpha-1}$$

so putting things together:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{\partial g^{-1}(y)}{\partial y}\right| = \mathbb{1}_{[0,1]}(y^\alpha) \cdot \alpha y^{\alpha-1}$$

Now we need to express the indicator $\mathbb{1}_{[0,1]}(y^\alpha)$ in terms of y, therefore:

$$\begin{aligned} 0 \le y^\alpha \le 1 \\ 0 \le y \le 1 \end{aligned}$$

Finally:

$$f_Y(y) = \mathbb{1}_{[0,1]}(y) \cdot \alpha y^{\alpha-1} = \begin{cases} \alpha y^{\alpha-1} & \text{if } y \in [0,1] \\ 0 & \text{elsewhere} \end{cases}$$

If $\alpha = 1$, as expected

$$f_Y(y) = \begin{cases} 1 & \text{if } y \in [0,1] \\ 0 & \text{elsewhere} \end{cases} = \mathbb{1}_{[0,1]}(y) \implies Y \sim \text{Unif}\,(0,1)$$

Example 2.4.8 (Esercizio virol)**.** If $X \sim \text{Unif}\,(0,1)$ and $Y = -2\log X$, show that $Y \sim \chi^2_2$. We apply 2.28 and compare with χ^2_n one.
We have the transformation $y = -2\log x$ so to obtain the inverse

$$-\frac{1}{2}y = \log x \iff x = e^{-\frac{1}{2}y}$$

therefore $g^{-1}(Y) = \exp\left(-\frac{Y}{2}\right)$. We have, being X a uniform on 0,1, that $f_X(x) = 1 \cdot \mathbb{1}_{[0,1]}(x)$. Now

$$\frac{\partial}{\partial y} g^{-1}(y) = -\frac{1}{2}e^{-y/2}$$

So applying the formula we arrive at

$$f_Y(y) = \mathbb{1}_{[0,1]}\left(e^{-y/2}\right) \cdot \frac{1}{2}e^{-y/2}$$

Now we need to express $\mathbb{1}_{[0,1]}\left(e^{-y/2}\right)$ in terms of y. The domain of y so

$$\begin{aligned}
&0 \leq e^{-y/2} \leq 1 \\
&-\infty < -y/2 \leq 0 \\
&0 < y \leq +\infty
\end{aligned}$$

Finally

$$f_Y(y) = \mathbb{1}_{[0,+\infty)}(y) \cdot \frac{1}{2} e^{-y/2} = \begin{cases} \frac{1}{2} e^{-y/2} & \text{if } y \in [0,+\infty) \\ 0 & \text{elsewhere} \end{cases}$$

which is a χ^2 with 2 degrees of freedom.

2.5 Independence

2.5.1 Independence

Remark 61 (Notation)*.* We can write intersections of events as follows

$$\begin{aligned}
&\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A \cap Y \in B) \\
&\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x \cap Y \leq y)
\end{aligned}$$

Remark 62*.* The concept of independence for random variables is similar to events independence.

Definition 2.5.1 (RVs independence (general case))**.** Given *any* collection (finite, countable, non countable) of random variables $\mathcal{V} = \{X_1, X_2, \ldots\}$, the elements of $\mathcal{V}$ are said to be independent if, for any *finite* subset of events $\mathcal{X} \subset \mathcal{V}$

$$\begin{aligned}
&\mathbb{P}(X_j \in B_j, \ldots, X_k \in B_k) = \mathbb{P}(X_j \in B_j) \cdot \ldots \cdot \mathbb{P}(X_k \in B_k) \\
&X_j, \ldots, X_k \in \mathcal{X} \quad \forall B_j, \ldots, B_k \in \mathscr{B}
\end{aligned}$$

or equivalently

$$\begin{aligned}
&\mathbb{P}(X_j \leq x_j, \ldots, X_k \leq x_k) = \mathbb{P}(X_j \leq x_j) \cdot \ldots \cdot \mathbb{P}(X_k \leq x_k) \qquad (2.30) \\
&X_j, \ldots, X_k \in \mathcal{X}, \quad \forall x_j, \ldots, x_k \in \mathbb{R}
\end{aligned}$$

Example 2.5.1 (Independence of two RVs $X \perp\!\!\!\perp Y$)**.** Two rvs X, Y are independent, and we write $X \perp\!\!\!\perp Y$, if

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y), \qquad \forall x, y \in \mathbb{R} \qquad (2.31)$$

NB: La definizione generale di sopra vale anche qui perché in un set finito chiediamo un subset sempre finito di 2 variabili poi lui credo usi $\mathcal{X} \subset \mathcal{V}$ per poter intendere anche $\mathcal{X} \subseteq \mathcal{V}$

Remark 63*.* In the discrete case 2.31 is equivalent to

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y), \qquad \forall x, y \in \mathbb{R}$$

Example 2.5.2. Let be X the result of first dice thrown and Y the second; sum and difference of results random variables $X+Y$, $X-Y$ are not independent considered that:

$$\begin{aligned}
&\mathbb{P}(X+Y = 12, X-Y = 1) = 0 \\
&\mathbb{P}(X+Y = 12) \cdot \mathbb{P}(X-Y = 1) = \frac{1}{6} \cdot \frac{5}{6}
\end{aligned}$$

This does make sense: knowing that the sum is 12, tells that their difference must be 0 so the two rv gives information of each other

Proposition 2.5.1. *If $X_1, \dots, X_n$ are independent, then they are pairwise, 3-wise, $\dots (n-1)$-wise independent. Viceversa implication does not hold.*

Proof. If $X_1, \dots, X_n$ are independent si ha (considerando a titolo di esempio la coppia X_1, X_2) che

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) = \mathbb{P}(X_1 \leq x_1) \cdot \mathbb{P}(X_2 \leq x_2)$$

Per vedere perché sia così basta far tendere a $+\infty$ gli $x_3, \dots, x_n$ in maniera tale che a sinistra dell'uguale, nella definizione 2.30, entro parentesi si abbiano eventi certi e a destra dell'uguale si moltiplichi per 1.
The example of why contrary implication does not hold can be done via counterexample. $\square$

Example 2.5.3. Example of three variables which are pairwise independent but not independent. Let X, Y be iid with

$$\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$$

and $Z = XY$ so that

$$\begin{aligned}
\mathbb{P}(Z = 1) &= \mathbb{P}(X = Y) = \mathbb{P}(X = 1, Y = 1) + \mathbb{P}(X = -1, Y = -1) \\
&= \mathbb{P}(X = 1) \cdot \mathbb{P}(Y = 1) + \mathbb{P}(X = -1) \cdot \mathbb{P}(Y = -1) \\
&= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}
\end{aligned}$$

and thus $\mathbb{P}(Z = -1) = 1/2$.
The set $\{X, Y, Z\}$ is not independent since, for example

$$\mathbb{P}(X = 1, Y = -1, Z = 1) = \mathbb{P}(\emptyset) = 0 \neq \mathbb{P}(X = 1) \cdot \mathbb{P}(Y = -1) \cdot \mathbb{P}(Z = 1) = \frac{1}{8}$$

However the three random variables are pairwise independent since

$$\begin{aligned}
&\mathbb{P}(X = 1, Y = 1) = \mathbb{P}(X = 1) \cdot \mathbb{P}(Y = 1) = \frac{1}{2}\frac{1}{2} = \frac{1}{4} \\
&\mathbb{P}(X = 1, Y = -1) = \mathbb{P}(X = 1) \cdot \mathbb{P}(Y = -1) = \frac{1}{2}\frac{1}{2} = \frac{1}{4} \\
&\mathbb{P}(X = 1, Z = 1) = \mathbb{P}(X = 1) \cdot \mathbb{P}(Z = 1) = \frac{1}{2}\frac{1}{2} = \frac{1}{4} \\
&\mathbb{P}(X = 1, Z = -1) = \mathbb{P}(X = 1) \cdot \mathbb{P}(Z = -1) = \frac{1}{2}\frac{1}{2} = \frac{1}{4} \\
&\dots
\end{aligned}$$

and in general one obtains

$$\mathbb{P}(X = a, Z = b) = \mathbb{P}(X = a)\, \mathbb{P}(Z = b), \quad \forall a, b \in \{-1, 1\}$$

Proposition 2.5.2 (Transform of independent rv)**.** *If X and Y are independent, then any transformation of X and Y are independent as well.*

Proof. Not shown. $\square$

2.5.2 IID RVs

Remark 64. A very important case is IID variables; this assumption is involved in the *law of large number* and *central limit theorem*.

Definition 2.5.2 (i.i.d. rvs)**.** Random variables in the set $\mathcal{V} = \{X_1, X_2, \dots\}$ are *independent* and *identically* distributed if

- the elements of $\mathcal{V}$ are independent
- $X_i \sim X_j$, forall $X_i, X_j \in \mathcal{V}$ (have the same distribution function).

Important remark 26 (Notation)*.* If the elements of $\mathcal{X} = \{X_1, X_2, \dots\}$ are iid, to communicate the common distribution of the X_i it suffices to write $X_i \sim \nu$

2.5.3 Conditional independence

Definition 2.5.3 (Conditional independence)**.** X and Y are conditional independent given Z if $\forall x, y \in \mathbb{R}$ and $\forall z \in R_Z$ it is:

$$\mathbb{P}(X \leq x, Y \leq y | Z = z) = \mathbb{P}(X \leq x | Z = z) \cdot \mathbb{P}(Y \leq y | Z = z) \tag{2.32}$$

Remark 65. For discrete rvs, an equivalent definition based on the mass function is

$$\mathbb{P}(X = x, Y = y | Z = z) = \mathbb{P}(X = x | Z = z) \cdot \mathbb{P}(Y = y | Z = z) \tag{2.33}$$

Proposition 2.5.3. *Rvs indipendence does not imply conditional independence and viceversa.*

Proof. By counterexamples, see Blitzstein pag 121. □

2.6 Moments

Remark 66. Distribution functions are the unifying concepts for continuous and discrete rvs; furthermore knowing F_X is to know the entire probabilistic structure of the rv.
In order to compare different rv, however, often synthetic indicator are needed and these are the moments.

Definition 2.6.1 (Moment of a rv)**.** A statistic of this kind, if it exists

$$\begin{aligned} &\sum_{i=1}^{\infty} g(x_i) \cdot p_X(x_i) && \text{if } X \text{ is discrete} \\ &\int_{-\infty}^{+\infty} g(x) \cdot f_X(x)\, \mathrm{d}x && \text{if } X \text{ is abs. continuous} \end{aligned}$$

Different g functions defines different moments

Important remark 27 (Important moments)*.* These are expected value, variance, asymmetry and kurtosis; all can be seen as a specialized (for g) version of the equations above.

2.6.1 Expected value

Remark 67 (Expected value existence check). Let X be a real r.v.; we aim to define its expectation. Before doing this however, it should be noted that such expectation (which involves series or integrals) may fail to exists (not finite).
To define the expectation of X (whether it is discrete or continuous), one should previously evaluate the expectation of $|X|$, that is $\mathbb{E}[|X|]$; this can be done through the formula

$$\mathbb{E}[|X|] = \int_0^{+\infty} \mathbb{P}(|X| > t)\, \mathrm{d}t$$

Incidentally if X is absolutely continuous, the above integral can be written as

$$\mathbb{E}[|X|] = \int_0^{+\infty} \mathbb{P}(|X| > t)\, \mathrm{d}t = \int_{-\infty}^{+\infty} |x|\, f(x)\, \mathrm{d}x$$

where f is the density of X. Now there are two situations:

$$\begin{cases} \mathbb{E}[|X|] = +\infty & \implies \text{we stop: expectation of } X \text{ does not exists} \\ \mathbb{E}[|X|] < \infty & \implies \text{expectation of } X \text{ exists and may be evaluated with following formulas} \end{cases}$$

Definition 2.6.2 (Expected value)**.** If $\mathbb{E}[|X|] < +\infty$ the expectation of X, denoted by $\mathbb{E}[X]$ or μ, gives a probability weighted mean of X and can be evaluated by

$$\mathbb{E}[X] = \begin{cases} \sum_i x_i \cdot \mathbb{P}(X = x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{+\infty} x \cdot f_X(x)\, \mathrm{d}x & \text{if } X \text{ is abs. continuous} \\ \int_0^{+\infty} \mathbb{P}(X > t)\, \mathrm{d}t & \text{if } X \geq 0 \end{cases}$$

Remark 68. The cases above don't cover all the possible cases (eg there are other formulas if X is not discrete, absolutely continuous or non negative) but are more than enough for us

Example 2.6.1 (Single dice)**.** Let X be the result of a single fair dice with $p_X(1) = \ldots = p_X(6) = 1/6$:

$$\mathbb{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}$$

Example 2.6.2. For the Cauchy random variable, the expected value does not exists. If $X \sim \mathit{Cauchy}$, X is absolutely continuous with support $\mathbb{R}$ and density

$$f(x) = \frac{1}{\pi} \frac{1}{1 + x^2}$$

In order to check it, we start evaluating the test for expected value existence

$$\begin{aligned} \mathbb{E}[|X|] &= \int_{-\infty}^{+\infty} |x| \cdot \frac{1}{\pi} \frac{1}{1 + x^2} \overset{(1)}{=} 2 \int_0^{+\infty} x \cdot \frac{1}{\pi} \frac{1}{1 + x^2} \\ &= 2 \cdot \frac{1}{\pi} \int_0^{+\infty} \frac{x}{1 + x^2}\, \mathrm{d}x \overset{(2)}{=} +\infty \end{aligned}$$

where in:

- (1) because it's an even function (symmetryc with respect to y axis) so we can double the integral on the positive part (taking x out of absolute value;
- (2) if we want to check very well , integrating by parts we have:

  $$\int \frac{x}{1+x^2}\,\mathrm{d}x = \frac{1}{2}\int \frac{2x}{1+x^2}\,\mathrm{d}x = \frac{1}{2}\log(1+x^2)+c$$

  Therefore

  $$\mathbb{E}\left[|X|\right] = \frac{2}{\pi}\left(\left[\frac{1}{2}\log(1+x^2)\right]_0^{+\infty}\right) = \frac{2}{\pi}(+\infty - 0) = +\infty$$

NB: rigo non ha fatto l'integrazione

Therefore the expected value does not exists.

Remark 69. Generalizing a bit, expectation is the *first* moment of a random variable X.

Definition 2.6.3 (Moment of order r (r-th moment) of X)**.** Adopting as g the r-power of X in the definition 2.6.1

$$\mu_r = \mathbb{E}\left[X^r\right] = \begin{cases} \sum_i {x_i}^r \cdot \mathbb{P}\left(X = x_i\right) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{+\infty} x^r \cdot f_X(x)\,\mathrm{d}x & \text{if } X \text{ is abs. continue} \end{cases} \tag{2.34}$$

Definition 2.6.4 (Moment of order r existence)**.** In general moment of order r for X exists (or X has moment of order r) if $\mathbb{E}\left[|X|^r\right] < +\infty$.

Remark 70. A useful results is the following.

Theorem 2.6.1. *If $\mathbb{E}\left[|X|^r\right] < +\infty$ for some $r > 0$, then all the moments of order $q \leq r$ exists/are finite as well:*

$$\mathbb{E}\left[|X|^q\right] < +\infty, \quad \forall q \in (0, r]$$

Remark 71. From now on, all the involved rv are assumed to have the mean. The following properties are very useful since they hold for any rv (regardless the type). The only needed assumption is that the involved rv havs the mean.

Proposition 2.6.2 (Main properties of the operator $\mathbb{E}\left[\cdot\right]$)**.** *We have*

1. $\mathbb{E}\left[aX + bY\right] = a\,\mathbb{E}\left[X\right] + b\,\mathbb{E}\left[Y\right]$ *(linearity)*
2. *if* $c \in \mathbb{R}$, $\mathbb{E}\left[c\right] = c$ *(expval of constant/dirac)*
3. $X \geq 0$ *a.s.* $\mathbb{E}\left[X\right] \geq 0$ *(positivity, just* $\geq$*)*
4. *if* $X \geq 0$ *and* $\mathbb{P}\left(X > 0\right) > 0$ *then* $\mathbb{E}\left[X\right] > 0$ *(strict positivity)*

Proposition 2.6.3 (Expected value properties (old non Rigo version))**.**

$$\mathbb{E}\left[aX+b\right] = a\,\mathbb{E}\left[X\right]+b \tag{2.35}$$
$$\mathbb{E}\left[X+Y\right] = \mathbb{E}\left[X\right]+\mathbb{E}\left[Y\right] \tag{2.36}$$
$$X \geq 0 \implies \mathbb{E}\left[X\right] \geq 0 \tag{2.37}$$
$$X \geq 0, \mathbb{P}\left(X>0\right)>0 \implies \mathbb{E}\left[X\right] > 0 \tag{2.38}$$
$$\mathbb{E}\left[g(X)\right] = \sum_i g(x_i)\cdot p_X(x_i) \tag{2.39}$$
$$X \perp\!\!\!\perp Y \implies \mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \tag{2.40}$$
$$\min\left(X\right) \leq \mathbb{E}\left[X\right] \leq \max\left(X\right) \tag{2.41}$$
$$\mathbb{E}\left[X-\mathbb{E}\left[X\right]\right] = 0 \tag{2.42}$$
$$\textit{minimizes}\ \mathbb{E}\left[\left(X-\mathbb{E}\left[X\right]\right)^2\right] \tag{2.43}$$

Remark 72. Congiuntamente alle 2.35 e 2.36 ci si riferisce come linearità del valore atteso, che torna spesso comodo per il calcolo soprattutto se si riesce a scrivere una vc come somma di due o più vc. La linearità è un mero fatto algebrico e di bello c'è che, ad esempio per 2.36, non è necessaria l'indipendenza tra X e Y affinché valga.

TODO: da chiarire sta nota di colore

Important remark 28. If $f:\mathbb{R}\to\mathbb{R}$ is a measurable function, to evaluate the expectation of $f(X)$, that is $E(f(X))$, we can repeat the previous properties with $f(X)$ instead of X.

Proof. Mostriamo con riferimento alle variabili discrete. Per la 2.35

$$\begin{aligned}
\mathbb{E}\left[aX+b\right] &= \sum_i (ax_i+b)\cdot\mathbb{P}\left(aX+b=ax_i+b\right) = \sum_i (ax_i+b)\cdot\mathbb{P}\left(X=x_i\right)\\
&= \sum_i ax_i\cdot\mathbb{P}\left(X=x\right) + \sum_i b\cdot\mathbb{P}\left(X=x\right)\\
&= a\sum_i x_i\cdot\mathbb{P}\left(X=x\right) + b\underbrace{\sum_i\mathbb{P}\left(X=x\right)}_{1}\\
&= a\,\mathbb{E}\left[X\right]+b
\end{aligned}$$

Viceversa nel caso continuo

$$\mathbb{E}\left[aX+b\right] = \int_{D_x}(ax+b)f(x)\,\mathrm{d}x = a\int_{D_x}xf(x)\,\mathrm{d}x + b\underbrace{\int_{D_x}f(x)\,\mathrm{d}x}_{=1} = a\,\mathbb{E}\left[X\right]+b$$

Per 2.36 facendo un passo indietro, possiamo scrivere un generico valore atteso facendo riferimento all'evento $s\in\Omega$ e applicando la funzione X ad esso, al fine di ottenere x_i:

$$\mathbb{E}\left[X\right] = \sum_i x_i\cdot\mathbb{P}\left(X=x_i\right) = \sum_s X(s)\cdot\mathbb{P}\left(\{s\}\right)$$

Da questa possiamo generalizzare alla somma di due funzioni:

$$\begin{aligned}\mathbb{E}[X+Y] &= \sum_s (X+Y)(s) \cdot \mathbb{P}(\{s\}) = \sum_s (X(s)+Y(s)) \cdot \mathbb{P}(\{s\}) \\ &= \sum_s X(s) \cdot \mathbb{P}(\{s\}) + \sum_s Y(s) \cdot \mathbb{P}(\{s\}) \\ &= \mathbb{E}[X] + \mathbb{E}[Y]\end{aligned}$$

Per il valore atteso della trasformazione g, 2.39, sfruttiamo la stessa tecnica facendo un passo indietro (rispetto all'applicazione della funzione X agli eventi dello spazio campionario): sia $s \in \Omega$ un evento dello spazio campionario e X la vc considerata. Come detto possiamo scrivere il valore atteso $\mathbb{E}[X]$ come prodotto del risultato di X per la probabilità che si verifichi quell'evento:

$$\mathbb{E}[X] = \sum_s X(s)\,\mathbb{P}(\{s\})$$

L'applicazione della trasformazione g porta il valore atteso $\mathbb{E}[g(X)]$:

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_s g(X(s)) \cdot \mathbb{P}(\{s\}) \\ &\overset{(1)}{=} \sum_i \sum_{s:X(s)=x_i} g(X(s))\,\mathbb{P}(\{s\}) \\ &= \sum_i g(x_i) \sum_{s:X(s)=x_i} \mathbb{P}(\{s\}) \\ &= \sum_i g(x_i) \cdot \mathbb{P}(X = x_i) \\ &= \sum_i g(x_i) \cdot p_X(x_i)\end{aligned}$$

dove in (1) semplicemente raggruppiamo per i diversi s che attraverso X forniscono lo stesso x_i.
Per 2.40 (mostrando il caso delle discrete) se $X \perp\!\!\!\perp Y$, allora $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$, da questo

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x \in D_x} \sum_{y \in D_y} x \cdot y \cdot \mathbb{P}(X = x, Y = y) = \sum_{x \in D_x} \sum_{y \in D_y} x \cdot y \cdot \mathbb{P}(X = x)\,\mathbb{P}(Y = y) \\ &= \sum_{x \in D_x} x \cdot \mathbb{P}(X = x) \sum_{y \in D_y} y \cdot \mathbb{P}(Y = y) = \mathbb{E}[X] \cdot \mathbb{E}[Y]\end{aligned}$$

La 2.41 è ovvia essendo $\mathbb{E}[X]$ una media pesata da probabilità dei valori assunti da X; l'uguaglianza vale in caso di variabili degeneri.
La 2.42 è una applicazione della linearità

$$\mathbb{E}[X - \mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$$

□

Example 2.6.3 (Valore atteso di trasformazione)**.** Supponiamo che X sia una vc che assuma i valori $-1, 0, 1$ con probabilità pari a $\mathbb{P}(x = -1) = 0.2$,

$\mathbb{P}(x=0) = 0.5$, $\mathbb{P}(x=1) = 0.3$. Calcoliamo $\mathbb{E}\left[X^2\right]$ applicando prima la trasformazione e poi moltiplicando per la probabilità:

$$\mathbb{E}\left[X^2\right] = (-1)^2(0.2) + 0^2 \cdot (0.5) + 1^2(0.3) = 0.5$$

Proposition 2.6.4 (Valore atteso di funzioni non lineari di vc)**.** *In generale non vale* $\mathbb{E}[g(X)] = g\,\mathbb{E}[X]$ *per una qualsiasi funzione* g*.*

Example 2.6.4. Sia X il lancio di un dado: calcoliamo $\exp(\mathbb{E}[X])$ e $\mathbb{E}[\exp X]$; ricordando che $\mathbb{E}[X] = 7/2$ si ha

$$\begin{aligned} g(\mathbb{E}[X]) &= \exp(7/2) \approx 33.12 \\ \mathbb{E}[g(X)] &= e^1 \cdot \frac{1}{6} + \ldots + e^6 \cdot \frac{1}{6} \approx 106.1 \end{aligned}$$

Considerando invece una trasformazione lineare $g(x) = 2x + 1$ i due risultati coincidono, come in mostrato 2.35. Si ha:

$$\begin{aligned} g(\mathbb{E}[X]) &= 2 \cdot \frac{7}{2} + 1 = 8 \\ \mathbb{E}[g(X)] &= 3\frac{1}{6} + 5\frac{1}{6} + 7\frac{1}{6} + 9\frac{1}{6} + 11\frac{1}{6} + 13\frac{1}{6} = 8 \end{aligned}$$

2.6.2 Variance

Definition 2.6.5 (Variance)**.** If $\mathbb{E}\left[|X|^2\right] = \mathbb{E}\left[X^2\right] < +\infty$ we can define the variance of X as

$$\bar{\mu}_2 = \mathrm{Var}[X] = \sigma^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \tag{2.44}$$

measure dispersion of the rv around its mean value.

Proposition 2.6.5 (Formula to use for evaluation)**.**

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \tag{2.45}$$

Remark 73. In the computation formula 2.45, its easier to see that to have a variance it must be $\mathbb{E}\left[|X|^2\right] \mathbb{E}\left[X^2\right] < +\infty$

Proof. We expand $(X - \mathbb{E}[X])^2$ and used expected value linearity:

$$\begin{aligned} \mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] &= \mathbb{E}\left[X^2 - 2\,\mathbb{E}[X]\,X + (\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \end{aligned}$$

□

Example 2.6.5 (Dice variance)**.** If X is result of a dice throw, previously we computed $\mathbb{E}[X] = 7/2$; furthermore we have

$$\mathbb{E}\left[X^2\right] = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) = \left(\frac{1}{6}\right)(91)$$

Therefore

$$\mathrm{Var}[X] = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

Remark 74 (Interpretation). We have that:

- X can be regarded as the outcome of a numerical experiment
- $\mathbb{E}[X]$ our best prediction of X (before making the experiment)
- $X - \mathbb{E}[X]$ can be seen as the error
- the variance is $\mathbb{E}\left[\text{error}^2\right]$ if we adopt the best prediction possible (which minimizes error squared) for outcome of our experiment (which is $\mathbb{E}[X]$). Infact, in general, if we predict X by a real number t the error becomes $X - t$. Defining the function

  $$e(t) = \mathbb{E}\left[\text{error}^2\right] = \mathbb{E}\left[(X-t)^2\right]$$

  we aim to minimize e. To this end, we note that

  $$\begin{aligned} e(t) &= \mathbb{E}\left[(X-t)^2\right] = \mathbb{E}\left[\left((X-\mathbb{E}[X]) + (\mathbb{E}[X]-t)\right)^2\right] \\ &= \mathbb{E}\left[\left(X-\mathbb{E}[X]\right)^2\right] + \left(\mathbb{E}[X]-t\right)^2 + 2(\mathbb{E}[X]-t)\underbrace{\mathbb{E}[X-\mathbb{E}[X]]}_{=0} \\ &= \mathrm{Var}[X] + (t-\mathbb{E}[X])^2 \end{aligned}$$

  Hence e attains its minimum at the point $t = \mathbb{E}[X]$ and $\mathrm{Var}[X]$ is our estimate of error in prediction.

Proposition 2.6.6. *Si ha che*

$$X \text{ è degenere} \iff \mathrm{Var}[X] = 0$$

Proof. La si può provare utilizzando una disuguaglianza (come faremo a tempo debito) o con le proprietà del valore atteso:

NB: dimostrazione dalla triennale che non fa uso di Jensen

- supposing $X = a$ almost surely ($\mathbb{P}(X = a) = 1$), then $\mathbb{E}[X] = a$ and also $\mathbb{E}\left[X^2\right] = a^2$, thus

  $$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}[X]\right)^2 = a^2 - a^2 = 0$$

- otherwise suppose

  $$0 = \mathrm{Var}[X] = \mathbb{E}\left[(X-\mathbb{E}[X])^2\right]$$

  Ma $(X - \mathbb{E}[X])^2 \geq 0$ e quindi per la proprietà 4 del valore atteso (strict positivity), se fosse $\mathbb{P}(X \neq \mathbb{E}[X]) > 0$ si avrebbe $\mathbb{P}\left((X - \mathbb{E}[X])^2 > 0\right) > 0$ e quindi $\mathrm{Var}[X] = \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] > 0$.
  Quindi deve essere $\mathbb{P}(X \neq \mathbb{E}[X]) = 0$ ossia $\mathbb{P}(X = \mathbb{E}[X]) = 1$ ovvero X degenere con $a = \mathbb{E}[X]$

□

Remark 75. Generalizing a bit, variance is the *second* moment of a random variable with respect to its mean.

Definition 2.6.6 (r-th moments of X with respect to mean)**.** In the definition 2.6.1 is obtained by adopting as g the r-power of difference between X and its expected value, $g = (x - \mathbb{E}[X])^r$:

$$\bar{\mu_r} = \mathbb{E}[(X - \mathbb{E}[X])^r] = \begin{cases} \sum_i (x_i - \mathbb{E}[X])^r \cdot p_X(x_i) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^r \cdot f_X(x)\,\mathrm{d}x & \text{se } X \text{ è continua} \end{cases} \tag{2.46}$$

Remark 76. Since $\bar{\mu_0} = 1, \bar{\mu_1} = 0$, these moments become interesting starting from $r = 2$.

Proposition 2.6.7 (Properties of variance)**.** *Given $a, b, c \in \mathbb{R}$:*

$$\operatorname{Var}[X] \geq 0 \tag{2.47}$$

$$\operatorname{Var}[X] = 0 \iff \mathbb{P}(X = c) = 1 \tag{2.48}$$

$$\operatorname{Var}[aX + b] = a^2 \operatorname{Var}[X] \tag{2.49}$$

$$X \perp\!\!\!\perp Y \implies \operatorname{Var}[X + Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] \tag{2.50}$$

Proof. Per la 2.47, la varianza è il valore atteso della vc nonnegativa $(X - \mathbb{E}[X])^2$, motivo per cui è non negativa date le proprietà del valore atteso.
Per 2.48 se $\mathbb{P}(X = c) = 1$ per qualche costante c allora $\mathbb{E}[X] = c$ e $\mathbb{E}\left[X^2\right] = c^2$, pertanto $\operatorname{Var}[X] = 0$; viceversa se $\operatorname{Var}[X] = 0$ allora $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = 0$ che mostra che $(X - \mathbb{E}[X])^2 = 0$ ha probabilità 1, che a sua volta mostra che X è uguale alla sua media con probabilità 1.
Per la 2.49 e per la linearità del valore atteso si ha:

$$\begin{aligned} \operatorname{Var}[aX + b] &= \mathbb{E}\left[(aX + b - (a\,\mathbb{E}[X] + b))^2\right] \\ &= \mathbb{E}\left[(aX + b - a\,\mathbb{E}[X] - b)^2\right] \\ &= \mathbb{E}\left[(aX - a\,\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[a^2(X - \mathbb{E}[X])^2\right] \\ &= a^2\,\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= a^2 \operatorname{Var}[X] \end{aligned}$$

La 2.50 verrà dimostrata/generalizzata in seguito, per ora verifichiamola:

$$\begin{aligned} \operatorname{Var}[X + Y] &= \mathbb{E}\left[(X + Y)^2\right] - (\mathbb{E}[X + Y])^2 = \mathbb{E}\left[X^2 + 2XY + Y^2\right] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\ &\overset{(1)}{=} \mathbb{E}\left[X^2\right] + 2\,\mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}\left[Y^2\right] - \mathbb{E}[X]^2 - 2\,\mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\ &= \operatorname{Var}[X] + \operatorname{Var}[Y] \end{aligned}$$

where in (1) we used that if $X \perp\!\!\!\perp Y$ we have $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$. ☐

Remark 77 (Variance is nonlinear)*.* Differently from expected value a is squared and b omitted, therefore variance of sum of different random variable could be different from sum of their variance.

Definition 2.6.7 (Standard deviation)**.**

$$\sigma = \sigma_X = \sqrt{\operatorname{Var}[X]} \tag{2.51}$$

2.6.3 Asymmetry/skewness and kurtosis

Definition 2.6.8 (Standardized rvs)**.** Given any RV X such that $\mathbb{E}\left[X^2\right] < +\infty$ and variance $\mathrm{Var}[X] \in (0, +\infty)$, standardized rv Z is defined as:

$$Z = \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}[X]}} = \frac{X - \mathbb{E}[X]}{\sigma} \tag{2.52}$$

Remark 78. Note that $\mathbb{E}[Z] = 0$ and $\mathrm{Var}[Z] = 1$:

$$\begin{aligned} \mathbb{E}[Z] &= \frac{\mathbb{E}[X] - \mathbb{E}[X]}{\sqrt{\mathrm{Var}[X]}} = 0 \\ \mathrm{Var}[Z] &= \mathrm{Var}\left[\frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}[X]}}\right] = \frac{1}{\mathrm{Var}[X]} \cdot \mathrm{Var}[X - \mathbb{E}[X]] = \frac{\mathrm{Var}[X]}{\mathrm{Var}[X]} = 1 \end{aligned}$$

Thus transform make rv independent from measure unit.

Definition 2.6.9 (r-th standardized moments of X)**.** We have them if $g = \left(\frac{x - \mathbb{E}[X]}{\sigma}\right)^r$:

$$\bar{\bar{\mu}}_r = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^r\right] = \begin{cases} \displaystyle\sum_i \left(\frac{x_i - \mathbb{E}[X]}{\sigma}\right)^r \cdot p_X(x_i) & \text{se } X \text{ è discreta} \\ \displaystyle\int_{-\infty}^{+\infty} \left(\frac{x - \mathbb{E}[X]}{\sigma}\right)^r \cdot f_X(x)\,\mathrm{d}x & \text{se } X \text{ è continua} \end{cases} \tag{2.53}$$

Remark 79. Since for any rv $\bar{\bar{\mu}}_0 = 1$, $\bar{\bar{\mu}}_1 = 0$, $\bar{\bar{\mu}}_2 = 1$ moments of interest are where $r = 3$ and $r = 4$.

2.6.3.1 Asymmetry/Skewness

Definition 2.6.10 (Symmetric rv)**.** X is symmetric (respect to $\mathbb{E}[X]$) if $X - \mathbb{E}[X]$ has the same distribution of $\mathbb{E}[X] - X$.

Remark 80 (Intuizione significato)*.* $X - \mathbb{E}[X]$ sposta la densità/probabilità, così com'è, centrandola sullo 0. Intuitivamente $-X$ ha l'effetto di ottenere la densità probabilità simmetrica/specchiata rispetto a $x = 0$; infine $-X + \mathbb{E}[X]$ specchia la densità/probabilità rispetto a 0 e poi la ricentra su 0. Pertanto se $X - \mathbb{E}[X]$ e $-X + \mathbb{E}[X]$ coincidono, è perché la distribuzione di partenza X è simmetrica rispetto al centro.

Proposition 2.6.8 (Simmetria di una vc continua (PDF))**.** *Sia X una vc continua con PDF f. Allora è simmetrica su $\mathbb{E}[X]$ se e solo se $f(x) = f(2\,\mathbb{E}[X] - x)$.*

Remark 81. La definizione è meramente quella di una funzione simmetrica rispetto a $x = \mu$ (vedi calcolo).

Proof. Sia F la CDF di X; dimostriamo la doppia implicazione.
Se la simmetria vale ($X - \mathbb{E}[X] = \mathbb{E}[X] - X$) abbiamo:

$$\begin{aligned} F(x) &= \mathbb{P}(X - \mathbb{E}[X] \le x - \mathbb{E}[X]) \overset{(1)}{=} \mathbb{P}(\mathbb{E}[X] - X \le x - \mathbb{E}[X]) \overset{(2)}{=} \mathbb{P}(X \ge 2\,\mathbb{E}[X] - x) \\ &= 1 - F(2\,\mathbb{E}[X] - x) \end{aligned}$$

dove in (1) abbiamo sfruttato la simmetria ($X - \mathbb{E}[X] = \mathbb{E}[X] - X$) e in (2) abbiamo elaborato algebricamente. Facendo la derivata dei membri estremi dell'equazione si ottiene $f(x) = f(2\mathbb{E}[X] - x)$.
Viceversa supponendo che $f(x) = f(2\mathbb{E}[X] - x)$ valga $forallx$, vogliamo dimostrare che $\mathbb{P}(X - \mathbb{E}[X] \leq t) = \mathbb{P}(\mathbb{E}[X] - X \leq t)$, ossia vi è simmetria e le cumulate CDF coincidono. Si ha

$$\mathbb{P}(X - \mathbb{E}[X] \leq t) = \mathbb{P}(X \leq \mathbb{E}[X] + t) = \int_{-\infty}^{\mathbb{E}[X]+t} f(x)\,\mathrm{d}x \overset{(1)}{=} \int_{-\infty}^{\mathbb{E}[X]+t} f(2\mathbb{E}[X] - x)\,\mathrm{d}x$$
$$\overset{(2)}{=} \int_{\mathbb{E}[X]-t}^{\infty} f(w)\,\mathrm{d}w = \mathbb{P}(\mathbb{E}[X] - X \leq t)$$

dove in abbiamo sfruttato che $f(x) = f(2\mathbb{E}[X] - x)$, mentre in (2) deve avvenire qualche trick di integrazione (integra $f(-x)$ ad indici invertiti e moltiplicati direi). □

Definition 2.6.11 (Skewness)**.** It's the 3-rd standardized moment:

$$\mathrm{Asym}(X) = \bar{\bar{\mu}}_3 = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^3\right] \tag{2.54}$$

Remark 82. A negative skewness means a left longer tail, while positive a right longer one.

2.6.3.2 Kurtosis

Definition 2.6.12 (Kurtosis)**.** It's the 4-th standardized moment

$$\mathrm{Kurt}(X) = \bar{\bar{\mu}}_4 = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^4\right] \tag{2.55}$$

Remark 83. Some defines kurtosis by centering on 3 (value assumed by the normal) as in:

$$\mathrm{Kurt}(X) = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^4\right] - 3 \tag{2.56}$$

In this way the normal will have 0 kurtosis and the remaining a value a negative or positive value, related to givin less or more weight to the tail of the distribution.

Remark 84. Una distribuzione con eccesso di curtosi (2.56) negativo (detta *platicurtica*) tende ad avere un profilo più piatto della normale e una minore importanza delle code. Produce outlier in misura minore o meno estremi rispetto alla normale. Un esempio è l'uniforme.
Viceversa una distribuzione con eccesso di curtosi positivo è detta *leptocurtica* (ad esempio distribuzione T di Student, logistica, Laplace): ha code che si avvicinano allo zero più lentamente rispetto una gaussiana, per cui produce più outlier della stessa.
In fig 2.3 alcune distribuzioni (con media 0 e varianza 1) e relativa curtosi.

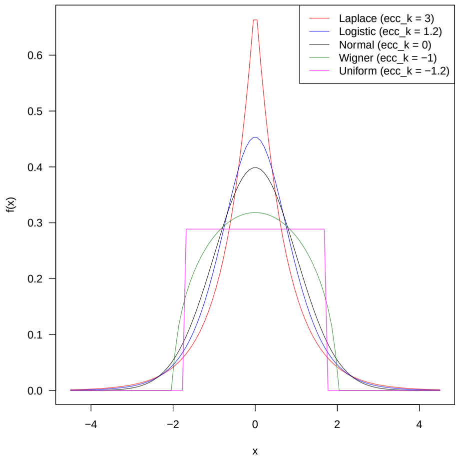


Figure 2.3: PDF for some rv (mean 0, variance 1) and their kurtosis

2.7 Random vectors

2.7.1 Random vectors and their distribution

Definition 2.7.1. A random vector X (or n-variate random variable) is a function $X : \Omega \to \mathbb{R}^n$ that maps the occurrence of the experiment to a real vector of n components. It's denoted as

$$X = \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix}$$

where $X_1, \dots X_n$ are real random variables.

Example 2.7.1 (Two dice roll)**.** With $\Omega = \{\{1,1\}, \dots, \{6,6\}\}$, we could construct following are *bivariate* random vectors:

- $X = (X_1, X_2)$ with X_1 outcome for the first dice, X_2 outcome of the second one;
- $X = (X_1, X_2)$ with X_1 sum of the two dice, X_2 difference

Important remark 29 (Probability of event E). It is defined as

$$\nu(E) = \mathbb{P}(X \in E), \qquad \forall E \in \beta(\mathbb{R}^n)$$

Definition 2.7.2 (Distribution function)**.** Distribution of random vector X is a function $F : \mathbb{R}^n \to \mathbb{R}$ defined as

$$F(x_1, \dots, x_n) = \mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) \tag{2.57}$$

Remark 85 (Remarks on distribution function). Again

- there's a 1-to-1 correspondance between F and ν expressed by
  $$F(x_1, \dots, x_n) = \nu((-\infty, x_1] \times \dots \times (-\infty, x_n]), \qquad \forall (x_1, \dots, x_n) \in \mathbb{R}^n$$
- F determines the probability distribution ν of X in the sense that
  $$X \sim Y \iff X \text{ and } Y \text{ have the same distribution function}$$

2.7.2 Type of random vectors

Important remark 30 (Types of random vectors). Random vector X is

- **multivariate discrete** iff $\exists B \subset \mathbb{R}^n$, B finite or countable such that $\mathbb{P}(X \in B) = 1$
- **multivariate absolutely continuous** iff exists a density (called *joint*) $f : \mathbb{R}^n \to \mathbb{R}$ such that
  1. $f \geq 0$
  2. f is integrable
  3. distribution is defined as integral of density

  $$F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \dots, t_n)\, \mathrm{d}t_1 \dots\, \mathrm{d}t_n, \qquad \forall (x_1, \dots, x_n) \in \mathbb{R}^n$$

  and for which

  $$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(t_1, \dots, t_n)\, \mathrm{d}t_1 \dots\, \mathrm{d}t_n = 1$$

- **singular continuous** (ignored) is not easily to handle (splits in several cases)

Remark 86. Essentially the same remarks done for random variable holds.
Next we extend to random vector the characterization theorem (already discussed in the $n = 1$ case) useful for proving that X is absolutely continuous. For this we need the concept of Lebesgue measure in 2+ dimension.

Definition 2.7.3 (Lebesgue measure on $\mathbb{R}^n$)**.** It's the only measure on $\beta(\mathbb{R}^n)$ such that the measure of the cartesian product of interval is equal to the product of the length of the intervals:

$$m(I_1 \times \ldots \times I_n) = len(I_1) \cdot \ldots \cdot len(I_n), \quad \forall I_i$$

where $len(I_i)$ is the legth of the interval I_i (eg if $I_i = [a, b]$ that is $b - a$).

Example 2.7.2. Intuitively, if $A \in \beta(\mathbb{R}^2)$ then $m(A)$ is the area of A; in $\beta(\mathbb{R}^3)$ is a volume and so on.

Theorem 2.7.1 (Absolutely continuous random vector characterization)**.** *A random vector X is absolutely continuous if and only if any set (event) with null lebesgue measure has null probability as well*

$$X \textit{ is absolutely continuous} \iff \begin{cases} \mathbb{P}(X \in E) = 0 \\ \forall E \in \beta(\mathbb{R}^n), \textit{such that } m(E) = 0 \end{cases}$$

Example 2.7.3. In a distribution in 2d (X_1, X_2), if any point x_i, y_i (which has a 2d lebesgue measure of 0) has zero probability then that is an absolutely continuous random variable.

Theorem 2.7.2. *If $X_1, \dots, X_n$ are absolutely continuous, this does not imply that the vector X is absolutely continuous.*

Example 2.7.4. As example of X not being absolutely continuous even if $X_1, \ldots, X_n$ are follows.
With $n = 2$, consider $X_1 \sim \mathrm{N}(0, 1)$ (absolutely continuos because it's a standard normal), $X_2 = X_1$ (equal, so absolutely continuous). Is $\mathbf{X} = (X_1, X_2)$ absolutely continuous?
To check that X is not absolutely continuous, we apply the theorem, letting the event to be we extracted a point on the diagonal $y = x$

$$E = \left\{(x, y) \in \mathbb{R}^2 : x = y\right\}$$

We have that:

- $\mathbb{P}(X \in E) = 1$: infact once we extracted $X_1 = x_1$ we have $X_2 = x_1$ as well so the vector will be on the diagonal $y = x$;
- however $m(E) = 0$ (that is the area of the line $y = x$ compared to 2 space dimension $\mathbb{R}^2$ is 0).

So X is not absolutely continuous

2.7.3 Marginals

Definition 2.7.4 (Marginal of random vector $X = (X_1, \ldots, X_n)^\top$)**.** It is any subvector $(X_{j_1}, \ldots, X_{j_k})$ where $\{j_1, \ldots, j_k\}$ is a subset of $1, \ldots, n$.

Remark 87. It is just a vector with less random variables; there are

- n marginals of only 1 variable (these are $X_1 \ldots X_n$);
- $\binom{n}{2}$ marginals of 2 random variables $(X_i, X_j)^\top$;
- $\binom{n}{3}$ marginals of 3 random variables $(X_i, X_j, X_k)^\top$;

Theorem 2.7.3 (Density of marginal of absolute continuous random vectors)**.** *If $X = (X_1, \ldots, X_n)$ is absolutely continuous with multivariate density f*

- *all marginal of X are still absolutely continuous (converse is not true in general but special case, see thm 2.7.5);*
- *the density g of the marginal $(X_1, \ldots, X_k)^t$ is obtained by making $n - k$ integral of f, that is integrating on the remaining $n-k$ variables one wants to eliminate:*

$$g(x_1, \ldots, x_k) = \underbrace{\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty}}_{n-k \ \text{integrals}} f(t_1, \ldots, t_n) \, \mathrm{d}t_{k+1} \ldots \, \mathrm{d}t_n \tag{2.58}$$

Example 2.7.5. If $n = 3$, $\mathbf{X} = (X, Y, Z)^\top$

- the density of $(Y, Z)^\top$ is
$$g(y, z) = \int_{-\infty}^{+\infty} f(x, y, z) \, \mathrm{d}x$$
- densitity of Y is
$$g(y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y, z) \, \mathrm{d}x \, \mathrm{d}z$$

2.7.4 Independence

Theorem 2.7.4 (Independence)**.** *Let $X = (X_1, \dots, X_n)^\top$ be* any *random vector with distribution function F then $X_1, \dots, X_n$ are indipendent if and only if the joint distribution function F is the product of the marginal distribution functions F_i:*

$$X_1, \dots, X_n \text{ are independent} \iff F(x_1, \dots, x_n) = F_1(x_1) \cdot \ldots \cdot F_n(x_n), \quad \forall \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix} \in \mathbb{R}^n$$

Similarly if X is absolutely continuous *we can replace distribution with densities, that is:*

$$X_1, \dots, X_n \text{ are independent} \iff f(x_1, \dots, x_n) = f_1(x_1) \cdot \ldots \cdot f_n(x_n), \quad \forall \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix} \in \mathbb{R}^n$$

Theorem 2.7.5 (Independence and absolutely continuous random vectors)**.** *In $X = (X_1, \dots, X_n)$, if $X_1, \dots, X_n$ are independent then*

$$X \text{ is absolutely continuous} \iff X_1, \dots, X_n \text{ are absolutely continuous}$$

NB: non mostrato da rigo ma usato, trovato su `https://math.stackexchange.com/questions/606205`

Theorem 2.7.6 (Independence and expected value)**.** *If $X \perp\!\!\!\perp Y$ then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$*

Proof. In case of absolutely continuous random variables, by definition they are independent if $f_{XY}(x, y) = f_X(x) f_Y(y)$. Then we have

$$\begin{aligned}
\mathbb{E}[XY] &= \int_{-\infty}^{+\infty} x \cdot y \cdot f_{XY}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_{-\infty}^{+\infty} x \cdot y \cdot f_X(x) f_Y(y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_{-\infty}^{+\infty} x \cdot f_X(x)\, \mathrm{d}x \int_{-\infty}^{+\infty} y \cdot f_Y(y)\, \mathrm{d}y \\
&= \mathbb{E}[X]\,\mathbb{E}[Y]
\end{aligned}$$

while the proof in the dicsrete case is analogous □

2.8 Relationship between RVs

2.8.1 Covariance

Definition 2.8.1 (Covariance)**.** Considering two random variables X and Y, if $\mathbb{E}[|X|] < +\infty$, $\mathbb{E}[|Y|] < +\infty$ and $\mathbb{E}[|XY|] < +\infty$ we can define the covariance as

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{2.59}$$

Remark 88. La covarianza misura la forza del legame lineare tra X e Y

Important remark 31. A sufficient condition for the existence of $\operatorname{Cov}(X, Y)$ is that $\mathbb{E}\left[X^2\right] < +\infty$ and $\mathbb{E}\left[Y^2\right] < +\infty$

Proposition 2.8.1 (Another useful formula)**.**

$$\mathrm{Cov}\left(X,Y\right)=\mathbb{E}\left[XY\right]-\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \tag{2.60}$$

Proof. We have

$$\begin{aligned}\mathrm{Cov}\left(X,Y\right)&=\mathbb{E}\left[(X-\mathbb{E}\left[X\right])(Y-\mathbb{E}\left[Y\right])\right]\\&=\mathbb{E}\left[XY-\mathbb{E}\left[X\right]\cdot Y-\mathbb{E}\left[Y\right]\cdot X+\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]\right]\\&=\mathbb{E}\left[XY\right]+\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]-2\,\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]\\&=\mathbb{E}\left[XY\right]-\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]\end{aligned}$$

□

Example 2.8.1. In particular if $Y=X$

$$\mathrm{Cov}\left(X,X\right)=\mathbb{E}\left[X^2\right]-\left(\mathbb{E}\left[X\right]\right)^2=\mathrm{Var}\left[X\right]$$

Proposition 2.8.2 (Covariance and independence)**.** *Assuming the covariance exists,* $X\perp\!\!\!\perp Y \implies \mathrm{Cov}\left(X,Y\right)=0$*. The converse implication is false.*

Proof. If $X\perp\!\!\!\perp Y$, then $\mathbb{E}\left[XY\right]=\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$ so the covariance is 0. A counterexample for counterimplication follows. □

Example 2.8.2 (Counterexample where $\mathrm{Cov}\left(X,Y\right)=0$ but X,Y are not independent)**.** Let $X\sim\mathrm{N}\left(0,1\right)$ and $Y=X^2$. Let's prove:

- $\mathrm{Cov}\left(X,Y\right)=0$. We have that

  $$\mathrm{Cov}\left(X,Y\right)=\mathbb{E}\left[XY\right]-\underbrace{\mathbb{E}\left[X\right]}_{=0}\mathbb{E}\left[Y\right]=\mathbb{E}\left[XY\right]=\mathbb{E}\left[X^3\right]$$

  Since X is absolutely continuous (normal) the expectation of X to the power 3 can be wrritten as

  $$\mathbb{E}\left[X^3\right]=\int_{-\infty}^{+\infty}x^3\cdot\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

  and since the integrand it's an odd function evaluated on a simmetric interval, the integral is 0

- $X\not\perp\!\!\!\perp Y$. It's intuitive these are not independent, however let's prove it formally. To prove that we consider the probability $\mathbb{P}\left(-1\leq X\leq 1,Y>1\right)$ which under independence should be equal to $\mathbb{P}\left(-1\leq X\leq 1\right)\cdot\mathbb{P}\left(Y>1\right)$. Now actually:

  $$\begin{aligned}&\mathbb{P}\left(|X|\leq 1,Y>1\right)\overset{(1)}{=}\mathbb{P}\left(|X|\leq 1,|X|>1\right)=\mathbb{P}\left(\emptyset\right)=0\\&\mathbb{P}\left(-1\leq X\leq 1\right)\cdot\mathbb{P}\left(Y>1\right)=\underbrace{\mathbb{P}\left(|X|\leq 1\right)}_{>0}\cdot\underbrace{\mathbb{P}\left(|X|>1\right)}_{>0}>0\end{aligned}$$

  where in (1) since $Y=X^2$.
  Since the first is null and the second positive, they can't be equal and so random variables are not independent

Important remark 32 (Special case). There is an important special case in which independence amounts to null covariance:

$$\text{If } \begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathrm{N}\,, \text{ then } X \perp\!\!\!\perp Y \iff \mathrm{Cov}\,(X,Y) = 0$$

Proposition 2.8.3 (Variance of sum of RVs). *Assuming the covariances exists, the variance of the sum of random variables is*

$$\begin{aligned} \mathrm{Var}\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}\,[X_i] + \sum_{i \neq j} a_i a_j \,\mathrm{Cov}\,(X_i, X_j) \\ &\overset{(1)}{=} \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}\,[X_i] + 2 \sum_{1 \leq i < j \leq n} a_i a_j \,\mathrm{Cov}\,(X_i, X_j) \end{aligned} \tag{2.61}$$

where (1) *because* $\mathrm{Cov}\,(X_i, X_j) = \mathrm{Cov}\,(X_j, X_i)$

Important remark 33. If $X_1, \ldots, X_n$ are independent, this formula reduces to

$$\mathrm{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}\,[X_i]$$

Example 2.8.3. With $n = 2$, by just letting $a_1 = 1, a_2 = \pm 1$

$$\begin{aligned} \mathrm{Var}\,[X+Y] &= \mathrm{Var}\,[X] + \mathrm{Var}\,[Y] + 2\,\mathrm{Cov}\,(X,Y) \\ \mathrm{Var}\,[X-Y] &= \mathrm{Var}\,[X] + \mathrm{Var}\,[Y] - 2\,\mathrm{Cov}\,(X,Y) \end{aligned}$$

In case X, Y are indipendent covariance is null and 1) variance of sum is sum of variance 2) $\mathrm{Var}\,[X+Y] = \mathrm{Var}\,[X-Y]$

Example 2.8.4. Sia $X \perp\!\!\!\perp Y$,

$$\mathrm{Var}\,[X - Y + 2Z] = \mathrm{Var}\,[X] + \mathrm{Var}\,[Y] + 4\,\mathrm{Var}\,[Z] + 4\,\mathrm{Cov}\,(X,Z) - 4\,\mathrm{Cov}\,(Y,Z)$$

Proposition 2.8.4 (Proprietà covarianza (wikipedia, non fatte da rigo)). *If* X, Y, W, V *are real-valued random variables and* $a, b, c, d \in \mathbb{R}$*, then the following facts are a consequence of the definition of covariance:*

$$\mathrm{Cov}\,(X,a) = 0 \tag{2.62}$$

$$\mathrm{Cov}\,(X,X) = \mathrm{Var}\,[X] \tag{2.63}$$

$$\mathrm{Cov}\,(X,Y) = \mathrm{Cov}\,(Y,X) \tag{2.64}$$

$$\mathrm{Cov}\,(aX,bY) = ab\,\mathrm{Cov}\,(X,Y) \tag{2.65}$$

$$\mathrm{Cov}\,(X+a,Y+b) = \mathrm{Cov}\,(X,Y) \tag{2.66}$$

$$\mathrm{Cov}\,(aX+bY,cW+dV) = ac\,\mathrm{Cov}\,(X,W) + ad\,\mathrm{Cov}\,(X,V) + bc\,\mathrm{Cov}\,(Y,W) + bd\,\mathrm{Cov}\,(Y,V) \tag{2.67}$$

$$X \perp\!\!\!\perp Y \implies \mathrm{Cov}\,(X,Y) = 0 \tag{2.68}$$

Example 2.8.5 (Esame vecchio viroli). Let X_1 and X_2 be two random variables with distribution $X_1 \sim \mathrm{N}\,(0,2)$ and $X_n \sim \mathrm{N}\,(-2,1)$ (parameters are mean and variance) and covariance -1. Compute $\mathrm{Cov}\,(X_1 + X_2, X_1 - X_2)$.
We have that

$$\begin{aligned} \mathrm{Cov}\,(X_1 + X_2, X_1 - X_2) &= \mathrm{Cov}\,(X_1,X_1) - \mathrm{Cov}\,(X_1,X_2) + \mathrm{Cov}\,(X_2,X_1) - \mathrm{Cov}\,(X_2,X_2) \\ &= \mathrm{Var}\,[X_1] - \mathrm{Var}\,[X_2] = 2 - 1 = 1 \end{aligned}$$

Example 2.8.6 (Esame vecchio viroli)**.** Let X_1, X_2 be two standard gaussian variables with covariance -1. Compute $\text{Cov}\left(X_1 + X_2, X_1 - X_2\right)$.
With the same developmet as above we have:

$$\text{Cov}\left(X_1 + X_2, X_1 - X_2\right) = \text{Var}\left[X_1\right] - \text{Var}\left[X_2\right] = 1 - 1 = 0$$

Example 2.8.7 (Esame vecchio viroli)**.** Let X and Y be two independent bernoulli random variables with same parameter p. Compute $\text{Cov}\left(Y - X, 2X + 2Y\right)$.

$$\begin{aligned}\text{Cov}\left(Y - X, 2X + 2Y\right) &= 2\,\text{Cov}\left(X, Y\right) + 2\,\text{Cov}\left(Y, Y\right) - 2\,\text{Cov}\left(X, X\right) - 2\,\text{Cov}\left(X, Y\right) \\ &= 2\,\text{Var}\left[X\right] - 2\,\text{Var}\left[Y\right] = 0\end{aligned}$$

taluni suggeriscono -1 ma mi pare na gran cacata

```
p = 0.5
x = rbinom(100000, 1, 0.5)
y = rbinom(100000, 1, 0.5)
cov(y-x, 2*x+2*y)

## [1] -3.897039e-06
```

Example 2.8.8 (Esame vecchio viroli)**.** Let $X = (X_1, X_2)$ be a bivariate gaussian vector with $\mu = [0, 0]$ and

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

What is the distribution of $Y = 3X_1 - 2X_2$?
Si ha che

$$\begin{aligned}\mathbb{E}\left[Y\right] &= \mathbb{E}\left[3X_1 - 2X_2\right] = 3\,\mathbb{E}\left[X_1\right] - 2\,\mathbb{E}\left[X_2\right] = 0 \\ \text{Var}\left[Y\right] &= \text{Var}\left[3X_1 - 2X_2\right] \\ &= 3^2\,\text{Var}\left[X_1\right] + (-2)^2\,\text{Var}\left[X_2\right] + 2(3\cdot(-2))\,\text{Cov}\left(X_1, X_2\right) \\ &= 9\cdot 1 + 4\cdot 1 + 2\cdot(-6)\cdot 0.5 = 7\end{aligned}$$

quindi è $Y \sim \text{N}\left(0, 7\right)$ come confermato da taluni

2.8.2 Correlation coefficient

Definition 2.8.2 (Correlation coefficient)**.** Considered two rv X, Y, if $\mathbb{E}\left[X^2\right] < +\infty$, $\mathbb{E}\left[Y^2\right] < +\infty$, $\text{Var}\left[X\right] > 0$, $\text{Var}\left[Y\right] > 0$, we can define the correlation coefficent as

$$\rho(X, Y) = \text{Corr}\left(X, Y\right) = \frac{\text{Cov}\left(X, Y\right)}{\sqrt{\text{Var}\left[X\right]}\sqrt{\text{Var}\left[Y\right]}} \tag{2.69}$$

Proposition 2.8.5. *Some properties:*

- *it can be easily seen that correlation coefficient can be written as covariance between the two standardized variables (def 2.6.8)*

$$\text{Corr}\left(X, Y\right) = \text{Cov}\left(\frac{X - \mathbb{E}\left[X\right]}{\sqrt{\text{Var}\left[X\right]}}, \frac{Y - \mathbb{E}\left[Y\right]}{\sqrt{\text{Var}\left[Y\right]}},\right)$$

quindi $\operatorname{Cov}(X,Y)$ *e* $\operatorname{Corr}(X,Y)$ *danno essenzialmente la stessa informazione, entrambi misurano l'intensità del legame lineare tra* X *ed* Y. $\operatorname{Corr}(\cdot,\cdot)$ *ha il vantaggio di essere normalizzato*

- *it ranges in*

$$-1 \leq \operatorname{Corr}(X,Y) \leq 1$$

Note that this amounts to

$$\operatorname{Cov}(X,Y)^2 \leq \operatorname{Var}[X] \operatorname{Var}[Y]$$

Quindi se il legame lineare tra X *ed* Y *è molto alto* $\operatorname{Corr}(X,Y)$ *è vicino a 1 in termini assoluti, ma non riesco ad esprimere questo fatto in termini di covarianza perché quest'ultima dipende dalle unità di misura di* X *ed* Y

- *the limit cases are the following:*

$$\begin{aligned} \operatorname{Corr}(X,Y) = 1 &\iff Y = a + bX, b > 0 \\ \operatorname{Corr}(X,Y) = -1 &\iff Y = a + bX, b < 0 \end{aligned}$$

Example 2.8.9 (Esame vecchio viroli)**.** Let X and Y be two gaussian variables with zero mean $\operatorname{Var}[X] = 1$, $\operatorname{Var}[Y] = 9$, covariance -1, compute $\rho(X+Y,X)$.

$$\begin{aligned} \operatorname{Corr}(X+Y,X) &= \frac{\operatorname{Cov}(X+Y,X)}{\sqrt{\operatorname{Var}[X+Y]}\sqrt{\operatorname{Var}[X]}} = \frac{\operatorname{Cov}(X,X) + \operatorname{Cov}(Y,X)}{\sqrt{\operatorname{Var}[X] + \operatorname{Var}[Y] + 2\operatorname{Cov}(X,Y)}\sqrt{\operatorname{Var}[X]}} \\ &= \frac{\operatorname{Var}[X] + \operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}[X] + \operatorname{Var}[Y] + 2\operatorname{Cov}(X,Y)}\sqrt{\operatorname{Var}[X]}} = \frac{1 + (-1)}{\sqrt{1 + 9 + 2(-1)}\sqrt{1}} \\ &= 0 \end{aligned}$$

Il risultato è confermato dal Bigo.

Example 2.8.10 (Esame vecchio viroli)**.** Let X and Y be two gaussian variables with zero mean $\operatorname{Var}[X] = 1$, $\operatorname{Var}[Y] = 9$, covariance -1, compute $\rho(1-2X+2, 3+Y)$.
We have:

$$\begin{aligned} \operatorname{Corr}(1-2X+2, 3+Y) = \operatorname{Corr}(3-2X, 3+Y) &= \frac{\operatorname{Cov}(3-2X, 3+Y)}{\sqrt{\operatorname{Var}[3-2X]}\sqrt{\operatorname{Var}[3+Y]}} \\ &= \frac{\operatorname{Cov}(3,3) + \operatorname{Cov}(3,Y) + \operatorname{Cov}(-2X,3) + \operatorname{Cov}(-2X,Y)}{\sqrt{4\operatorname{Var}[X]}\sqrt{\operatorname{Var}[Y]}} \\ &= \frac{0+0+0+2}{2 \cdot 1 \cdot 3} = \frac{1}{3} \end{aligned}$$

come confermato da taluni

2.9 Exercises

Example 2.9.1 (Es crash course, giorno 1)**.** Let X be a rv that has the density

$$f(x) = \begin{cases} ce^{-\lambda x} & if x \geq 0 \\ 0 & x < 0 \end{cases}$$

Find:

1. c
2. $\mathbb{E}[X]$
3. $\mathrm{Var}[X]$
4. $F(X)$

We have

1. it must be that
$$1 = \int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = \int_0^{+\infty} ce^{-\lambda x}\,\mathrm{d}x = c\int_0^{+\infty} e^{-\lambda x}\,\mathrm{d}x = c\left[\left(\frac{1}{-\lambda}e^{-\lambda x}\right)\right]_0^{+\infty}$$
$$= 0 - \frac{c}{-\lambda}\cdot 1$$
therefore $c = \lambda$ (this is the exponential distribution)

2. we have,
$$\mathbb{E}[X] = \int_0^{+\infty} x\cdot\lambda e^{-\lambda x}\,\mathrm{d}x = \lambda\int_0^{+\infty} x\cdot e^{-\lambda x}\,\mathrm{d}x$$
using integration by parts we have
$$\begin{aligned}\int xe^{-\lambda x}\,\mathrm{d}x &= x\left(-\frac{1}{\lambda}e^{-\lambda x}\right) - \int -\frac{1}{\lambda}e^{-\lambda x}\\ &= \left(-\frac{x}{\lambda}e^{-\lambda x}\right) + \frac{1}{\lambda}\int e^{-\lambda x}\\ &= \left(-\frac{x}{\lambda}e^{-\lambda x}\right) + \frac{1}{\lambda}\left(-\frac{1}{\lambda}e^{-\lambda x}\right)\end{aligned}$$
che opportunamente valutato
$$\left[\left(-\frac{x}{\lambda}e^{-\lambda x}\right) + \frac{1}{\lambda}\left(-\frac{1}{\lambda}e^{-\lambda x}\right)\right]_0^{+\infty} = 0 + 0 - \left(0 - \frac{1}{\lambda^2}\right)$$
Per cui tornando al valore atteso
$$\mathbb{E}[X] = \lambda\left(\frac{1}{\lambda^2}\right) = \frac{1}{\lambda}$$

3. first we find $\mathbb{E}\left[X^2\right]$
$$\mathbb{E}\left[X^2\right] = \lambda\int_{-\infty}^{+\infty} x^2e^{-\lambda x}\,\mathrm{d}x \overset{(1)}{=} \lambda\left[(x^2\frac{-1}{\lambda}e^{-\lambda x})\Big|_0^{\infty} + \frac{2}{\lambda}\int_0^{+\infty} xe^{-\lambda x}\,\mathrm{d}x\right]$$
$$= 2\int_0^{+\infty} xe^{-\lambda x}\,\mathrm{d}x = \frac{2}{\lambda^2}$$
where in (1) again by integration by parts. So
$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

4. we have

$$F(x) = \int_0^x f(s)\ \mathrm{d}s = \lambda \int_0^x e^{-\lambda s}\ \mathrm{d}s = \lambda\left(\frac{-1}{\lambda}e^{-\lambda s}\right)\Big|_0^x$$
$$= 1 - e^{\lambda x}, \quad \text{for } x \geq 0$$

for $x < 0, F(x) = \int_{-\infty}^x f(s)\ \mathrm{d}s = 0$ so

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{\lambda x} & x \geq 0 \end{cases}$$

Example 2.9.2 (crash course, day 1 es 3 pag 6)**.** Let $f(k) = \frac{c^k e^{-\lambda}}{k!}$ for $k \in \{0, 1, \ldots\}$ be the pmf that X satisfies:

1. find c
2. find $\mathbb{E}[X]$
3. find $\mathrm{Var}[X]$

we have

1.

$$\sum_{k=0}^{\infty} f(k) = 1 = e^{-\lambda} \underbrace{\sum_{k=0}^{\infty} \frac{c^k}{k!}}_{e^c} = e^{-\lambda}e^c = e^{c-\lambda} = 1 = e^0 \implies c = \lambda$$

2.

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} kf(k) = \sum_{k=0}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!}$$
$$\overset{(1)}{=} \lambda \underbrace{\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!}}_{F(\infty)=1} = \lambda$$

with (1) substituting $u = k - 1$. This is the poisson distribution, we say $X \sim \mathrm{Pois}(\lambda)$

3. first we find $\mathbb{E}\left[X^2\right]$, but first consider the following

$$\mathbb{E}[X(X-1)] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X] = \sum_{k=0}^{\infty} k(k-1)f(k) = \sum_{k=2}^{\infty} k(k-1)f(k)$$
$$= \sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=2}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-2)!} = \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2} e^{-\lambda}}{(k-2)!}$$
$$\overset{(1)}{=} \lambda^2 \underbrace{\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!}}_{F(\infty)=1} = \lambda^2 = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]$$

where in (1) doin subst $u = k - 2$. Therefore

$$\mathbb{E}\left[X^2\right] = \lambda^2 + \lambda \implies \mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Example 2.9.3 (crashcourse, day 1 es 3 pag 7)**.** Let $X \sim \text{Bin}(n,p)$, that is $\mathbb{P}(X = k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$.

TODO: da finire ma valuta se ne vale la pena, la binomiale è già sviluppata nella prossima sezione

Example 2.9.4 (crashcourse, day 1 es 4 pag 7)**.** Let $F(x) = \frac{c}{2}\left(1 - \frac{1}{x^2}\right)$ for $x \in [1, \infty)$:

1. obtain $f(x)$
2. obtain c
3. $\mathbb{E}[X]$
4. $\text{Var}[X]$

we have

1. $$f(x) = \frac{\partial}{\partial x} F(x) = \frac{\partial}{\partial x} \frac{c}{2}\left(1 - \frac{1}{x^2}\right) = \frac{c}{x^3}$$

2. $$c \int_1^\infty \frac{1}{x^3}\, \mathrm{d}x = c\left[-\frac{1}{2}\frac{1}{x^2}\right]_1^\infty = \frac{c}{2} = 1 \implies c = 2$$

3. $$2 \int_1^\infty x \frac{1}{x^3}\, \mathrm{d}x = 2 \int_1^\infty \frac{1}{x^2}\, \mathrm{d}x = 2\left[\frac{-1}{x}\right]_1^\infty = 2$$

4. first we find
$$\mathbb{E}\left[X^2\right] = 2 \int_1^\infty x^2 \frac{1}{x^3} = 2 \int_1^\infty \frac{1}{x}\, \mathrm{d}x = 2\left[\log x\right]_1^\infty = +\infty$$

Example 2.9.5 (crashcourse, day 1 es 5 pag 8)**.** Let $f(x) = ce^{-\frac{x^2}{2}}$, $x \in \mathbb{R}$:

1. find c
2. $\mathbb{E}[X]$
3. $\text{Var}[X]$

Respectively

1. we know $\int_{-\infty}^{+\infty} cf(x)\, \mathrm{d}x = 1$ so we can do this trick
$$\begin{aligned} 1 &= \underbrace{\int_{-\infty}^{+\infty} cf(x)\, \mathrm{d}x}_{1} \underbrace{\int_{-\infty}^{+\infty} cf(y)\, \mathrm{d}y}_{1} \\ &= c^2 \int_{-\infty}^{+\infty} f(x)f(y)\, \mathrm{d}x\, \mathrm{d}y = c^2 \int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}}\, \mathrm{d}x\, \mathrm{d}y \end{aligned}$$

Now trasforming variable to polar coordinates that is applying

$$\begin{cases} x = r\cos\theta \\ y = r\sin\theta \end{cases}, \quad r \in [0,\infty), \theta \in [0, 2\pi)$$

so that $x^2 + y^2 = r^2$ and $\mathrm{d}x\ \mathrm{d}y = r\ \mathrm{d}r\ \mathrm{d}\theta$ we have

$$1 = c^2 \int_0^{2\pi} \int_0^{\infty} e^{-\frac{r^2}{2}} r\ \mathrm{d}r\ \mathrm{d}\theta \overset{(1)}{=} c^2 \int_0^{2\pi} \underbrace{\int_0^{+\infty} e^{-u}\ \mathrm{d}u}_{=1}\ \mathrm{d}\theta$$
$$= c^2 \int_0^{2\pi} \mathrm{d}\theta = c^2 2\pi = 1 \implies c = \frac{1}{\sqrt{2\pi}}$$

where in (1) we substitute $u = \frac{r^2}{2}$ so $\mathrm{d}u = r\ \mathrm{d}r$.

2. $\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x)\ \mathrm{d}x$. We have that $f(x)$ is an even function:

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

it's symmetric. However we are interested in $\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x)\ \mathrm{d}x$ that is trying to find the area under an odd function. Now in general if we're trying to find

- odd functions: given that it's symmetric around origin, positive areas compensates with negative areas so it's integral (over $\mathbb{R}$) is 0 (this holds for any odd function).

- even functions: synce symmetric around y axis to calculate integral on region $(-\infty, \infty)$ we can double the integral on region $(0, \infty)$

Therefore our $\mathbb{E}[X] = 0$.

3. for the variance first get

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{+\infty} \overbrace{\underbrace{x^2}_{even} \underbrace{f(x)}_{even}}^{even}$$
$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}}\ \mathrm{d}x \overset{(1)}{=} \frac{2}{\sqrt{2\pi}} \int_0^{+\infty} x^2 e^{-\frac{x^2}{2}}\ \mathrm{d}x \overset{(2)}{=} \frac{2}{\sqrt{2\pi}} \int_0^{+\infty} \sqrt{2} u^{1/2} e^{-u}\ \mathrm{d}u$$
$$= \frac{2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} u^{1/2} e^{-u}\ \mathrm{d}u = \frac{2}{\sqrt{\pi}} \Gamma\left(\frac{3}{2}\right) = 1$$

where in (1) since it's even and (2) using the variable change $u = \frac{x^2}{2}$ therefore $\mathrm{d}u = x\ \mathrm{d}x$ and $x = \sqrt{2u}$.
$\Gamma((x))$ is called gamma function, we will be familiar with it in the next years, for now trust me that $\Gamma(x) = (x-1)\Gamma(x-1)$ and $\Gamma(1) = 1$ $\Gamma(1/2) = \sqrt{\pi}$ so for integers n $\Gamma(n) = (n-1)!$ but for our case $\Gamma(3/2) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2}$

Therefore in the end for $X \sim \mathrm{N}(0,1)$, $\mathbb{E}[X] = 0$ and $\mathrm{Var}[X] = 1$. This is called a standard rv. But in general normal rvs can have different mean and variance: the general case is denoted as $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+$ and this correspond to translation of a standard normal rv and then scaling it.
Let $Z \sim \mathrm{N}(0,1)$ and $X = \sigma Z + \mu$ then

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\sigma Z + \mu] = \sigma \underbrace{\mathbb{E}[Z]}_{=0} + \mu = \mu \\ \mathrm{Var}[X] &= \mathrm{Var}[\sigma Z + \mu] = \sigma^2 \underbrace{\mathrm{Var}[Z]} = 1 = \sigma^2\end{aligned}$$

2.9.1 Random vectors

Example 2.9.6 (Esame vecchio viroli)**.** Let $\mathbf{X} = (X, Y)^\top$ be a random vector with joint density

$$f(x, y) = ky$$

where $0 < x < y < 1$. Compute k.
In order to compute k it must be:

$$\begin{aligned}1 &= \int_0^1 \int_0^y ky \, \mathrm{d}x \, \mathrm{d}y = k \int_0^1 y \int_0^y 1 \, \mathrm{d}x \, \mathrm{d}y \\ &= k \int_0^1 y[x]_0^y \, \mathrm{d}y = k \int_0^1 y^2 \, \mathrm{d}y = k \left[\frac{y^3}{3}\right]_0^1 = \frac{k}{3}\end{aligned}$$

da cui $k = 3$

Example 2.9.7. Consider the function

$$f(x|y) = \begin{cases} \frac{y^x e^{-y}}{x!} & \text{for } x = 0, 1, 2, \ldots \text{ and } y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

1. if the marginal pdf of Y is $\mathrm{Exp}(1)$, what is the joint probability function of (X, Y)
2. derive the marginal probability function of X

We have:

1. for the joint probability

$$f_{X,Y}(x, y) = f(y) \cdot f(x|y) = e^{-y} \frac{y^x e^{-y}}{x!} = \frac{y^x e^{-2y}}{x!}$$

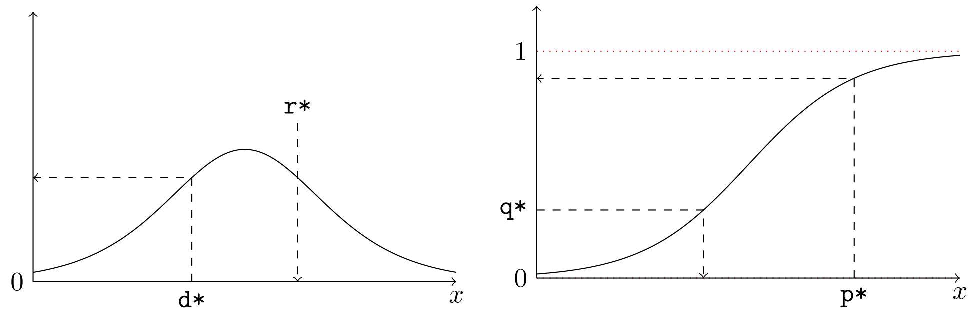
2. for the marginal probability of X

$$\begin{aligned}f_X(x) &= \int_0^{+\infty} \frac{y^x e^{-2y}}{x!} \, \mathrm{d}y = \frac{1}{x!} \underbrace{\int_0^{+\infty} y^x e^{-2y} \, \mathrm{d}y}_{(1)} \\ &= \frac{1}{x!} \frac{\Gamma(x+1)}{2^{x+1}} = \frac{1}{2^{x+1}}\end{aligned}$$

where (1) is the kernel of a $\mathrm{Gamma}(\alpha = x+1, \beta = 2)$

Function	Prefix	Family	Suffix	Family	Suffix
Density/Probability	`d`	Bernoulli	`binom`	Uniforme cont.	`unif`
PDF	`p`	Binomiale	`binom`	Esponenziale	`exp`
Quantile	`q`	Geometrica	`geom`	Normale	`norm`
RNG	`r`	Binomiale neg.	`nbinom`	Gamma	`gamma`
		Ipergeometrica	`hyper`	Chi-quadrato	`chisq`
		Poisson	`pois`	Beta	`beta`
		Uniforme disc.	$*$	T di Student	`t`
				F	`f`
				Logistica	`logis`
				Lognormale	`lnorm`
				Weibull	`weibull`
				Pareto (pac. `VGAM`)	`pareto`

Table 2.2: Utilities for family of rvs in R



(a) PMF/PDF (b) CDF

Figure 2.4: Funzioni in R

2.10 Probability models and R

Remark 89. In the following chapters we study main probabilistic models, which are the most used family of distribution

Remark 90. In:

- table 2.2 we report prefix of main functions and suffixes of main families;
- figure 2.4 function input needed (where arrows starts) and output returned (where arrow ends) for the 4 main functions.

Remark 91 (Variabili discrete con supporto finito in R). Per quanto riguarda la simulazione di queste (tra le quali l'uniforme discreta) si fa utilizzo della funzione `sample` alla quale, oltre a specificare l'urna `x`, il numero `size` di estrazioni desiderate, l'estrazione con reinserimento (`replace`) o meno, si possono specificare le probabilità `prob` di ciascun elemento nell'urna.

```
## DUnif(100)
sample(x = 1:100, size = 10, replace = TRUE)
```

```
##  [1] 38 40 36 75 62 92 53 79 44 29

## Urna discreta custom
sample(x = 1:3, prob = c(0.4, 0.4, 0.2), size = 10, replace = TRUE)

##  [1] 1 1 2 1 2 1 2 2 3 3
```

Chapter 3

Discrete random variables

Definition 3.0.1 (Family of random variables)**.** Set of distribution function $F(x; \boldsymbol{\Theta})$ having the same functional form+ but different for one or more parameters.

Definition 3.0.2 (Parameters space)**.** $\boldsymbol{\Theta}$, it's the set of possible value for the parameters of a distribution function.

3.1 Dirac

Definition 3.1.1 (Dirac rv (degenere))**.** $X \sim \delta_c$ if $\mathbb{P}(X = c) = 1$.

Proposition 3.1.1.

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \textit{if } x < c \\ 1 & \textit{if } x \geq c \end{cases} \tag{3.1}$$

Proposition 3.1.2 (Moments)**.**

$$\begin{aligned} &\mathbb{E}[X] = c \\ &\operatorname{Var}[X] = 0 \end{aligned}$$

Proof.

$$\begin{aligned} &\mathbb{E}[X] = c \cdot 1 = c \\ &\operatorname{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = c^2 \cdot 1 - c^2 = 0 \end{aligned}$$

□

Remark 92. Dirac is the only random variable having null variance.

3.2 Bernoulli

3.2.1 Definition

Remark 93. Viene utilizzata quando si ha a che fare con un esperimento il cui esito possibile è dicotomico (es $X = 1$ successo, $X = 0$ insuccesso).

Definition 3.2.1 (vc di Bernoulli). X is distributed as Bernoulli with parameter $0 \leq p \leq 1$, written $X \sim \text{Bern}(p)$, if $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$.

Remark 94. If $p = 0 \vee p = 1$ we obtain a Dirac.

3.2.2 Functions

Remark 95 (Support and parametric space).

$$\begin{aligned} R_X &= \{0, 1\} \\ \boldsymbol{\Theta} &= \{p \in \mathbb{R} : 0 \leq p \leq 1\} \end{aligned}$$

Definition 3.2.2 (PMF).

$$p_X(x) = \mathbb{P}(X = x) = p^x \cdot (1-p)^{1-x} \cdot \mathbb{1}_{R_X}(x) \tag{3.2}$$

Definition 3.2.3 (PDF).

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{se } x < 0 \\ 1 - p & \text{se } 0 \leq x < 1 \\ 1 & \text{se } x \geq 1 \end{cases} \tag{3.3}$$

3.2.3 Moments

Proposition 3.2.1 (Momenti caratteristici).

$$\mathbb{E}[X] = p \tag{3.4}$$
$$\text{Var}[X] = p(1-p) \tag{3.5}$$
$$\text{Asym}(X) = \frac{1-2p}{\sqrt{p(1-p))}} \tag{3.6}$$
$$\text{Kurt}(X) = \frac{3p^2 - 3p + 1}{p(1-p)} \tag{3.7}$$

Proof. Per il valore atteso

$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1-p) = p$$

Per la varianza, dato che $X^2 = X$ e dunque $\mathbb{E}\left[X^2\right] = \mathbb{E}[X]$ si ha:

$$\text{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$$

□

Remark 96. In particolare il valore atteso coincide con la probabilità di successo e la varianza è sempre compresa nell'intervallo $[0; 0.25]$, raggiungendo il massimo per $p = 1/2$.

3.3 Indicator rv for an event

3.3.1 Definition, properties

Important remark 34. Any event A is associated to a Bernoulli indicator random variable.

Definition 3.3.1 (Indicator rv of event A)**.** Let $\Omega = \{\omega_1, \omega_2, \ldots\}$ be the sample space of the experiment considered and $A \subseteq \Omega$ a possible event; suppose that ω is the outcome that currently happens as a result of the experiment. Then:

$$I_A = I(A) = \begin{cases} 1 & \text{if } A \text{ verifies: } \omega \in A \\ 0 & \text{if } A \text{ does not: } \omega \notin A \end{cases}$$

therefore if $\mathbb{P}(A) = p$, then $I_A \sim \mathrm{Bern}(p)$

Proposition 3.3.1 (Indicator rv properties)**.**

$$(I_A)^n = I_A, \quad \forall n \in \mathbb{N} : n > 0 \tag{3.8}$$

$$I_{\overline{A}} = 1 - I_A \tag{3.9}$$

$$I_{A \cap B} = I_A \cdot I_B \tag{3.10}$$

$$I_{A \cup B} = I_A + I_B - I_A \cdot I_B \tag{3.11}$$

Proof. La 3.8 vale dato che $0^n = 0$ e $1^n = 1$ per qualsiasi intero positivo n. La 3.9 vale dato che $1 - I_A$ è 1 se A non accade e 0 se accade. Per la 3.10, $I_A \cdot I_B$ è 1 solo se sia I_A che I_B sono 1 e 0 altrimenti. Per la 3.11,

$$\begin{aligned} I_{A \cup B} &\overset{(1)}{=} 1 - I_{\overline{A} \cap \overline{B}} = 1 - I_{\overline{A}} \cdot I_{\overline{B}} = 1 - (1 - I_A)(1 - I_B) \\ &= I_A + I_B - I_A I_B \end{aligned}$$

dove in (1) abbiamo sfruttato De Morgan. □

3.3.2 Probability/expected value link

Remark 97. Indicator function/rv provide a link between probability of an event and expected value

Proposition 3.3.2 (Fundamental bridge)**.** *There's a 1-1 link between events and indicator rv: probability of an event A and the expected value of its indicator rv I_A:*

$$\mathbb{P}(A) = \mathbb{E}[I_A] \tag{3.12}$$

Proof. For any event A we have a rv I_A, and viceversa for each I_A there's one event A (that is $A = \{\omega \in \Omega : I_A(\omega) = 1\}$).
Considered $I_A \sim \mathrm{Bern}(p)$ with $p = \mathbb{P}(A)$, we have

$$\mathbb{E}[I_A] = \mathbb{E}[\mathrm{Bern}(p)] = 0 \cdot \mathbb{P}(I(A) = 0) + 1 \cdot \mathbb{P}(I(A) = 1)\, \mathbb{P}(A) = p$$

□

Remark 98 (Usefulness). Previous result enable to express any probability as expected value; some examples come in the following section.
Furthermore indicator rvs are useful in exercises on expected value: often we can define a complex rv of unknown/complex distribution function as sum of indicator function (simpler). The so-called fundamental bridge enable then, applying expected value properties, to find expected value of unknown complex distribution function

3.3.3 Some application: probability

Proposition 3.3.3 (Boole inequality). *If $E_1, \ldots, E_n$ are events we have:*

$$\mathbb{P}(E_1 \cup \ldots \cup E_n) \leq \mathbb{P}(E_1) + \ldots + \mathbb{P}(E_n) \tag{3.13}$$

Proof. Let $E_1, \ldots, E_n$ be the events considered; we note that

$$I_{E_1 \cup \ldots \cup E_n} \leq I_{E_1} + \ldots + I_{E_n}$$

since left branch is 1 is all the events occur while right one is 1 even if only one does. Taking expected value:

$$\begin{aligned}
&\mathbb{E}[I_{E_1 \cup \ldots \cup E_n}] \leq \mathbb{E}[I_{E_1} + \ldots + I_{E_n}] && \text{by linearity of expectation} \ldots \\
&\mathbb{E}[I_{E_1 \cup \ldots \cup E_n}] \leq \mathbb{E}[I_{E_1}] + \ldots + \mathbb{E}[I_{E_n}] && \text{applying 3.12} \ldots \\
&\mathbb{P}(E_1 \cup \ldots \cup E_n) \leq \mathbb{P}(E_1) + \ldots + \mathbb{P}(E_n)
\end{aligned}$$

□

Proposition 3.3.4 (Bonferroni inequality). *If $E_1, \ldots, E_n$ are events:*

$$\mathbb{P}(E_1 \cap \ldots \cap E_n) \geq 1 - \sum_{i=1}^{n} \mathbb{P}\left(\overline{E_i}\right) \tag{3.14}$$

Proof. Similarty to the Boole inequality, applying DeMorgan

$$I_{E_1 \cap \ldots \cap E_n} = 1 - I_{\overline{E_1} \cup \ldots \cup \overline{E_n}}$$

Taking expected value:

$$\begin{aligned}
&\mathbb{E}[I_{E_1 \cap \ldots \cap E_n}] = \mathbb{E}\left[1 - I_{\overline{E_1} \cup \ldots \cup \overline{E_n}}\right] && \text{per linearità} \ldots \\
&\mathbb{E}[I_{E_1 \cap \ldots \cap E_n}] = 1 - \mathbb{E}\left[I_{\overline{E_1} \cup \ldots \cup \overline{E_n}}\right] && \text{passando alle probabilità} \ldots \\
&\mathbb{P}(E_1 \cap \ldots \cap E_n) = 1 - \mathbb{P}\left(\overline{E_1} \cup \ldots \cup \overline{E_n}\right)
\end{aligned}$$

Finally applying 3.13

$$\mathbb{P}(E_1 \cap \ldots \cap E_n) = 1 - \mathbb{P}\left(\overline{E_1} \cup \ldots \cup \overline{E_n}\right) \geq 1 - \mathbb{P}\left(\overline{E_1}\right) - \ldots - \mathbb{P}\left(\overline{E_n}\right)$$

□

Proposition 3.3.5 (Inclusion/exclusion principle). *In case of two events*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \tag{3.15}$$

In general:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n}(-1)^{r+1} \sum_{i_1<\ldots<i_r} \mathbb{P}\left(E_{i1} \cap E_{i2} \cap \ldots \cap E_{ir}\right) \tag{3.16}$$

$$= \sum_{i} \mathbb{P}\left(E_i\right) - \sum_{i<j} \mathbb{P}\left(E_i \cap E_j\right) + \sum_{i<j<k} \mathbb{P}\left(E_i \cap E_j \cap E_k\right) - \ldots + (-1)^{n+1}\,\mathbb{P}\left(E_1 \cap \ldots \cap E_n\right) \tag{3.17}$$

Proof. Given 3.15 we take expected value of both branch of 3.11. Considering 3.16, we can apply indicator rv properties

$$\begin{aligned} 1 - I_{E_1\cup\ldots\cup E_n} &= I_{\overline{E_1}\cap\ldots\cap\overline{E_n}} \\ &= I_{\overline{E_1}} \cdot \ldots \cdot I_{\overline{E_n}} \\ &= (1 - I_{E_1}) \cdot \ldots \cdot (1 - I_{E_n}) \\ &\overset{(1)}{=} 1 - \sum_i I_{E_i} + \sum_{i<j} I_{E_i} I_{E_j} - \ldots + (-1)^n I_{E_1} \cdot \ldots \cdot I_{E_n} \end{aligned}$$

where in (1):

- il 1 significa selezionare tutti gli 1 negli n fattori;
- il $\sum_i I_{E_i}$ si ottiene selezionando tutti gli 1 a meno di un fattore a turno che ha sempre il segno $-$ davanti;
- $\sum_{i<j} I_{E_i} I_{E_j}$ si ottiene selezionando tutti gli 1 ad eccezione di due fattori.

Prendendo i valori attesi di ambo i membri si ha

$$\begin{aligned} \mathbb{E}\left[1 - I_{E_1\cup\ldots\cup E_n}\right] &= \mathbb{E}\left[1 - \sum_i I_{E_i} + \sum_{i<j} I_{E_i} I_{E_j} - \ldots + (-1)^n I_{E_1} \cdot \ldots \cdot I_{E_n}\right] \\ 1 - \mathbb{E}\left[I_{E_1\cup\ldots\cup E_n}\right] &\overset{(1)}{=} 1 - \mathbb{E}\left[\sum_i I_{E_i} - \sum_{i<j} I_{E_i} I_{E_j} + \ldots + (-1)^{n+1} I_{E_1} \cdot \ldots \cdot I_{E_n}\right] \\ \mathbb{E}\left[I_{E_1\cup\ldots\cup E_n}\right] &= \mathbb{E}\left[\sum_i I_{E_i}\right] - \mathbb{E}\left[\sum_{i<j} I_{E_i} I_{E_j}\right] + \ldots + \mathbb{E}\left[(-1)^{n+1} I_{E_1} \cdot \ldots \cdot I_{E_n}\right] \\ \mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) &= \sum_i \mathbb{P}\left(E_i\right) - \sum_{i<j} \mathbb{P}\left(E_i \cap E_j\right) + \ldots + (-1)^{n+1}\,\mathbb{P}\left(E_1 \cap \ldots \cap E_n\right) \end{aligned}$$

dove in (1) abbiamo raccolto un meno al secondo membro entro parentesi. □

3.3.4 Applications: expected value evaluation

Example 3.3.1 (Matching carte)**.** Abbiamo un mazzo di n carte numerate da 1 a n ben mischiato. Una carta è un match se la sua posizione nell'ordine del mazzo matcha con il suo numero. Sia X il numero totale di match nel mazzo:

Fisso ...	con reinserimento	senza reinserimento
n trial	binomiale	ipergeometrica
n successi	binomiale negativa	ipergeometrica negativa

Table 3.1

qual è il valore atteso di X?
Se scriviamo $X = I_1 + \ldots + I_n$ con

$$I_i = \begin{cases} 1 & \text{se l'}i\text{-esima carta matcha col proprio numero} \\ 0 & \text{altrimenti} \end{cases}$$

Si ha che, non condizionando a nulla e pensando ad un singolo shuffle/match

$$\mathbb{E}[I_i] = \frac{1}{n}$$

pertanto per linearità

$$\mathbb{E}[X] = \mathbb{E}[I_1] + \ldots + \mathbb{E}[I_n] = n \cdot \frac{1}{n} = 1$$

Quindi il numero di match medi è 1, indipendentemente da n. Anche se I_i sono dipendenti in maniera complicata, la linearità del valore atteso vale sempre.

Example 3.3.2 (Valore atteso di Ipergeometrica Negativa)**.** Un'urna contiene w palline bianche e b palline nere che sono estratte senza reinserimento. Il numero di palline nere estratte prima di pescare la prima bianca ha una distribuzione Ipergeometrica negativa (in tab 3.1 una sintesi dei casi). Trovare il valore atteso.
Trovarlo dalla definizione della variabile è complicato, ma possiamo esprimere la variabile come somma di indicatrici. Etichettiamo le palline nere con $1, 2, \ldots b$ e sia I_i l'indicatrice che la pallina nera i è stata estratta prima di qualsiasi bianca. Si ha che

$$\mathbb{P}(I_i = 1) = \frac{1}{w+1}$$

dato nel listare l'ordine in cui la pallina nera i e le altre bianche son pescate (ignorando le altre) tutti gli ordine sono equiprobabili. Pertanto per linearità

$$\mathbb{E}\left[\sum_{i=1}^{b} I_i\right] = \sum_{i=1}^{b} \mathbb{E}[I_i] = \frac{b}{w+1}$$

La risposta ha n senso dato che aumenta con b, diminuisce con w ed è corretta nei casi estremi $b = 0$ (nessuna pallina nera sarà estratta) e $w = 0$ (tutte le palline nere saranno esaurite prima di pescare una non esistente bianca).

3.4 Binomial

3.4.1 Definition

Remark 99. Used to know the probability of having x success among $n \geq x$ independent Bernoulli trial with common probabily success p.

Definition 3.4.1 (vc binomiale)**.** Eseguiamo n prove bernoulliane indipendenti, aventi comune probabilità di successo p (es estrazioni di palline bianche/nere da una urna con reimmissione). Sia X la somma dei successi ottenuti: allora X si distribuisce come una vc binomiale di parametri n e p, e si scrive $X \sim \text{Bin}(n, p)$.

Remark 100. Se $n = 1$ la distribuzione Binomiale coincide con quella di Bernoulli, ossia $\text{Bin}(1, p) = \text{Bern}(p)$

Proposition 3.4.1. *La binomiale può essere generata sommando bernoulliane iid; se X_i, $i = 1, \ldots, n$ sono vc bernoulliane iid $X_i \sim \text{Bern}(p)$ allora la loro somma $X = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$*

Proof. Sia $X_i = 1$ se l'i-esimo trial ha successo o 0 in caso contrario. Se pensiamo di avere una persona per ciascun trial, chiediamo di alzare la mano se si ha successo e contiamo le mani alzate (che equivale a sommare X_i) otteniamo il numero totale di successi in n trial che è X. □

3.4.2 Functions

Remark 101 (Supporto e spazio parametrico)*.*

$$\begin{aligned} R_X &= \{0, 1, \ldots, \ n\} \\ \boldsymbol{\Theta} &= \{n \in \mathbb{N} \setminus \{0\}, \ p \in \mathbb{R} : 0 \leq p \leq 1\} \end{aligned}$$

Definition 3.4.2 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = \binom{n}{x} \cdot p^x (1-p)^{n-x} \cdot \mathbb{1}_{R_X}(x) \tag{3.18}$$

con: x è il numero di successi, n è il numero di esperimenti, p probabilità di successo in ogni esperimento.

Remark 102. Nella 3.18 la prima parte (il coefficiente binomiale) serve per quantificare il numero di casi in cui si verificano il numero di successi desiderati; questa viene moltiplicata per la seconda che costituisce la probabilità di un tale esito (determinato come probabilità di eventi indipendenti di successo/insuccesso).

Definition 3.4.3 (Funzione di ripartizione)**.**

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{k=0}^{x} \binom{n}{k} \cdot p^k (1-p)^{n-k}$$

Validità PMF. Si ha che

$$\sum_{x=0}^{n} p_X(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \overset{(1)}{=} (p + (1-p))^n = 1$$

dove in (1) si è sfruttata la proprietà del coefficiente binomiale:

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$$

□

3.4.3 Moments

Proposition 3.4.2 (Momenti caratteristici)**.**

$$\mathbb{E}\left[X\right] = np \tag{3.19}$$
$$\mathrm{Var}\left[X\right] = np(1-p) \tag{3.20}$$
$$\mathrm{Asym}\left(X\right) = \frac{1-2p}{\sqrt{np(1-p))}} \tag{3.21}$$
$$\mathrm{Kurt}\left(X\right) = 3 + \frac{1-6p+6p^2}{np(1-p)} \tag{3.22}$$

Proof. Per il valore atteso, sfruttando il fatto che $X \sim \mathrm{Bin}\left(n, p\right)$ sia descrivibile come la somma di n vc $X_i \sim \mathrm{Bern}\left(p\right)$, sfruttando la linearità del valore atteso, il risultato è la somma di n valori attesi uguali:

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = n\,\mathbb{E}\left[X_i\right] = np$$

Alternativamente potevamo sviluppare l'algebra:

$$\begin{aligned}\mathbb{E}\left[X\right] &= \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x (1-p)^{(n-x)} = \sum_{x=0}^{n} x \cdot \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)} \\ &= \sum_{x=0}^{n} x \cdot \frac{n(n-1)!}{x(x-1)!\left[(n-1)-(x-1)\right]!} pp^{x-1}(1-p)^{\left[(n-1)-(x-1)\right]}\end{aligned}$$

Ora dato che per $x = 0$ il termine entro sommatoria è nullo possiamo portare avanti di uno l'indice inferiore della stessa:

$$\mathbb{E}\left[X\right] = \sum_{x=1}^{n} x \cdot \frac{n(n-1)!}{x(x-1)!\left[(n-1)-(x-1)\right]!} pp^{x-1}(1-p)^{\left[(n-1)-(x-1)\right]}$$

ponendo $y = x - 1$ si giunge

$$\begin{aligned}\mathbb{E}\left[X\right] &= np \sum_{y=0}^{n-1} \underbrace{\frac{(n-1)!}{y!\left[(n-1)-y\right]!} p^y (1-p)^{\left[(n-1)-y\right]}}_{\mathrm{Bin}\left(n-1, p\right)} \\ &\overset{(1)}{=} np\end{aligned}$$

con (1) dato che la sommatoria è $= 1$. ☐

Proof. Sfruttando sempre il fatto che $X \sim \mathrm{Bin}\left(n, p\right)$ sia descrivibile come la somma di n vc iid $X_i \sim \mathrm{Bern}\left(p\right)$, con varianza comune $p(1-p)$, e applicando le proprietà della varianza:

$$\mathrm{Var}\left[X\right] = \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \overset{(1)}{=} \sum_{i=1}^{n} \mathrm{Var}\left[X_i\right] = n\,\mathrm{Var}\left[X_i\right] = n \cdot p(1-p) \tag{3.23}$$

where in (1) there's no covariance since they are independent. ☐

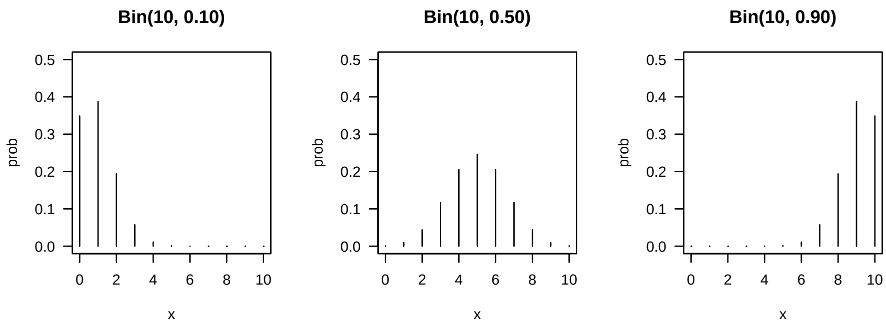


Figure 3.1: Forma distribuzione Bin (n, p)

3.4.4 Shape

```
plot_binom <- function(n, p, plot_main = TRUE, ...){
    the_seq <- seq(from = 0, to = n)
    probs <- dbinom(x = the_seq, size = n, p = p)
    plot(x = the_seq, y = probs, type = 'h', las = 1,
         xlab = 'x', ylab = 'prob',
         main = if (plot_main) sprintf('Bin(%d, %.2f)', n, p) else '',
         ...)
}

par(mfrow = c(1,3))
ylim <- c(0, 0.5)
plot_binom(n = 10, p = 0.1, ylim = ylim)
plot_binom(n = 10, p = 0.5, ylim = ylim)
plot_binom(n = 10, p = 0.9, ylim = ylim)
```

```
par(mfrow = c(2,2))
first_n <- 10
second_n <- 40
first_p <- 0.1
second_p <- 0.35
ylim <- c(0, 0.5)
plot_binom(n = first_n,  p = first_p,  ylim = ylim)
plot_binom(n = second_n, p = first_p,  ylim = ylim)
plot_binom(n = first_n,  p = second_p, ylim = ylim)
plot_binom(n = second_n, p = second_p, ylim = ylim)
```

Proposition 3.4.3 (Shape)**.** *La distribuzione è simmetrica se $p = 0.5$, è asimmetrica positiva (coda a destra) se $p < 0.5$, asimmetrica negativa (a sinistra) se $p > 0.5$. (Figura 3.1)*

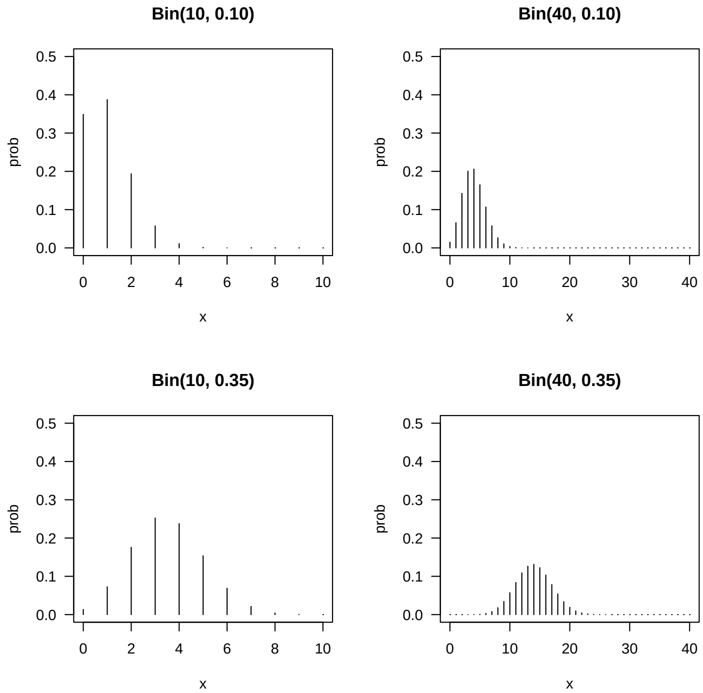


Figure 3.2: Convergenza alla normale della binomiale

Proof. Per $p = 0.5$ è simmetrica in quanto $p = 1 - p = \frac{1}{2}$ e

$$p_X(x) = \binom{n}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{n-x} = p_X(n-x) = \binom{n}{n-x}\left(\frac{1}{2}\right)^{n-x}\left(\frac{1}{2}\right)^x \quad (3.24)$$

per le proprietà del coefficiente binomiale. E dato che $p_X(x) = p_X(n-x)$, $\forall x \in R_X$, allora la distribuzione è simmetrica attorno al centro del supporto. $\square$

Proposition 3.4.4. *In una binomiale di parametri n, p, la funzione di densità (per x che varia da 0 a n) è inizialmente strettamente crescente e successivamente strettamente decrescente. Si raggiunge il massimo in corrispondenza del più grande intero $x \leq (n+1)p$*

Proof. Consideriamo il rapporto $\mathbb{P}(X = x)/\mathbb{P}(X = x-1)$ e determiniamo per quali valori di x esso risulti maggiore (funzione crescente) o minore (decrescente) di 1:

$$\frac{\mathbb{P}(X = x)}{\mathbb{P}(X = x-1)} = \frac{\frac{n!}{(n-x)!x!}p^x(1-p)^{n-x}}{\frac{n!}{(n-x+1)!(x-1)!}p^{x-1}(1-p)^{n-x+1}} = \frac{(n-x+1)p}{x(1-p)}$$

Quindi tale rapporto ≥ 1 se e solo se:

$$(n-x+1)p \geq x(1-p)$$
$$np - xp + p \geq x - xp$$

ossia $x \leq (n+1)p$ □

Remark 103 (Convergenza alla normale). La distribuzione converge verso la Normale (diviene simmetrica e la curtosi tende a 3) al crescere di $n \to \infty$; la convergenza è tanto più veloce per quanto più p è prossimo a 0.5. (figura 3.2)

3.4.5 Variabili derivate

Proposition 3.4.5 (Vc numero di insuccessi). *Sia* $X \sim \mathrm{Bin}(n,p)$. *Allora* $n - X \sim \mathrm{Bin}(n, 1-p)$.

Proof. Ad intuito basta invertire i ruoli di successo e insuccesso (si inverte anche la probabilità). Volendo tuttavia verificare, sia $Y = n - X$, la PMF è:

$$\begin{aligned}\mathbb{P}(Y = x) &\overset{(1)}{=} \mathbb{P}(X = n-x) = \binom{n}{n-x} p^{n-x}(1-p)^x \\ &\overset{(1)}{=} \binom{n}{x}(1-p)^x p^{n-x} = \mathrm{Bin}(n, 1-p)\end{aligned}$$

dove in (1) diciamo che in n estrazioni la probabilità di avere x fallimenti è uguale alla probabilità di avere $n-x$ successi , mentre in (2) abbiamo sfruttato la proprietà del coefficiente binomiale. □

Remark 104. Un fatto importante della binomiale è che la somma di binomiali indipendenti aventi la stessa probabilità di successo è un'altra binomiale

Proposition 3.4.6 (Somma di binomiali). *Se* $X \sim \mathrm{Bin}(n,p)$, $Y \sim \mathrm{Bin}(m,p)$ *e* X *è indipendente da* Y, *allora* $X + Y \sim \mathrm{Bin}(n+m, p)$

Proof. Un modo semplice è rappresentare X e Y come le somma di $X = X_1 + \ldots + X_n$ e $Y = Y_1 + \ldots + Y_n$ con $X_i, Y_i \sim \mathrm{Bern}(p)$ iid. Allora $X+Y$ è la somma di $n+m$ $\mathrm{Bern}(p)$ iid, pertanto la distribuzione è $\mathrm{Bin}(n+m,p)$ per teorema 3.4.1.
Alternativamente, mediante la legge delle probabilità totali, possiamo trovare la PMF di $X+Y$ condizionando su X (oppure ugualmente su Y) e sommando:

$$\begin{aligned}\mathbb{P}(X+Y=k) &= \sum_{j=0}^{k} \mathbb{P}(X+Y=k|X=j)\cdot\mathbb{P}(X=j) \\ &= \sum_{j=0}^{k} \mathbb{P}(Y=k-j|X=j)\cdot\mathbb{P}(X=j) \\ &\overset{(1)}{=} \sum_{j=0}^{k} \mathbb{P}(Y=k-j)\cdot\mathbb{P}(X=j) \\ &= \sum_{j=0}^{k} \binom{m}{k-j} p^{k-j}(1-p)^{m-k+j} \cdot \binom{n}{j} p^j (1-p)^{n-j} \\ &= p^k(1-p)^{n+m-k}\sum_{j=0}^{k}\binom{m}{k-j}\binom{n}{j} \\ &\overset{(2)}{=} \binom{n+m}{k} p^k (1-p)^{n+m-k} = \mathrm{Bin}(n+m,p)\end{aligned}$$

dove in (1) abbiamo sfruttato l'indipendenza tra X e Y e in (2) l'identità di Vandermonde (eq ??). $\square$

3.5 Hypergeometric

3.5.1 Definition

Remark 105. La variabile ipergeometrica descrive l'estrazione *senza reinserimento* di palline dicotomiche da un urna. A differenza della binomiale dove la probabilità di successo p non cambiava da una sottoprova Bernoulliana all'altra, qui il non reinserimento fa si che la probabilità di successo vari ad ogni prova.

Definition 3.5.1 (Distribuzione ipergeometrica)**.** Supponiamo di dover estrarre un campione di n palline senza reinserimento da un'urna che contiene w palline bianche (successo) e b nere. Il numero X di palline bianche (successi) tra le estratte si distribuisce come una ipergeometrica con parametri w, b ed n e si scrive $X \sim \mathrm{HGeom}\,(w, b, n)$.

3.5.2 Functions

Remark 106 (Supporto e spazio parametrico).

$$\begin{aligned} R_X &= \{0, 1, \dots, n\} \\ \mathbf{\Theta} &= \{w, b \in \mathbb{N} : w + b \geq 1;\ n \in \{0, \dots, w + b\}\} \end{aligned}$$

Definition 3.5.2 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}\,(X = x) = \frac{\dbinom{w}{x}\dbinom{b}{n-x}}{\dbinom{w+b}{n}} \cdot \mathbb{1}_{R_X}(x) \tag{3.25}$$

Remark 107 (Interpretazione). Al denominatore sono quantificati il numero di modi con cui posso estrarre n palline qualsiasi dall'urna. Di queste estrazioni, al numeratore sono quantificati il numero di modi in cui nelle n palline estratte ci sono x bianche (successi); ossia devo averne x bianche scelte tra w ($\binom{w}{x}$ modi per farlo), e $n - x$ nere scelte tra b ($\binom{b}{n-x}$ modi).

Remark 108 (Binomiale e ipergeometrica). Come è intuitivo, se il numero di palline estratte n è molto vicino al numero di palline totali nell'urna $w + b$ la binomiale e l'ipergeometrica forniscono risultati molto diversi.
Tuttavia se n è molto più piccolo di $w + b$ binomiale e ipergeometrica danno risultati simili: nella pratica le estrazioni sono quasi sempre senza reimmissione ma ciò nonostante si usa la binomiale al posto dell'ipergeometrica perché l'errore che si commette è trascurabile.

Validità PMF. Facendo la somma del numeratore si ha:

$$\sum_{x=0}^{n} \binom{w}{x}\binom{b}{n-x} \overset{(1)}{=} \binom{w+b}{n}$$

con (1) per l'identità di Vandermonde (eq ??), per cui la PMF somma a 1. $\square$

Remark 109. In R per la PMF si usa `dhyper(x, m, n, k)` dove `x` è il supporto (ossia il numero di palline bianche estratte), `m` il numero di palline bianche nell'urna, `n` il numero di palline nere e k il numero di estrazioni.

Remark 110. L'argomento che abbiamo usato per l'ipergeometrica si adatta immediatamente al seguente caso più generale (equivalente estensione dalla binomiale alla multinomiale per la ipergeometrica)

Definition 3.5.3 (Multivariate Hypergeometric distribution)**.** Supponiamo di avere una urna contenente palline di k colori, in numerosità $n_1, \ldots, n_k$ e di effettuare $n \leq \sum_i n_i$ estrazioni senza reimmissione. In tal caso la probabilità di estrarre j_1 palline del primo colore, $\ldots j_k$ palline del k-esimo è

$$\mathbb{P}\left(j_1, \ldots, j_k\right) = \frac{\binom{n_1}{j_1} \cdot \ldots \cdot \binom{n_1}{j_1}}{\binom{\sum_i n_i}{\sum_i j_i}}$$

3.5.3 Moments

Proposition 3.5.1 (Momenti caratteristici)**.**

$$\mathbb{E}\left[X\right] = n\frac{w}{w+b} \tag{3.26}$$

$$\operatorname{Var}\left[X\right] = np(1-p)\left(\frac{w+b-n}{w+b-1}\right), \qquad con\ p = \frac{w}{w+b} \tag{3.27}$$

Proof. Per il valore atteso, come nel caso binomiale possiamo scrivere X come somma di Bernoulliane $I_i \sim \operatorname{Bern}(p)$ con $p = w/(w+b)$.

$$X = I_1 + \ldots + I_n$$

A differenza della binomiale le I_i non sono indipendenti, tuttavia la linearità del valore atteso non lo richiede, quindi

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[I_1 + \ldots + I_n\right] = \mathbb{E}\left[I_1\right] + \ldots + \mathbb{E}\left[I_n\right] = np = n\frac{w}{w+b}$$

□

Proof. Per la varianza invece essendo variabili non indipendenti non possiamo sommare le varianze direttamente. Vedremo in seguito la dimostrazione della formula riportata. □

3.5.4 Struttura essenziale ed esperimenti assimilabili

Remark 111. L'idea dell'Ipergeometrica è classificare una popolazione utilizzando due set di tag consecutivi (entrambi dicotomici successo/insuccesso) e ottenere il numero degli elementi caratterizzati dal successo in entrambi i tag. Nell'esempio delle palline il primo tag è il colore della pallina (bianco = successo), mentre il secondo è estrazione (estratta = sucesso).
Problemi aventi la stessa struttura presenteranno medesima distribuzione.

Example 3.5.1. Il numero A di assi estratti (sono 4 in un mazzo di 52 carte) in una mano di poker (5 carte estratte) si distribuirà come $A \sim \operatorname{HGeom}(4, 48, 5)$.

Remark 112. La struttura essenziale ci permette di dimostrare facilmente l'uguaglianza di due ipergeometriche dove l'ordine dei set di tag viene invertito

Proposition 3.5.2. HGeom (w, b, n) *e* HGeom $(n, w+b-n, w)$ *sono identiche.*

Proof. Sia $X \sim$ HGeom (w, b, n) è il numero di palline bianche tra le estratte campione; sia $Y \sim$ HGeom $(n, w+b-n, w)$ il numero di palline estratte tra le bianche (pensando ad estratto/non estratto come il primo tag e al colore come secondo. Entrambe X, Y contano il numero di bianche estratte pertanto avranno la stessa distribuzione.
Alternativamente possiamo controllare algebricamente che

$$\mathbb{P}(X = x) = \frac{\binom{w}{x}\binom{b}{n-x}}{\binom{w+b}{n}} = \frac{\frac{w!}{k!(w-k)!}\frac{b!}{(n-k)!(b-n+k)!}}{\frac{(w+b)!}{n!(w+b-n)!}} = \frac{w!b!n!(w+b-n)!}{k!(w-k)!(n-k)!(b-n+k)!}$$

$$\mathbb{P}(Y = y) = \frac{\binom{n}{y}\binom{w+b-n}{w-y}}{\binom{w+b}{w}} = \frac{\frac{n!}{k!(n-k)!}\frac{(w+b-n)!}{(w-k)!(b-n+k)!}}{\frac{(w+b)!}{w!b!}} = \frac{w!b!n!(w+b-n)!}{k!(w-k)!(n-k)!(b-n+k)!}$$

e dunque $\mathbb{P}(X = x) = \mathbb{P}(Y = y)$. □

3.5.5 Connessioni con la binomiale

Remark 113. Binomiale ed ipergeometrica sono connesse: possiamo ottenere la binomiale calcolando un limite sull'ipergeometrica, oppure ottenere una ipergeometrica condizionando una binomiale.

3.5.5.1 Dall'ipergeometrica alla binomiale

Proposition 3.5.3. *Se* $X \sim$ HGeom (w, b, n) *e* $w + b \to \infty$ *ma* $p = w/(w+b)$ *rimane fisso, allora la PMF di* X *converge a* Bin (n, p).

Proof. Sviluppiamo algebricamente per essere comodi prima di applicare il limite:

$$\mathbb{P}(X = x) = \frac{\binom{w}{x}\binom{b}{n-x}}{\binom{w+b}{n}} \overset{(1)}{=} \binom{n}{x}\frac{\binom{w+b-n}{w-x}}{\binom{w+b}{w}}$$

dove in (1) abbiamo sfruttato che HGeom $(w, b, n) =$ HGeom $(n, w+b-n, w)$ come nella dimostrazione di 3.5.2. Ora sviluppiamo il rapporto al secondo fattore ricordando che $\binom{n}{d} = \frac{n!}{d!(n-d)!}$; si ha:

$$\begin{aligned}
\frac{\binom{w+b-n}{w-x}}{\binom{w+b}{w}} &= \frac{(w+b-n)!}{(w-x)!(w+b-n-w+x)!} : \frac{(w+b)!}{w!(w+b-w)!} \\
&= \frac{(w+b-n)!}{(w-x)!(b-n+x)!} \cdot \frac{w!b!}{(w+b)!} \\
&= \frac{w!}{(w-x)!}\frac{b!}{(b-n+x)!}\frac{(w+b-n)!}{(w+b)!} \\
&= \frac{w \cdot \ldots \cdot (w-x+1)(w-x)!}{(w-x)!}\frac{b \cdot \ldots \cdot (b-n+x+1)(b-n+x)!}{(b-n+x)!}\frac{(w+b-n)!}{(w+b) \cdot \ldots \cdot (w+b-n+1}\\
&= \frac{w \cdot \ldots \cdot (w-x+1)}{1}\frac{b \cdot \ldots \cdot (b-n+x+1)}{1}\frac{1}{(w+b) \cdot \ldots \cdot (w+b-n+1)}
\end{aligned}$$

ora al numeratore del primo rapporto abbiamo $w-(w-x+1)+1=x$ fattori, al numeratore del secondo ne abbiamo $b-(b-n+x+1)+1=n-x$ elementi. Pertanto complessivamente al numeratore abbiamo n fattori. Al denominatore invece abbiamo $(w+b)-(w+b-n+1)+1=n$ fattori anche qui. Pertanto possiamo dividere per $(w+b)$, applicandolo n volte sia al numeratore che al denominatore, ottenendo

$$\frac{\binom{w+b-n}{w-x}}{\binom{w+b}{w}}=\frac{\frac{w}{w+b}\cdot\ldots\cdot\left(\frac{w}{w+b}-\frac{x-1}{w+b}\right)\cdot\left(\frac{b}{w+b}\right)\cdot\ldots\cdot\left(\frac{b}{w+b}-\frac{n-x-1}{w+b}\right)}{1\cdot\ldots\cdot\left(1-\frac{n-1}{w+b}\right)}$$

ora sostituendo $p=\frac{w}{w+b}$, $1-p=\frac{b}{w+b}$ e al denominatore $w+b=N$ dove utile si ha:

$$\frac{\binom{w+b-n}{w-x}}{\binom{w+b}{w}}=\frac{p\cdot\ldots\cdot\left(p-\frac{x-1}{N}\right)\cdot(1-p)\cdot\ldots\cdot\left(1-p-\frac{n-x-1}{N}\right)}{\left(1-\frac{1}{N}\right)\ldots\left(1-\frac{n-1}{N}\right)}$$

Ora tornando da dove siamo partiti abbiamo:

$$\mathbb{P}(X=x)=\binom{n}{x}\frac{p\cdot\ldots\cdot\left(p-\frac{x-1}{N}\right)\cdot(1-p)\cdot\ldots\cdot\left(1-p-\frac{n-x-1}{N}\right)}{\left(1-\frac{1}{N}\right)\ldots\left(1-\frac{n-1}{N}\right)}$$

Infine per $N\to+\infty$ il denominatore va a 1 mentre il numeratore va a $p^x(1-p)^{n-x}$ pertanto

$$\mathbb{P}(X=x)\to\binom{n}{x}p^x(1-p)^{n-x}$$

che è la $\mathrm{Bin}(n,p)$.
Intuitivamente data un'urna con w palline bianche e b nere, la binomiale sorge dall'estrarre n palline con replacement, mentre l'ipergeometrica senza. Se il numero di palline nell'urna sale notevolmente rispetto al numero di palline estratte, il campionamento con ripetizione e senza diventano essenzialmente equivalenti. (l'estrazione di una pallina non cambia la probabilità delle prossime estrazioni perché data la grande numerosità nell'urna non modifica praticamente la probabilità di successo) □

Remark 114. In termini pratici il teorema ci dice che se $N=w+b$ è grande rispetto a n possiamo approssimare la PMF di $\mathrm{HGeom}(w,b,n)$ con $\mathrm{Bin}(n,w/(w+b))$.

3.5.5.2 Dalla binomiale all'ipergeometrica

Proposition 3.5.4. *Se $X\sim\mathrm{Bin}(n,p)$, $Y\sim\mathrm{Bin}(m,p)$ e X è indipendente da Y, allora la distribuzione condizionata di X dato che $X+Y=r$ è $\mathrm{HGeom}(n,m,r)$*

Remark 115. Dimostriamo attraverso un esempio (distribuzione del test esatto di Fisher).

Proof. Un ricercatore vuole studiare se la prevalenza di una data malattia sia uguale o meno tra maschi e femmine. Raccoglie un campione di n donne ed m uomini e testa la malattia. Sia $X\sim\mathrm{Bin}(n,p_1)$ il numero di donne con la malattia nel campione e $Y\sim\mathrm{Bin}(n,p_2)$ il numero di uomini. Qui p_1 e p_2 sono sconosciuti.

	Donne	Uomini	Tot
Malato	x	$r-x$	r
Sano	$n-x$	$m-r+x$	$n+m-r$
Tot	n	m	$n+m$

Table 3.2

Supponiamo che siano osservate $X+Y=r$ persone malate. Siamo interessati a testare se $p_1 = p_2 = p$ (la cd ipotesi nulla); il test di Fisher si fonda sul condizionare sui totali di riga e colonna (quindi n, m, r sono considerati fissi) e verificare se il valore osservato X (numero di donne malate) sia estremo (dato che il tot malati è r) sotto ipotesi nulla. Assumendo l'ipotesi nulla vera troviamo la PMF condizionale di X dato che $X+Y=r$.
La tabella 2×2 di riferimento è la 3.2. Costruiamo PMF condizionata attraverso la regola di Bayes:

$$\begin{aligned}\mathbb{P}\left(X=x|X+Y=r\right) &= \frac{\mathbb{P}\left(X+Y=r|X=x\right)\mathbb{P}\left(X=x\right)}{\mathbb{P}\left(X+Y=r\right)} = \frac{\mathbb{P}\left(Y=r-x|X=x\right)\mathbb{P}\left(X=x\right)}{\mathbb{P}\left(X+Y=r\right)} \\ &\stackrel{(1)}{=} \frac{\mathbb{P}\left(Y=r-x\right)\mathbb{P}\left(X=x\right)}{\mathbb{P}\left(X+Y=r\right)}\end{aligned}$$

dove in (1) abbiamo sfruttato l'indipendenza di X e Y. Assumendo per buona l'ipotesi nulla e impostando $p_1 = p_2 = p$ si hanno le vc indipendenti $X \sim \text{Bin}\,(n,p)$ e $Y \sim \text{Bin}\,(m,p)$, per cui $X+Y \sim \text{Bin}\,(n+m,p)$ (per il risultato 3.4.6). Pertanto sostituendo le formule per esteso si ha

$$\begin{aligned}\mathbb{P}\left(X=x|X+Y=r\right) &= \frac{\binom{m}{r-x}p^{r-x}(1-p)^{m-r+x}\cdot\binom{n}{x}p^x(1-p)^{n-x}}{\binom{n+m}{r}p^r(1-p)^{n+m-r}} \\ &= \frac{\binom{n}{x}\binom{m}{r-x}}{\binom{n+m}{r}} = \text{HGeom}\,(n,m,r)\end{aligned}$$

Intuitivamente questo avviene perché condizionatamente ad avere $X+Y=r$ malati (primo tag), X è il numero di donne (secondo tag) tra quelli. □

3.6 Geometric

3.6.1 Definition

Remark 116. Supponiamo di ripetere in maniera indipendente diverse prove bernoulliane, ciascuna avente p probabilità di successo, sino a che che si verifica il primo successo. Sia X il numero di prove *fallimentari* necessari per ottenere il primo successo; X si distribuisce come una variabile geometrica con parametro p e si scrive $X \sim \text{Geom}\,(p)$.

Example 3.6.1. Il numero di croci sino alla prima testa si distribuisce come $\text{Geom}\,(1/2)$.

3.6.2 Functions

Remark 117 (Supporto e spazio parametrico).

$$R_X = \{x \in \mathbb{N}\}$$
$$\boldsymbol{\Theta} = \{p \in (0,1)\}$$

Definition 3.6.1 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}\left(X = x\right) = (1-p)^x p \cdot \mathbb{1}_{R_X}(x) \tag{3.28}$$

Validità PMF. Si ha che

$$\sum_{x=0}^{\infty}(1-p)^x p = p\sum_{x=0}^{\infty}(1-p)^x \overset{(1)}{=} p \cdot \frac{1}{p} = 1$$

con l'uguaglianza (1) dovuta alla serie geometrica. □

Remark 118. Come il teorema binomiale mostra che la PMF binomiale sia valida, la serie geometrica mostra che la PMF Geometrica sia valida.

Remark 119 (Interpretazione). La probabilità di avere x fallimenti consecutivi seguiti da un successo è data dalla probabilità di x fallimenti per la probabilità di un successo.

Definition 3.6.2 (Funzione di ripartizione)**.** Si ha

$$F_X(x) = \mathbb{P}\left(X \leq x\right) = 1 - (1-p)^{x+1} \tag{3.29}$$

Derivazione della CDF. Si ha

$$F_X(x) = \mathbb{P}\left(X \leq x\right) = 1 - \mathbb{P}\left(X > x\right) = 1 - \sum_{k=x+1}^{\infty}(1-p)^k p$$

Espandendo la sommatoria:

$$\begin{aligned}
\sum_{k=x+1}^{\infty}(1-p)^k p &= (1-p)^{x+1}\cdot p + (1-p)^{x+2}\cdot p + \ldots + (1-p)^{\infty}\cdot p \\
&= p(1-p)^x\left[(1-p) + (1-p)^2 + \ldots + (1-p)^{\infty}\right] \\
&= p(1-p)^x\left[\sum_{i=1}^{\infty}(1-p)^i\right] \\
&= p(1-p)^x\left[\sum_{i=0}^{\infty}(1-p)^i - 1\right] \\
&= p(1-p)^x\left(\frac{1}{p} - 1\right) = p(1-p)^x\frac{1-p}{p} \\
&= (1-p)^{x+1}
\end{aligned}$$

Pertanto:

$$F_X(x) = 1 - (1-p)^{x+1}$$

□

3.6.3 Moments

Proposition 3.6.1 (Momenti caratteristici).

$$\mathbb{E}[X] = \frac{1-p}{p}$$
$$\text{Var}[X] = \frac{1-p}{p^2}$$

Proof. Per il valore atteso abbiamo

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \cdot (1-p)^x p$$

Non può essere ricondotta a serie geometrica direttamente per la presenza entro sommatoria di x come primo fattore. Ma notiamo che il termine entro sommatoria assomiglia a $x(1-p)^{x-1}$ ossia la derivata di $(1-p)^x$ rispetto a $1-p$, quindi partiamo da li:

$$\sum_{x=0}^{\infty}(1-p)^x = \frac{1}{p}$$

Questa serie converge dato che $0 < p < 1$. Derivando entrambi i membri rispetto a p.

$$\sum_{x=0}^{\infty} x(1-p)^{x-1} \cdot (-1) = -\frac{1}{p^2}$$
$$\sum_{x=0}^{\infty} x(1-p)^{x-1} = \frac{1}{p^2}$$

e se moltiplichiamo entrambi i lati per $p(1-p)$ otteniamo la somma dalla quale siamo partiti

$$p(1-p)\sum_{x=0}^{\infty} x(1-p)^{x-1} = \frac{1}{p^2}p(1-p)$$
$$\sum_{x=0}^{\infty} xp(1-p)^x = \frac{1-p}{p}$$

□

Proof. Per la varianza dobbiamo calcolare $\mathbb{E}\left[X^2\right]$:

$$\mathbb{E}\left[X^2\right] = \sum_{x=0}^{\infty} x^2 \cdot \mathbb{P}(X=x) = \sum_{x=0}^{\infty} x^2 \cdot (1-p)^x \cdot p \overset{(1)}{=} \sum_{x=1}^{\infty} x^2 \cdot (1-p)^x \cdot p$$

con (1) dato dal fatto che se $x = 0$ il termine entro sommatoria è nullo e si può portare avanti l'indice della stessa. Anche qui cerchiamo di sfruttare la serie geometrica per arrivare ad una espressione compatta equivalente all'ultimo termine di sopra. La serie è

$$\sum_{x=0}^{\infty}(1-p)^x = \frac{1}{p}$$

Derivando rispetto a p entrambi i membri, come visto in precedenza si ha:

$$\sum_{x=0}^{\infty} x \cdot (1-p)^{x-1} = \frac{1}{p^2}$$

Possiamo portare avanti di 1 l'indice di sommatoria dato che se $x = 0$ è nullo il termine dentro

$$\sum_{x=1}^{\infty} x \cdot (1-p)^{x-1} = \frac{1}{p^2}$$

Ora, derivando ancora si andrebbe a $x(x-1)$ entro sommatoria, invece di x^2 desiderato, pertanto moltiplichiamo per $(1-p)$ entrambi i membri giungendo a:

$$\sum_{x=1}^{\infty} x \cdot (1-p)^{x} = \frac{1-p}{p^2}$$

Derivando ambo i membri nuovamente rispetto a p si va a

$$\begin{aligned}
\sum_{x=1}^{\infty} x^2 \cdot (1-p)^{x-1} \cdot (-1) &= \frac{(-1) \cdot p^2 - 2p \cdot (1-p)}{p^4} \\
\sum_{x=1}^{\infty} x^2 \cdot (1-p)^{x-1} &= (-1) \frac{p^2 - 2p}{p^4} \\
\sum_{x=1}^{\infty} x^2 \cdot (1-p)^{x-1} &= \frac{2-p}{p^3}
\end{aligned}$$

Moltiplicando entrambi i membri per $(1-p) \cdot p$ si arriva al punto dove eravamo rimasti con $\mathbb{E}\left[X^2\right]$

$$\sum_{x=1}^{\infty} x^2 \cdot (1-p)^{x} \cdot p = \frac{2-p}{p^3} \cdot (1-p) \cdot p = \frac{(2-p)(1-p)}{p^2}$$

Per cui

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x^2 \cdot (1-p)^{x} \cdot p = \frac{(2-p)(1-p)}{p^2}$$

e dunque:

$$\begin{aligned}
\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 &= \frac{(2-p)(1-p)}{p^2} - \frac{(1-p)^2}{p^2} \\
&= \frac{(1-p)(2-p-1+p))}{p^2} = \frac{1-p}{p^2}
\end{aligned}$$

□

3.6.4 Shape

Remark 120 (Shape). Tutte le geometriche hanno forma simile: la funzione è decrescente, con probabilità più alte associate ai valori più piccoli di x. Ha asimmetria positiva che aumenta al crescere di p (più p è alto più velocemente la PMF discende verso 0). Ha una notevole curtosi (figura 3.3)

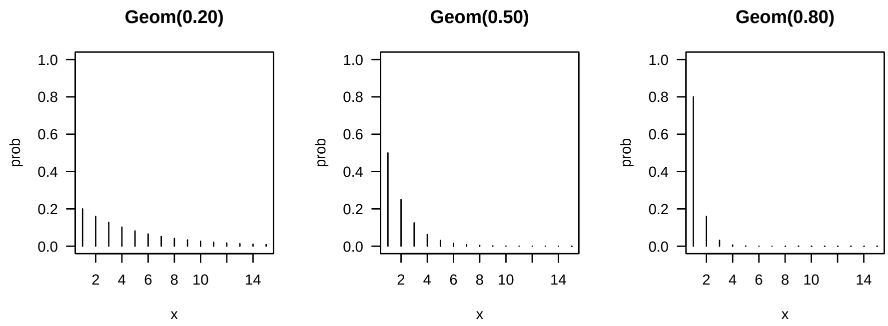


Figure 3.3: Forma distribuzione Geom (p)

```r
## occhio alla parametrizzazione di R per cui p(x) = p (1-p)^x
plot_geom <- function(p, plot_main = TRUE, ...){
    the_seq <- seq(from = 0, length = 15)
    probs <- dgeom(x = the_seq, prob = p)
    plot(x = the_seq + 1, y = probs, type = 'h', las = 1,
         xlab = 'x', ylab = 'prob',
         main = if (plot_main) sprintf('Geom(%.2f)', p) else '',
         ...)
}

all_p <- c(0.2, 0.5, 0.8)
par(mfrow = c(1, 3))
rm <- lapply(all_p, function(p) plot_geom(p = p, ylim = c(0, 1)))
```

3.6.5 Assenza di memoria

Remark 121. Una proprietà peculiare della geometrica è di esser l'unica vc discreta senza memoria (a parte la sua riformulazione).

Proposition 3.6.2 (Assenza di memoria)**.**

$$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s) \tag{3.30}$$

Proof. Si ha:

$$\begin{aligned}\mathbb{P}(X > t + s | X > t) &= \frac{\mathbb{P}(X > t + s)}{\mathbb{P}(X > t)} = \frac{1 - F_X(t+s)}{1 - F_X(t)} = \frac{1 - 1 + (1-p)^{t+s+1}}{1 - 1 + (1-p)^{t+1}} \\ &= (1-p)^s = 1 - F_X(s) = \mathbb{P}(X > s)\end{aligned}$$

□

3.6.6 Alternative definition (first success distribution)

Remark 122. Altri definiscono X come il numero di *prove* necessarie per ottenere il primo successo (incluso quest'ultimo). Qui la chiamiamo FS distribution e la indichiamo con $X \sim \text{FS}(p)$

Remark 123. Se $Y \sim \text{FS}(p)$ allora $Y - 1 \sim \text{Geom}(p)$ e possiamo convertire tra le PMF di Y e $Y - 1$ scrivendo

$$\mathbb{P}(Y = k) = \mathbb{P}(Y - 1 = k - 1)$$

Viceversa se $X \sim \text{Geom}(p)$ allora $X + 1 \sim \text{FS}(p)$

Remark 124 (Supporto e spazio parametrico).

$$\begin{aligned} R_X &= \{x \in \mathbb{N} \setminus \{0\}\} \\ \boldsymbol{\Theta} &= \{p \in (0, 1)\} \end{aligned}$$

Definition 3.6.3 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = (1 - p)^{x-1} p \cdot \mathbb{1}_{R_X}(x) \tag{3.31}$$

Remark 125 (Interpretazione). La probabilità di avere il primo successo all'n-esima estrazione e data dalla probabilità di $n - 1$ fallimenti per la probabilità di un successo.

Definition 3.6.4 (Funzione di ripartizione)**.**

$$\begin{aligned} F_X(x) = \mathbb{P}(X \le x) &= \sum_{k=1}^{x} \mathbb{P}(X = k) = \sum_{k=1}^{x} (1 - p)^{k-1} p \\ &= 1 - (1 - p)^x \end{aligned} \tag{3.32}$$

Proposition 3.6.3 (Momenti caratteristici)**.**

$$\begin{aligned} \mathbb{E}[X] &= \frac{1}{p} \\ \text{Var}[X] &= \frac{1-p}{p^2} \\ \text{Asym}(X) &= \frac{2-p}{\sqrt{1-p}} \\ \text{Kurt}(X) &= 9 + \frac{p^2}{1-p} \end{aligned}$$

Proof. Sia $Y = X + 1 \sim \text{FS}(p)$ con $X \sim \text{Geom}(p)$. Allora sfruttando le conoscenze sulla geometrica e le proprietà di valore atteso e varianza

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[X + 1] = \mathbb{E}[X] + 1 = \frac{1-p}{p} + 1 = \frac{1}{p} \\ \text{Var}[Y] &= \text{Var}[X + 1] = \text{Var}[X] = \frac{1-p}{p^2} \end{aligned}$$

□

Proposition 3.6.4 (Assenza di memoria)**.** *Analogamente a quanto avviene per la geometrica* $\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s)$*.*

Proof. Si ha:

$$\begin{aligned}
\mathbb{P}(X > t+s|X > t) &= \frac{\mathbb{P}(X > t+s)}{\mathbb{P}(X > t)} = \frac{1-F_X(t+s)}{1-F_X(t)} \\
&= \frac{(1-p)^{t+s}}{(1-p)^t} = (1-p)^s \\
&= \mathbb{P}(X > s)
\end{aligned}$$

ovvero il ritardo accertato di un evento in t sottoprove indipendenti non modifica la probabilità che esso si verifichi entro ulteriori s sottoprove. □

3.7 Negative binomial

Remark 126. Generalizza la distribuzione Geometrica: invece di aspettare il primo successo conta i fallimenti prima di ottenere il k-esimo successo.

Remark 127. In altre parole l'esperimento è lo stesso della binomiale. Però mentre nel caso della binomiale io fissavo il numero delle prove e contavo il numero di successi qui faccio il contrario. Vi sono due definizioni equivalenti:

- fisso il numero di successi e conto il numero delle *prove* che occorre fare per ottenere quel certo numero di successi
- fisso il numero di successi e conto il numero di *insuccessi* che occorre fare per ottenere quel certo numero di successi

3.7.1 Definition

Definition 3.7.1 (Definizione con numero di fallimenti)**.** In una sequenza di prove Bernoulliane indipendenti con probabilità di successo p, se X è il numero di *fallimenti* prima del k-esimo successo, allora X ha una distribuzione binomiale negativa con parametri k e p e si scrive $X \sim \text{Nb}(k, p)$

Remark 128. Anche a livello di notazione, nei parametri, si nota subito la differenza con la binomiale: questa fissa il numero di trial mentre la binomiale negativa fissa il numero di successi.

3.7.2 Functions

Remark 129 (Supporto e spazio parametrico).

$$\begin{aligned}
R_X &= \mathbb{N} \\
\boldsymbol{\Theta} &= \{k \in \mathbb{N} : k \geq 1, p \in \mathbb{R} : 0 \leq p \leq 1\}
\end{aligned}$$

Definition 3.7.2 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = \binom{x+k-1}{k-1} p^k (1-p)^x \cdot \mathbb{1}_{R_X}(x) \tag{3.33}$$

Remark 130 (Interpretazione). Ci sono $\binom{x+k-1}{k-1}$ sequenze possibili di x fallimenti e $k-1$ successi. Ciascuna di esse ha probabilità $p^{k-1}(1-p)^x$. Si termina con un success, quindi moltiplicando per p.

Remark 131. Come una binomiale può essere rappresentata da una somma di Bernoulli iid, una binomiale negativa può essere rappresentata come somma di Geometriche iid, come mostrato dal seguente teorema.

Proposition 3.7.1. *Sia $X \sim \mathrm{Nb}(k, p)$ il numero di fallimenti prima del k-esimo successo in una sequenza di probe bernoulliane indipendenti con probabilità di successo p. Allora possiamo scrivere $X = X_1 + \ldots + X_k$ dove gli X_i sono iid e $X_i \sim \mathrm{Geom}(p)$.*

Proof. Sia X_1 il numero di fallimenti prima del primo successo, X_2 il numero di fallimenti tra il primo successo e il secondo e, in generale, X_i il numero di fallimenti tra $(i-1)$-esimo succeso e l'i-esimo.
Allora $X_1 \sim \mathrm{Geom}(p)$ per la definizione della geometrica, $X_2 \sim \mathrm{Geom}(p)$ e così via. Inoltre le X_i sono indipendenti dato che le prove bernoulliane sono indipendenti l'un l'altra. Sommando gli X_i si ottiene il totale di fallimenti prima del k-esimo successo, che è X. □

3.7.3 Moments

Proposition 3.7.2 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = k\frac{1-p}{p} \tag{3.34}$$

$$\mathrm{Var}[X] = k\frac{1-p}{p^2} \tag{3.35}$$

Proof. Per il valore atteso sfruttiamo che X è scrivibile come somma di k vc Geometriche X_i. Il valore atteso è la somma dei valori attesi delle geometriche:

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \ldots + X_k] = \mathbb{E}[X_1] + \ldots \mathbb{E}[X_k] = k\frac{1-p}{p}$$

Per la varianza avviene lo stesso, dato che le variabili sono indipendenti:

$$\mathrm{Var}[X] = \mathrm{Var}[X_1 + \ldots + X_k] = \mathrm{Var}[X_1] + \ldots \mathrm{Var}[X_k] = k\frac{1-p}{p^2}$$

□

3.7.4 Shape

Remark 132 (Shape)*.* Si nota che così al crescere di k, la distribuzione diviene più simmetrica e la curtosi tende a 3 indicando convergenza alla normalità. All'aumentare di p assume asimmetria positiva. (figura 3.4)

```
## The probability of obtaining the fourth cross before the
## third head (and then after two head) is equal to 11.72%.

plot_binom_neg <- function(k, p, plot_main = TRUE, ...){
    fails <- seq(from = 0, length = 20)
    probs <- dnbinom(x = fails, size = k, p = p)
    plot(x = fails, y = probs, type = 'h', las = 1,
```

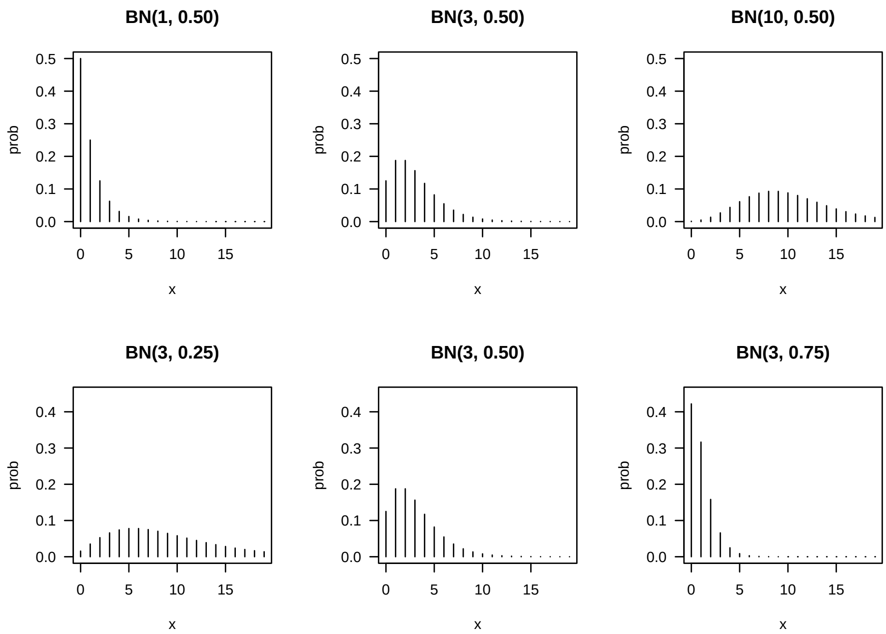


Figure 3.4: Distribuzione binomiale negativa

```
         xlab = 'x', ylab = 'prob', xlim = range(fails),
         main = if (plot_main) sprintf('BN(%d, %.2f)', k, p) else '',
         ...)
}

par(mfrow = c(2,3))
plot_binom_neg(k =  1, p = 0.5, ylim = c(0, 0.5))
plot_binom_neg(k =  3, p = 0.5, ylim = c(0, 0.5))
plot_binom_neg(k = 10, p = 0.5, ylim = c(0, 0.5))
## incremento di p
plot_binom_neg(k =  3, p = 0.25, ylim = c(0, 0.45))
plot_binom_neg(k =  3, p = 0.5, ylim = c(0, 0.45))
plot_binom_neg(k =  3, p = 0.75, ylim = c(0, 0.45))
```

3.7.5 Alternative definition

3.7.5.1 Definition

Definition 3.7.3 (Definizione con numero di prove)**.** Il numero di *prove* indipendenti X (ciascuna con probabilità p di essere successo) necessarie per avere $k \geq 1$ successi si distribuisce come una binomiale negativa di parametri k e p, ossia $X \sim \mathrm{Nb}\left(k, p\right)$.

3.7.5.2 Functions

Remark 133 (Supporto e spazio parametrico).

$$\begin{aligned} R_X &= \{k, k+1, \dots\} \\ \mathbf{\Theta} &= \{k \in \mathbb{N} \setminus \{0\}, p \in \mathbb{R} : 0 \le p \le 1\} \end{aligned}$$

Definition 3.7.4 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k} \cdot \mathbb{1}_{R_X}(x) \tag{3.36}$$

Remark 134 (Interpretazione)*.* La formula deriva dalla considerazione che per ottenere il k-esimo successo nella n-esima prova, ci dovranno essere $k-1$ successi nelle prime $n-1$ prove, la cui probabilità

$$\binom{n-1}{k-1} p^{k-1} (1-p)^{n-k}$$

è moltiplicata per la probabilità di un successo nella n-esima, ossia p.

3.7.5.3 Moments

Proposition 3.7.3 (Momenti caratteristici)**.**

$$\begin{aligned} \mathbb{E}[X] &= \frac{k}{p} \\ \mathrm{Var}[X] &= \frac{k(1-p)}{p^2} \\ \mathrm{Asym}(X) &= \frac{2-p}{\sqrt{k(1-p)}} \\ \mathrm{Kurt}(X) &= 3 + \frac{6}{k} + \frac{p^2}{k(1-p)} \end{aligned}$$

3.8 Recap generalizzazioni

Abbiamo che bernoulliana e geometrica sono specializzazioni di binomiale e binomiale negativa

NB: Rigo's realm here

$$\begin{aligned} \mathrm{Bern}(p) &= \mathrm{Bin}(1, p) \quad \text{ovvero } n = 1 \\ \mathrm{Geom}(p) &= \mathrm{Nb}(1, p) \quad \text{ovvero } k = 1 \end{aligned}$$

Ad esempio se $X \sim \mathrm{Geom}(p)$,

$$\mathbb{P}(X = j) = \binom{j-1}{k-1} p^k (1-p)^{j-k} \overset{(1)}{=} \binom{j-1}{0} p(1-p)^{j-1} = p(1-p)^{j-1}, \quad j = 1, 2, \dots$$

dove in (1) si è sostituito $k = 1$

3.9 Poisson

3.9.1 Definition

Remark 135. È una vc utilizzabile per modellare conteggi (motivo per cui il supporto è $\mathbb{N}$); sull'origine definizione ragioniamo in seguito. Per ora ci accontentiamo di definire la Poisson come la distribuzione caratterizzata dalle funzioni presentate in seguito: se la vc X è distribuita come una Poisson con parametro λ scriveremo $X \sim \text{Pois}(\lambda)$.

Remark 136. L'esperimento che si descrive mediante la Poisson è quello in cui si contano gli arrivi aleatori in un finito intervallo di tempo I. Ad esempio

- conto le visite ad un sito web nell'intervallo I
- conto il numero di autobus che arrivato ad una certa fermata nell'intervallo I

Tutte le volte che si hanno arrivi aleatori, e tali arrivi sodisfano certe ipotesi, la probabilità di avere x arrivi in un dato intervallo si calcola in base alla Poisson.

Remark 137. Un risultato che ci servirà per questa distribuzione è il seguente

Proposition 3.9.1 (Sviluppo di Maclaurin della funzione esponenziale)**.**

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{3.37}$$

Proof. Si ha:

$$e^x = e^0 + \frac{e^0}{1!}(x-0) + \frac{e^0}{2!}(x-0)^2 + \ldots \frac{e^0}{m!}(x-0)^m + \ldots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

□

3.9.2 Functions

Remark 138 (Supporto e spazio parametrico).

$$\begin{aligned} R_X &= \mathbb{N} \\ \mathbf{\Theta} &= \{\lambda \in \mathbb{R} : \lambda > 0\} \end{aligned}$$

Definition 3.9.1 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = \frac{e^{(-\lambda)} \cdot \lambda^x}{x!} \cdot \mathbb{1}_{R_X}(x) \tag{3.38}$$

Validità PMF. Si ha:

$$\sum_{x=0}^{\infty} p_X(x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \stackrel{(1)}{=} e^{-\lambda} e^{\lambda} = 1$$

dove in (1) abbiamo sfruttato la 3.37 con le dovute sostituzioni di lettere. □

3.9.3 Moments

Proposition 3.9.2 (Momenti caratteristici)**.**

$$\mathbb{E}\left[X\right] = \lambda \tag{3.39}$$
$$\mathrm{Var}\left[X\right] = \lambda \tag{3.40}$$
$$\mathrm{Asym}\left(X\right) = \frac{1}{\sqrt{\lambda}} \tag{3.41}$$
$$\mathrm{Kurt}\left(X\right) = 3 + \frac{1}{\lambda} \tag{3.42}$$

Proof. Per il valore atteso

$$\begin{aligned}\mathbb{E}\left[X\right] &= \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda}\lambda^x}{x!} \overset{(1)}{=} e^{-\lambda}\sum_{x=1}^{\infty} x\frac{\lambda^x}{x!} = \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \\ &\overset{(2)}{=} \lambda e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = \lambda e^{-\lambda}e^{\lambda} = \lambda\end{aligned}$$

dove in (1) abbiamo anche portato avanti di 1 la sommatoria dato che il primo termine è nullo e in (2) abbiamo sostituito $y = x - 1$ e sfruttato 3.37. □

Proof. Per la varianza troviamo innanzitutto $\mathbb{E}\left[X^2\right]$:

$$\mathbb{E}\left[X^2\right] = \sum_{x=0}^{\infty} x^2 \cdot \mathbb{P}\left(X = x\right) = \sum_{x=0}^{\infty} x^2\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty} x^2\frac{\lambda^x}{x!}$$

Ora prendiamo la serie dell'esponenziale e la deriviamo rispetto a λ ad entrambi i membri (x costante)

$$e^{\lambda} = \sum_{x=0}^{\infty}\frac{\lambda^x}{x!} \overset{(1)}{=} \sum_{x=0}^{\infty} x\frac{\lambda^{x-1}}{x!} \overset{(2)}{=} \sum_{x=1}^{\infty} x\frac{\lambda^{x-1}}{x!}$$

dove in (1) abbiamo effettuato la derivazione (il primo membro rimane invariato), in (2) abbiamo portato avanti l'indice di sommatoria perché il primo termine è nullo. Ora moltiplicando per λ entrambi i lati si ottiene

$$\lambda e^{\lambda} = \sum_{x=1}^{\infty} x\frac{\lambda^x}{x!}$$

Effettuando gli stessi passaggi, nell'ordine derivare entrambi i membri rispetto a λ e moltiplicandoli per λ si prosegue come

$$\begin{aligned}\sum_{x=1}^{\infty} x^2\frac{\lambda^{x-1}}{x!} &= e^{\lambda} + \lambda e^{\lambda} = e^{\lambda}(1+\lambda) \\ \sum_{x=1}^{\infty} x^2\frac{\lambda^x}{x!} &= e^{\lambda}\lambda(1+\lambda)\end{aligned}$$

E infine riprendendo da dove eravamo arrivati con la main quest

$$\mathbb{E}\left[X^2\right] = e^{-\lambda}\sum_{x=0}^{\infty} x^2\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda}\lambda(1+\lambda) = \lambda(1+\lambda)$$

per cui

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \lambda(1+\lambda) - \lambda^2 = \lambda$$

□

Proof. Dimostrazione alternativa per la varianza:

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}\left[X^2\right] - [\mathbb{E}[X]]^2 \\
&= \left(\sum_{x=0}^{\infty} x^2 \cdot \frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} (x^2 + x - x) \cdot \frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} (x(x-1) + x) \cdot \frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} (x(x-1)) \frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^2\lambda^{x-2}}{x(x-1)(x-2)!} + \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda\lambda^{x-1}}{x(x-1)!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} e^{-\lambda}\lambda^2 + \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda}\lambda\right) - \lambda^2 \\
&\overset{(1)}{=} \left(\sum_{z=0}^{\infty} \frac{\lambda^z}{z!} e^{-\lambda}\lambda^2 + \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda}\lambda\right) - \lambda^2 \\
&= (e^{\lambda}e^{-\lambda}\lambda^2 + e^{\lambda}e^{-\lambda}\lambda) - \lambda^2 \\
&= (\lambda^2 + \lambda) - \lambda^2 \\
&= \lambda
\end{aligned}$$

dove in (1) abbiamo posto $y = x-1$, $z = x-2$ per sfruttare 3.37 nel seguito. □

3.9.4 Shape

Remark 139 (Shape)*.* Quindi valore medio e varianza della vc di Poisson coincidono con il parametro λ; la distribuzione ha picco intorno a λ. Al crescere di questo, la distribuzione diventa più simmetrica e la curtosi tende a 3 (convergendo ad una Normale). Se $\lambda < 1$ la distribuzione ha un andamento decrescente, mentre se > 1 è prima crescente e poi decrescente. (figura 3.5)

```
plot_pois <- function(lambda, plot_main = TRUE, ...){
    x <- 0:10
    probs <- dpois(x = x, lambda = lambda)
    plot(x = x, y = probs, type = 'h', las = 1,
         xlab = 'x', ylab = 'prob', xlim = range(x),
         main = if (plot_main) sprintf('Pois(%.1f)', lambda) else '',
         ...)
}
```

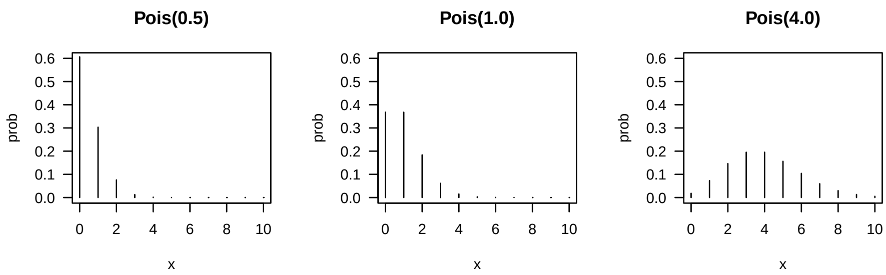


Figure 3.5: Distribuzione Poisson

```
par(mfrow = c(1,3))
tmp <- lapply(c(0.5, 1, 4), plot_pois, ylim = c(0, 0.6))
```

3.9.5 Origine e approssimazione

Remark 140. In pratica si utilizza la Poisson quando si contano gli arrivi aleatori in un fissato intervallo di tempo. Tuttavia, per poter utilizzare la Poisson, tali arrivi devono soddisfare alcune ipotesi. Vediamo quali.

NB: Rigo style here

Ad ogni intervallo $I \subset [0, 1]$ associamo una vc $N(I)$ il cui significato è

$$N(I) = \text{numero arrivi nell'intervallo } I$$

Supponiamo inoltre che

1. $N(I_1), \ldots, N(I_k)$ sono v.c. indipendenti se $I_1, \ldots, I_k$ sono intervalli a due a due disgiunti (si pensi a diversi sottoinsiemi/intervallini nel periodo tra 0 e 1). In altre parole, gli arrivi corrispondenti ad intervalli disgiunti danno danno luogo a c.v. indipendenti

2. $\mathbb{P}(N(I) = 1) = \lambda m(I) + f[m(i)]$ dove $m(I)$ è la lunghezza dell'intervallo I ed f è una funzione tale che
$$\lim_{x \to 0} \frac{f(x)}{x} = 0$$

3. $\mathbb{P}(N(I) > 1) = g[m(i)]$ dove g è una funzione tale che
$$\lim_{x \to 0} \frac{g(x)}{x} = 0$$

Se valgono le condizioni 1-3 allora

$$\mathbb{P}(N([0, 1]) = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad \forall x = 0, 1, 2, \ldots$$

dove λ è un parametro > 0 il cui significato è

$$\lambda = \lim_{n \to +\infty} \frac{N([0, n])}{n} \quad \text{q.c.}$$

dove

- per q.c. si intendeche l'equaglianza è solo "quasi certa"
- $N([0, n])$ denota il numero di arrivi nell'intervallo $[0, n]$

Analogamente λ può essere pensato come il valore medio di X cioè $\lambda = \mathbb{E}[X]$

Remark 141. In questo caso non dimostriamo il risultato precedente

Example 3.9.1 (Dalla binomiale alla poisson, Rigo style)**.** Se $X_n \sim \text{Bin}(n, p_n)$, $X \sim \text{Pois}(\lambda)$ ed $np_n \to \lambda$ allora

$$\mathbb{P}(X = x) = \lim_{n\to+\infty} \mathbb{P}(X_n = x), \forall x = 0, 1, 2, \dots$$

Per fare i conti, occorre richiamare un risultato di analisi

$$\lim_{n\to+\infty} \left(1 + \frac{1}{n}\right)^n = e,$$
$$\lim_{n\to+\infty} \left(1 + \frac{a}{n}\right)^n = e^a$$

Più in generale: se $\lim_{n\to+\infty} a_n = a$ allora $\lim_{n\to+\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a$. Allora:

$$\begin{aligned} \mathbb{P}(X_n = x) &= \binom{n}{x} p_n^x (1 - p_n)^{n-x} \\ &= \frac{1}{x!} \underbrace{\frac{n(n-1)\cdot \ldots \cdot (n-x+1)}{n^x}}_{=1} \underbrace{(np_n)^x}_{=\lambda^x} \underbrace{(1-p_n)^{-x}}_{(1-0)^{-j}=1} \underbrace{\left(1 - \frac{np_n}{n}\right)^n}_{e^{-\lambda}} \end{aligned}$$

Quindi

$$\lim_{n\to\infty} P(X_n = x) = \frac{1}{x!} \cdot 1 \cdot \lambda^x \cdot 1 \cdot e^{-\lambda} = \frac{e^{-\lambda}\lambda^x}{x!} = \mathbb{P}(X = x)$$

NB: Altri non-rigo appunti sotto

Remark 142. È utilizzata per modellare il numero di eventi registrati in un ambito circoscritto (*temporale* o *spaziale*), in cui vi è un largo numero di prove indipendenti (o quasi) caratterizzate ciascuna da una bassa probabilità di successo (per questa è chiamata legge degli eventi rari)

Proposition 3.9.3 (Paradigma di Poisson)**.** *Siano $E_1, \dots, E_n$ eventi con $p_i = \mathbb{P}(E_i)$, dove n è largo, p_i sono piccoli e gli E_i sono vc indipendenti o debolmente dipendenti. Sia*

$$X = \sum_{i=1}^{n} I_{E_i}$$

la somma di quanti eventi E_i siano accaduti. Allora X è abbastanza bene distribuita come una $\text{Pois}(\lambda)$ *con $\lambda = \sum_i p_i$.*

Proof. La prova dell'approssimazione di sopra è complessa, richiede definire la dipendenza debole e buona approssimazione; è omessa qui. □

Remark 143 (Ruolo di λ). Il parametro λ è interpretato come *rate di occorrenza*: ad esempio $\lambda = 2$ mail di spam per giorno.

Remark 144. Nell'esempio sopra il numero di eventi X non è esattamente distribuito come Poisson perché una variabile di Poisson non ha limite superiore, mentre $I_{E_1} + \ldots + I_{E_n}$ somma al più a n. Ma la distribuzione di Poisson da spesso una buona approssimazione e le condizioni per il verificarsi della situazione di sopra sono abbastanza flessibili: infatti i p_i non devono essere uguali e le prove non devono essere strettamente indipendenti. Questo fa sì che il modello di Poisson sia spesso un buon punto di partenza per dati che assumono valore intero non negativo (chiamati conteggi)
È comunque possibile quantificare l'errore commesso.

Proposition 3.9.4 (Errore di approssimazione)**.** *Se E_i sono indipendenti e sia $N \sim \text{Pois}(\lambda)$, allora l'errore di approssimazione che si fa nell'utilizzare la poisson per stimare la probabilità di un dato set di interi non negativi $I \subset \mathbb{N}$, è dato dalla seguente:*

$$\mathbb{P}(X \in I) - \mathbb{P}(N \in I) \leq \min\left(1, \frac{1}{\lambda}\right) \sum_{i=1}^{n} p_i^2 \tag{3.43}$$

Proof. Anche questa è per ora complessa (necessita di una tecnica chiamata metodo di Stein). □

Remark 145. La 3.43 fornisce un limite superiore dell'errore commesso nell'utilizzare unaapprossimazione di Poisson: non solo per l'intera distribuzione (se $I = \mathbb{N}$) ma per qualsiasi suo sottoinsieme. Altresì precisa quanto i p_i dovrebbero essere piccoli: vogliamo che $\sum_{i=1}^{n} p_i^2$ sia molto piccolo, o quanto meno lo sia rispetto a λ.

3.9.6 Legami con la binomiale

Remark 146. La relazione tra Poisson e Binomiale è simile a quella intercorrente tra Binomiale e Ipergeometrica: possiamo andare dalla Poisson alla binomiale condizionando, e viceversa dalla Binomiale alla Poisson prendendo un limite. Prima un risultato strumentale.

Proposition 3.9.5 (Somma di Poisson indipendenti)**.** *Siano $X \sim \text{Pois}(\lambda_1)$ e $Y \sim \text{Pois}(\lambda_2)$ vc indipendenti. Allora $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$*

Proof. Per ottenere la PMF di $X + Y$ condizioniamo su X e utilizziamo il teorema delle probabilità totali

$$\begin{aligned}
\mathbb{P}(X+Y=k) &= \sum_{j=0}^{k} \mathbb{P}(X+Y=k|X=j)\cdot\mathbb{P}(X=j) \\
&= \sum_{j=0}^{k} \mathbb{P}(Y=k-x|X=j)\cdot\mathbb{P}(X=j) \\
&\overset{(1)}{=} \sum_{j=0}^{k} \mathbb{P}(Y=k-x)\cdot\mathbb{P}(X=j) \\
&= \sum_{j=0}^{k} \frac{e^{-\lambda_2}\lambda_2^{k-j}}{(k-j)!}\frac{e^{-\lambda_1}\lambda_1^{j}}{(j)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{j=0}^{k}\binom{k}{j}\lambda_1^j\lambda_2^{k-j} \\
&\overset{(2)}{=} \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^k)}{k!} = \text{Pois}(\lambda_1+\lambda_2)
\end{aligned}$$

con (1) data l'indipendenza e in (2) si è utilizzato il teorema binomiale $(a+b)^n = \sum_{i=0}^{n}\binom{n}{i}a^ib^{n-i}$ □

Remark 147. A intuito se vi sono due tipi di eventi che accadono ai rate λ_1 e λ_2 indipendentemente, allora il rate complessivo di eventi è $\lambda_1+\lambda_2$.

3.9.6.1 Dalla Poisson alla binomiale

Proposition 3.9.6. *Se $X \sim \text{Pois}(\lambda_1)$ e $Y \sim \text{Pois}(\lambda_2)$ sono indipendenti, allora la ditribuzione condizionata di X dato che $XY = n$ è* $\text{Bin}(n, \lambda_1/(\lambda_1\lambda_2))$.

Proof. Utilizziamo la regola di Bayes per calcolare la PMF condizionata $\mathbb{P}(X=x|X+Y=n)$:

$$\begin{aligned}
\mathbb{P}(X=x|X+Y=n) &= \frac{\mathbb{P}(X+Y=n|X=x)\cdot\mathbb{P}(X=x)}{\mathbb{P}(X+Y=n)} \\
&= \frac{\mathbb{P}(Y=n-x|X=x)\cdot\mathbb{P}(X=x)}{\mathbb{P}(X+Y=n)} \\
&\overset{(1)}{=} \frac{\mathbb{P}(Y=n-x)\cdot\mathbb{P}(X=x)}{\mathbb{P}(X+Y=n)}
\end{aligned}$$

con (1) per indipendenza delle due. Ora sostituendo le PMF di X, Y e $X+Y$; questa al denominatore è distribuita come $\text{Pois}(\lambda_1+\lambda_2)$ per proposizione 3.9.5.

Si ha:

$$
\begin{aligned}
\mathbb{P}(X=k|X+Y=n) &= \frac{\left(\frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!}\right)\left(\frac{e^{\lambda_1}\lambda_2^{k}}{k!}\right)}{\frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}} = \frac{\frac{e^{-(\lambda_1+\lambda_2)}\cdot\lambda_1^k\cdot\lambda_2^{n-k}}{k!(n-k)!}}{\frac{e^{-(\lambda_1+\lambda_2)}\cdot(\lambda_1+\lambda_2)^n}{n!}} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}\cdot\lambda_1^k\cdot\lambda_2^{n-k}}{k!(n-k)!}\cdot\frac{n!}{e^{-(\lambda_1+\lambda_2)}\cdot(\lambda_1+\lambda_2)^n} \\
&= \frac{n!}{k!(n-k)!}\cdot\frac{\lambda_1^k\cdot\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} \\
&= \binom{n}{k}\left(\frac{\lambda_1^k}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2^k}{\lambda_1+\lambda_2}\right)^{n-k} \\
&= \mathrm{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)
\end{aligned}
$$

□

3.9.6.2 Dalla binomiale alla Poisson

Remark 148. Viceversa se prendiamo il limite della $\mathrm{Bin}(n,p)$ per $n\to\infty$ e $p\to 0$ con np fisso arriviamo alla Poisson.

Proposition 3.9.7 (Approssimazione Poissoniana della binomiale)**.** *Se $X\sim$ $\mathrm{Bin}(n,p)$ e facciamo tendere $n\to\infty$, $p\to 0$ ma $\lambda=np$ rimane fisso, allora la PMF di X converge a* $\mathrm{Pois}(\lambda)$.
La stessa conclusione si ha se $n\to\infty$, $p\to 0$ ed np converge *ad una costante λ.*

Remark 149. Questo è un *caso speciale* del paradigma di Poisson dove E_i sono indipendenti e hanno la stessa probabilità, quindi $\sum_{i=1}^n I_{E_i}$ ha distribuzione binomiale. In questo caso speciale possiamo dimostrare che l'approssimazione di Poisson ha senso limitandoci a prendere il limite della Binomiale.

Proof. Effettueremo la dimostrazione per $\lambda=np$ fisso (considerando $p=\lambda/n$), mostrando che la PMF $\mathrm{Bin}(n,p)$ converge alla $\mathrm{Pois}(\lambda)$. Per $0\le x\le n$:

$$
\begin{aligned}
\mathbb{P}(X=x) &= \binom{n}{x}p^x(1-p)^{n-x} \\
&= \frac{n(n-1)\cdot\ldots\cdot(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!}\frac{n(n-1)\cdot\ldots\cdot(n-x+1)}{n^x}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}
\end{aligned}
$$

Per $n\to\infty$ con k fisso

$$
\frac{\overbrace{n(n-1)\cdot\ldots\cdot(n-x+1)}^{x\text{ termini}}}{n^x} \overset{(1)}{=} \frac{n\cdot n\left(1-\frac{1}{n}\right)\cdot\ldots\cdot n\left(1-\frac{k-1}{n}\right)}{n^x}\to 1
$$

$$
\left(1-\frac{\lambda}{n}\right)^n\to e^{-\lambda}
$$

$$
\left(1-\frac{\lambda}{n}\right)^{-k} = \left[\left(1-\frac{\lambda}{n}\right)^n\right]^{-\frac{k}{n}}\to e^{-\frac{k}{n}} = 1
$$

dove in (1) abbiamo raccolto un n a partire dal secondo fattore, lasciando fuori parentesi k n che si moltiplicano. Pertanto

$$\mathbb{P}(X = x) \to \frac{e^{-\lambda}\lambda^x}{x!} = \text{Pois}(\lambda)$$

□

Remark 150. Il precedente risultato implica che se n è grande, p piccolo e np moderato, possiamo approssimare $\text{Bin}(n, p)$ con $\text{Pois}(np)$; come visto in precedenza l'errore nell'approssimare $\mathbb{P}(X \in I)$ con $\mathbb{P}(N \in I)$ per $X \sim \text{Bin}(n, p)$ e $N \sim \text{Pois}(np)$ è al massimo $\min(p, np^2)$.

Example 3.9.2. Il proprietario di un sito vuole studiare la distribuzione del numero di visitatori. Ogni giorno un millione di persone in maniera indipendente decide se visitare il sito o meno, con probabilità $p = 2 \times 10^{-1}$. Fornire una approssimazione della probabilità di avere almeno tre visitatori al giorno.
Se $X \sim \text{Bin}(n, p)$ è il numero di visitatori con $n = 10^6$, fare i calcoli con la binomiale va incontro a difficoltà computazionali ed errori numerici del pc (dato che n è largo e p molto basso). Ma data la situazione con n largo p basso e $np = 2$ moderato, $\text{Pois}(2)$ è una buona approssimazione. Questo porta a

$$\mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X < 3) \approx 1 - e^{-2} - e^{-2} \cdot 2 - e^{-2} \cdot \frac{2^2}{2!} = 1 - 5e^{-2} \approx 0.3233$$

che è una approssimazione molto accurata.

3.9.7 Processo di Poisson

Definition 3.9.2 (Processo di Poisson)**.** È una insieme di prove E_i che si possono verificare ciascuna in un dato arco temporale $[0, T]$. Le prove sono svolte nelle medesime condizioni e soddisfano di assiomi:

- il verificarsi di E nell'intervallo (t_1, t_2) è indipendente dal verificarsi di E nell'intervallo (t_3, t_4) (se gli intervalli non si sovrappongono);
- la probabilità del verificarsi di E in un intervallo infinitesimo $(t_0, t_0 + \mathrm{d}t)$ è proporzionale ad un parametro $\lambda > 0$ che caratterizza la prova;
- la probabilità che due eventi si verifichino nello stesso intervallo di tempo è un infinitesimo di ordine superiore rispetto alla probabilità che se ne verifichi soltanto uno.

3.10 Discrete uniform

3.10.1 Definition

Remark 151. La prova che genera la vc Uniforme discreta si può assimilare all'estrazione di una pallina da un'urna che contiene n palline identiche numerate da 1 a n. Viene in genere utilizzata quanto tutti i risultati dell'esperimento sono equiprobabili

Definition 3.10.1 (Uniforme discreta)**.** Il numero X della pallina estratta dall'urna contenente n palline numerate (da 1 a n) si distribuisce come Uniforme discreta $X \sim \text{DUnif}(n)$.

3.10.2 Functions

Remark 152 (Supporto e spazio parametrico).

$$R_X = \{1, \ldots, n\}$$
$$\boldsymbol{\Theta} = \{n \in \mathbb{N} \setminus \{0\}\}$$

Proposition 3.10.1 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = \frac{1}{n} \cdot \mathbb{1}_{R_X}(x) \tag{3.44}$$

Definition 3.10.2 (Funzione di ripartizione)**.**

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{se } x < 1 \\ \frac{k}{n} & \text{se } k \leq x < k+1, (k = 1, 2, \ldots, n-1) \\ 1 & \text{se } x \geq n \end{cases} \tag{3.45}$$

Remark 153. La funzione di ripartizione è nulla in $(\infty; 1)$ ed è una funzione a gradini di altezza costante pari a $1/n$, in corrispondenza di ogni valore intero $1 \leq x \leq n$ e vale 1 in $[n; +\infty)$.

3.10.3 Moments

Proposition 3.10.2 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{n+1}{2} \tag{3.46}$$
$$\operatorname{Var}[X] = \frac{n^2 - 1}{12} \tag{3.47}$$
$$\operatorname{Asym}(X) = 0 \tag{3.48}$$
$$\operatorname{Kurt}(X) = 1.8 \tag{3.49}$$

Proof.

$$\mathbb{E}[X] = \sum_{x=1}^{n} x \frac{1}{n} = \frac{1}{n}(1 + 2 + \ldots + n) = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$$

□

Proof.

$$\begin{aligned}
\mathrm{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - \left[\mathbb{E}\left[x\right]\right]^2 &= \left(\sum_{x=1}^{n} x^2 \frac{1}{n}\right) - \left(\frac{n+1}{2}\right)^2 \\
&= \left(\frac{1}{n}(1^2 + 2^2 + \ldots + n^2)\right) - \left(\frac{n+1}{2}\right)^2 \\
&= \left(\frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6}\right) - \left(\frac{n^2+1+2n}{4}\right) \\
&= \left(\frac{(n+1)(2n+1)}{6}\right) - \left(\frac{n^2+1+2n}{4}\right) \\
&= \frac{2(2n^2+2n+n+1) - 3(n^2+1+2n)}{12} \\
&= \frac{4n^2+4n+2n+2-3n^2-3-6n}{12} \\
&= \frac{n^2-1}{12}
\end{aligned}$$

□

Chapter 4

Absolute continuous random variables

4.1 Logistica

4.1.1 Origine/definizione

Remark 154. Viene utilizzata per modelli di crescita di grandezze nel tempo, dove la crescita segue le fasi di crescita esponenziale, saturazione e arresto. Un buon modello per rappresentare fenomeni di questo tipo è rappresentato dalla funzione di ripartizione logistica.

Remark 155. Deriva il nome dall'avere la funzione di ripartizione che soddisfa l'equazione logistica: $F'(x) = \frac{1}{s}F(x)(1 - F(x))$.

Remark 156. E' matematicamente semplice e ci permette di focalizzarci su aspetti non numerici; è altresì importante nella regressione logistica.

4.1.2 Funzioni

Definition 4.1.1 (Funzione di ripartizione)**.** Ha CDF

$$F_X(x) = \mathbb{P}(X \leq x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}, \qquad x \in \mathbb{R} \tag{4.1}$$

Remark 157. Si trovano entrambe le definizioni (si passa dall'una all'altra moltiplicando/dividendo a numeratore e denominatore per e^x)

Definition 4.1.2 (Funzione di densità)**.** Derivando entrambe le espressioni si hanno, equivalentemente:

$$f_x(x) = \frac{e^x}{(1+e^x)^2} = \frac{e^{-x}}{(1+e^{-x})^2} \tag{4.2}$$

4.1.3 Versione generale

Remark 158 (Supporto e spazio parametrico).

$$\begin{aligned} R_X &= \mathbb{R} \\ \boldsymbol{\Theta} &= \{\mu \in \mathbb{R}, s \in \mathbb{R} : s > 0\} \end{aligned}$$

Definition 4.1.3 (Funzione di ripartizione)**.** La funzione di densità di una vc $X \sim \text{Logistic}(\mu, \sigma)$ è

$$F_X(x) = \frac{e^{\frac{x-\mu}{\sigma}}}{\left(1 + e^{\frac{x-\mu}{\sigma}}\right)} \cdot \mathbb{1}_{R_X}(x) \tag{4.3}$$

Definition 4.1.4 (Funzione di densità)**.** La funzione di densità di una vc $X \sim \text{Logistic}(\mu, \sigma)$ è

$$f_X(x) = \frac{e^{\frac{x-\mu}{\sigma}}}{\sigma\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2} \cdot \mathbb{1}_{R_X}(x) \tag{4.4}$$

Proposition 4.1.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \mu$$
$$\text{Var}[X] = \frac{\pi^2}{3}\sigma^2$$

TODO: perché la varianza non è σ^2 applicando le regole su trasf lineari?

Mia dimostrazione, controllare. Sia $Z \sim \text{Logistic}(0, 1)$ e sia $X = \sigma Z + \mu$, con σ parametro di scala e μ di posizione. Allora si ha che

$$Z = \frac{X - \mu}{\sigma} \sim \text{Logistic}(0, 1)$$

Per cui possiamo scrivere che

$$F_X(x) = \frac{e^{\frac{x-\mu}{\sigma}}}{1 + e^{\frac{x-\mu}{\sigma}}}$$

Derivando per ottenere $f_X(x)$ si ha

$$\begin{aligned} f_X(x) &= \frac{\left(e^{\frac{x-\mu}{\sigma}} \cdot \frac{1}{\sigma}\right)\left(1 + e^{\frac{x-\mu}{\sigma}}\right) - \left(e^{\frac{x-\mu}{\sigma}} \cdot \frac{1}{\sigma}\right)\left(e^{\frac{x-\mu}{\sigma}}\right)}{\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2} = \frac{\left(e^{\frac{x-\mu}{\sigma}} \cdot \frac{1}{\sigma}\right)\left(1 + e^{\frac{x-\mu}{\sigma}} - e^{\frac{x-\mu}{\sigma}}\right)}{\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2} \\ &= \frac{e^{\frac{x-\mu}{\sigma}}}{\sigma\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2} \end{aligned}$$

□

```r
par(mfrow = c(1,2))
## mu <- c(5, 9, 9, 6, 2)
## s  <- c(2, 3, 4, 2, 1)
mu <- c(0, 2, 2 ,5, 5)
s  <- c(1, 3, 4, 3, 4)

tmp <- Map(function(mu, s, add, col) {
```

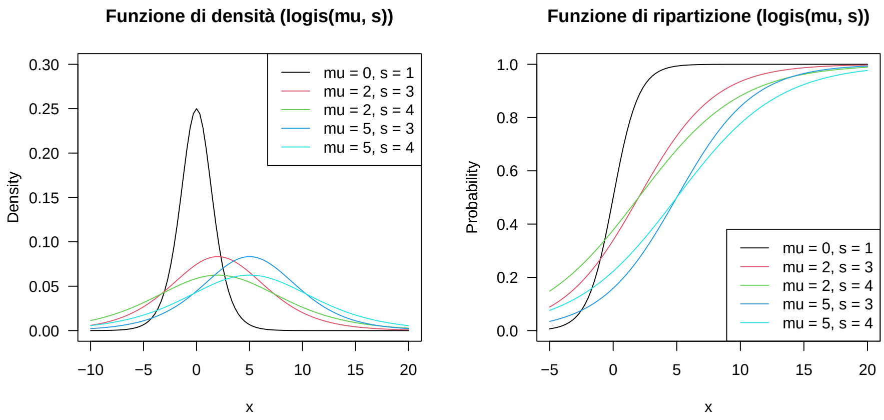


Figure 4.1: Distribuzione logistica

```
    plot_fun(function(x) dlogis(x, location = mu, scale = s),
             from = -10, to = 20,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 0.3),
             ylab = 'Density', las = 1,
             main = 'Funzione di densità (logis(mu, s))')
}, as.list(mu), as.list(s), as.list(c(F, T, T, T, T)), as.list(1:5))
leg <- unlist(Map(function(mu, s) sprintf('mu = %d, s = %d', mu, s), mu, s))
legend('topright', legend = leg,  col = 1:5, lty = 'solid')

tmp <- Map(function(mu, s, add, col) {
    plot_fun(function(x) plogis(x, location = mu, scale = s),
             from = -5, to = 20,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 1),
             ylab = 'Probability', las = 1,
             main = 'Funzione di ripartizione (logis(mu, s))')
}, as.list(mu), as.list(s), as.list(c(F, T, T, T, T)), as.list(1:5))
leg <- unlist(Map(function(mu, s) sprintf('mu = %d, s = %d', mu, s), mu, s))
legend('bottomright', legend = leg,  col = 1:5, lty = 'solid')
```

4.2 Uniforme continua

Remark 159. È una vc continua X definita sul supporto (a, b), con $a < b$ ed ed esiti aventi la medesima densità, indicata con $X \sim \text{Unif}\,(a, b)$

Remark 160. Una formulazione usuale per tale modello probabilistico è la uniforma continua sull'intervallo con $a = 0, b = 1$.

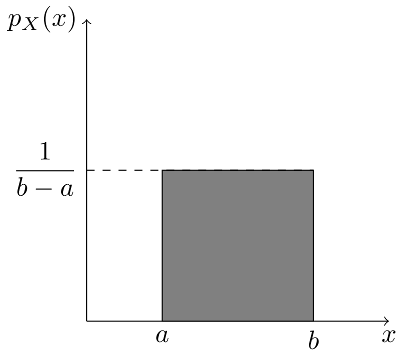


Figure 4.2: Uniforme continua

Remark 161 (Supporto e spazio parametrico).

$$R_X = [a, b]$$
$$\mathbf{\Theta} = \{a, b \in \mathbb{R}, a < b\}$$

Definition 4.2.1 (Funzione di densità)**.** In figura 4.2

$$f_X(x) = \frac{1}{b-a} \cdot \mathbb{1}_{R_X}(x) \tag{4.5}$$

Proposition 4.2.1. *L'area è 1.*

Proof.

$$(b-a) \cdot \frac{1}{(b-a)} = 1$$

□

Definition 4.2.2 (Funzione di ripartizione)**.**

$$F_X(x) = \begin{cases} 0 & \text{per } x \leq a \\ \frac{x-a}{b-a} & \text{se } a < x < b \\ 1 & \text{per } x \geq b \end{cases} \tag{4.6}$$

Proposition 4.2.2 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{a+b}{2} \tag{4.7}$$
$$\mathrm{Var}[X] = \frac{(b-a)^2}{12} \tag{4.8}$$
$$\mathrm{Asym}(X) = 0 \tag{4.9}$$
$$\mathrm{Kurt}(X) = 1.8 \tag{4.10}$$

Proof.

$$
\begin{aligned}
\mathbb{E}[X] &= \int_a^b x\frac{1}{b-a}\,\mathrm{d}x = \left[\frac{x^2}{2(b-a)}\right]_a^b \\
&= \left(\frac{b^2}{2(b-a)} + c\right) - \left(\frac{a^2}{2(b-a)} + c\right) \\
&= \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}
\end{aligned}
$$

□

Proof.

$$
\begin{aligned}
\mathrm{Var}[X] &= \left(\int_a^b x^2\frac{1}{b-a}\,\mathrm{d}x\right) - \left(\frac{a+b}{2}\right)^2 \\
&= \left[\frac{x^3}{3(b-a)}\right]_a^b - \left(\frac{a+b}{2}\right)^2 \\
&= \left(\frac{b^3}{3(b-a)} + c\right) - \left(\frac{a^3}{3(b-a)} + c\right) - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} \\
&= \frac{(b-a)(a^2+b^2+ab)}{3(b-a)} - \frac{(a+b)^2}{4} \\
&= \frac{a^2+b^2+ab}{3} - \frac{a^2+b^2+2ab}{4} \\
&= \frac{4a^2+4b^2+4ab-3a^2-3b^2-6ab}{12} \\
&= \frac{a^2+b^2-2ab}{12} = \frac{(a-b)^2}{12} = \frac{(b-a)^2}{12}
\end{aligned}
$$

□

Remark 162. Si tratta di una variabile simmetrica e platicurtica (ovvero con una distribuzione molto piatta).

4.3 Esponenziale

Remark 163. L'esponenziale è generalmente usata per fenomeni di cui interessa un tempo/durata t (di vita, resistenza, funzionamento).
La derivazione può avvenire se si ipotizza una funzione di rischio/azzardo costante $H(t) = \lambda > 0$, con λ tasso di occorrenza dell'evento (reciproco del numero di eventi per unità di tempo).

Remark 164 (Supporto e spazio parametrico).

$$
\begin{aligned}
R_X &= \{x \in \mathbb{R} : x > 0\} \\
\boldsymbol{\Theta} &= \{\lambda \in \mathbb{R} : \lambda > 0\}
\end{aligned}
$$

Definition 4.3.1 (Distribuzione esponenziale)**.** Se $H(t) = \lambda > 0$ la funzione di ripartizione si ricava dalla 2.25 come

$$
\begin{aligned}
F_X(t) &= 1 - \exp\left(-\int_0^t H(w)\,\mathrm{d}w\right) = 1 - \exp\left(-\int_0^t \lambda\,\mathrm{d}w\right) \\
&= 1 - \exp\left(-\lambda t\right)
\end{aligned}
$$

Definition 4.3.2 (Funzione di ripartizione)**.**

$$
F_X(x) = \begin{cases} 1 - \exp(-\lambda x) & \text{per } x \geq 0 \\ 0 & \text{per } x < 0 \end{cases} \tag{4.11}
$$

Remark 165. La funzione di densità si ottiene derivando dalla 4.11; pertanto una vc continua X si dice vc Esponenziale con parametro $\lambda > 0$, e si scrive $X \sim \mathrm{Exp}\left(\lambda\right)$ se caratterizzata dalla seguente funzione di densità.

Definition 4.3.3 (Funzione di densità)**.**

$$
f_X(x) = \lambda \exp(-\lambda x) \cdot \mathbb{1}_{R_X}(x) \tag{4.12}
$$

Proposition 4.3.1 (Momenti caratteristici)**.**

$$
\mathbb{E}\left[X\right] = \frac{1}{\lambda} \tag{4.13}
$$

$$
\mathrm{Var}\left[X\right] = \frac{1}{\lambda^2} \tag{4.14}
$$

$$
\mathrm{Asym}\left(X\right) = 2 \tag{4.15}
$$

$$
\mathrm{Kurt}\left(X\right) = 9 \tag{4.16}
$$

NB: rigo here

Proof. Per il valore atteso

$$
\begin{aligned}
\mathbb{E}\left[X\right] &= \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x = \int_0^{+\infty} x\lambda e^{-\lambda x}\,\mathrm{d}x = \lambda \left\{ \left[\frac{e^{-\lambda x}}{-\lambda} x\right]_0^{+\infty} + \frac{1}{\lambda}\int_0^{+\infty} e^{-\lambda x}\,\mathrm{d}x \right\} \\
&= \int_0^{+\infty} e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda}\int_0^{+\infty} e^{-y}\,\mathrm{d}y = \frac{1}{\lambda}
\end{aligned}
$$

□

Remark 166 (Forma distribuzione). Tale funzione è decrescente a partire da $x = 0$, in corrispondenza del quale si registra la moda; è asimmetrica positiva e fortemente leptocurtica (a punta), con asimmetria e curtosi costanti al variare di λ. (figura 4.3)

```
par(mfrow = c(1,2))
lambda <- c(0.5, 1, 1.5)
tmp <- Map(function(l, cp, add, col) {
    plot_fun(function(x) dexp(x, rate = l), from = 0, to = 5,
             cartesian_plane = cp, add = add, col = col, ylim = c(0, 1.5),
             ylab = 'Density', las = 1, main = 'Densità')
}, as.list(lambda), as.list(c(F, F, F)), as.list(c(F, T, T)), as.list(1:3))
```

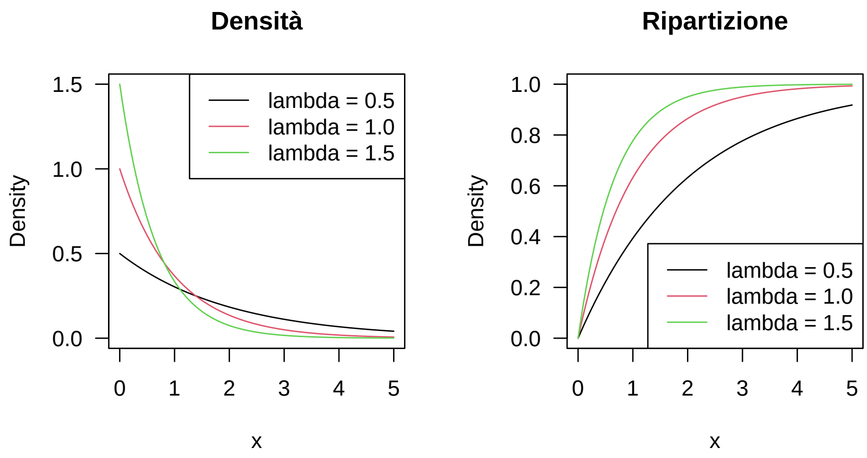


Figure 4.3: Distribuzione esponenziale

```
legend('topright', legend = sprintf("lambda = %.1f", lambda),
       col = 1:3, lty = 'solid' )

tmp <- Map(function(l, cp, add, col) {
    plot_fun(function(x) pexp(x, rate = l), from = 0, to = 5,
             cartesian_plane = cp, add = add, col = col, ylim = c(0, 1),
             ylab = 'Density', las = 1, main = 'Ripartizione')
}, as.list(lambda), as.list(c(F, F, F)), as.list(c(F, T, T)), as.list(1:3))
legend('bottomright', legend = sprintf("lambda = %.1f", lambda),
       col = 1:3, lty = 'solid' )
```

Proposition 4.3.2. *La vc esponenziale è l'unica rv assolutamente continua con* mancanza di memoria

$$\mathbb{P}\left(X > a + b | X > a\right) = \mathbb{P}\left(X > b\right), \quad \forall a, b > 0$$

Remark 167. il tipico fenomeno che si descrive mediante la va esponenziale è la durata in vita (di qualcuno o qualcosa). In tale interpretazione

$$\mathbb{P}\left(X > a + b | X > a\right) = \mathbb{P}\left(\text{sono vivo all'istante } a + b | \text{sono vivo all'istante } a\right) = \mathbb{P}\left(\text{sono vivo all'istante } b\right)$$

Per inciso da questo segue che l'esponenziale NON è un modello adatto a descrivere la durata in vida di un essere vivente,

$$\mathbb{P}\left(X > 85 | X > 80\right) \neq \mathbb{P}\left(X > 5\right)$$

Proof. Dimostriamo che se X è esponenziale, allora valoe la mancanza di memoria

$$\begin{aligned}\mathbb{P}\left(X > a + b | X > a\right) &= \frac{\mathbb{P}\left(X > a + b, X > a\right)}{\mathbb{P}\left(X > a\right)} = \frac{\mathbb{P}\left(X > a + b\right)}{\mathbb{P}\left(X > a\right)} = \frac{1 - F(a+b)}{1 - F(a)} \\ &= \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = \mathbb{P}\left(X > b\right)\end{aligned}$$

Dimostrare che se X è assolutamente contiunua e manca di memoria allora è esponenziale è molto più complicato □

Remark 168. La vc Esponenziale presenta una struttura molto semplice ma rigida, per cui non si adatta facilmente a tutte le situazioni reali; infatti, talvolta non è realistico assumere che la funzione di rischio si costante rispetto al tempo. Pertanto si hanno almeno due generalizzazioni: la Weibull e la Gamma.

4.4 Normale/Gaussiana

Remark 169. Viene utilizzata come prima approssimazione per descrivere variabili casuali a valori reali che tendono a concentrarsi attorno a un singolo valor medio.

Remark 170. Una vc continua si dice vc Normale con parametri μ e σ^2, e la si indica con $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ se è definita su tutto l'asse reale e presenta la seguente funzione di densità.

Remark 171 (Supporto e spazio parametrico).

$$\begin{aligned} R_X &= \{\mathbb{R}\} \\ \boldsymbol{\Theta} &= \left\{\mu \in \mathbb{R}; \sigma^2 \in \mathbb{R} : \sigma^2 > 0\right\} \end{aligned}$$

Definition 4.4.1 (Funzione di densità)**.**

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] \cdot \mathbb{1}_{R_X}(x) \tag{4.17}$$

Remark 172 (Forma della distribuzione). Ha una forma campanulare e simmetrica rispetto al punto di ascissa $x = \mu$, è crescente in $(-\infty, \mu)$ e decrescente in (μ, ∞). In corrispondenza di μ $f_X(x)$ ha il massimo (perché l'esponente negativo è minimo). Pertanto μ è il valore centrale la moda, mediana e valore medio della vc.
Si dimostra che $f_X(x)$ presenta due flessi in corrispondenza di $x = \mu \pm \sigma$. Ha come asintoto l'asse x
μ è un parametro di posizione mentre σ^2 misura la dispersione attorno a μ. La modifica di μ a parità di σ^2 implica una traslazione della funzione di densità lungo l'asse x; invece, al crescere di σ a parità di μ, i flessi si allontanano da μ e la funzione di den attribuisce maggiore probabilità ai valori lontani dal valore centrale (e viceversa al diminuire di σ^2). (figura 4.4)

Definition 4.4.2 (Normale standardizzata)**.** Se $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, la trasformazione lineare $Z = (X - \mu)/\sigma$ definisce la vc Normale standardizzata $Z \sim \mathrm{N}(0, 1)$

Definition 4.4.3 (Funzione di densità (Normale standardizzata))**.**

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \cdot \mathbb{1}_{R_X}(x) \tag{4.18}$$

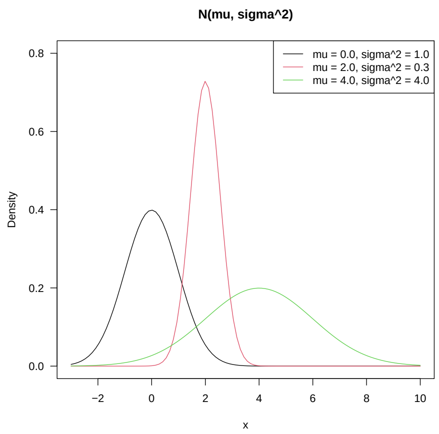


Figure 4.4: Distribuzione normale

```
params <- list(c('mu' = 0, 's2' = 1),
               c('mu' = 2, 's2' = 0.3),
               c('mu' = 4, 's2' = 4))

tmp <- Map(function(p, add, col) {
    ## browser()
    plot_fun(function(x) dnorm(x, mean = p["mu"], sd = sqrt(p["s2"])),
             from = -3, to = 10,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 0.8),
             ylab = 'Density', las = 1, main = 'N(mu, sigma^2)'
             )
}, params, as.list(c(F, T, T)), as.list(1:3))

leg <- unlist(lapply(params, function(x)
    sprintf('mu = %.1f, sigma^2 = %.1f', x['mu'], x['s2'])))
legend('topright', legend = leg,  col = 1:3, lty = 'solid' )
```

Definition 4.4.4 (Funzione di ripartizione (Normale standardizzata))**.**

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) \mathrm{d}w \tag{4.19}$$

Remark 173. La funzione di ripartizione della vc Z non ammette una formulazione esplicita ed è necessario predisporre delle tavole che per opportuni valori di z forniscano l'integrale con sufficiente accuratezza.

Remark 174. Sfruttando la simmetria della funzione di densità, è sufficiente conoscere $\Phi(z)$ per i soli valori di $z > 0$. Infatti $\Phi(0) = 0.5$ ed inoltre:

$$\Phi(-z) = 1 - \Phi(z) \quad \forall z \geq 0 \tag{4.20}$$

Remark 175. La conoscenza della funzione di ripartizione della vc $Z \sim \mathrm{N}(0,1)$ è sufficiente per calcolare la probabilità di qualsiasi vc $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ mediante una semplice trasformazione:

$$\begin{aligned}\mathbb{P}(x_0 < X \leq x_1) &= \mathbb{P}\left(\frac{x_0-\mu}{\sigma} < \underbrace{\frac{X-\mu}{\sigma}}_{Z} \leq \frac{x_1-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x_1-\mu}{\sigma}\right) - \Phi\left(\frac{x_0-\mu}{\sigma}\right)\end{aligned}$$

In pratica per calcolare la probabilità ce una vc normale assuma valori in un intervallo basta standardizzare gli estremi dell'intervallo ed utilizzare le tavole di $\Phi(z)$.

Proposition 4.4.1 (Momenti caratteristici (Normale standardizzata))**.**

$$\mathbb{E}[Z] = 0 \tag{4.21}$$
$$\mathrm{Var}[Z] = 1 \tag{4.22}$$
$$\mathrm{Asym}(Z) = 0 \tag{4.23}$$
$$\mathrm{Kurt}(Z) = 3 \tag{4.24}$$

Proposition 4.4.2 (Momenti caratteristici (Normale))**.** *Da $X = \mu + \sigma Z$ si ha*

$$\mathbb{E}[X] = \mu \tag{4.25}$$
$$\mathrm{Var}[X] = \sigma^2 \tag{4.26}$$
$$\mathrm{Asym}(X) = 0 \tag{4.27}$$
$$\mathrm{Kurt}(X) = 3 \tag{4.28}$$

Remark 176. Nel prosieguo tratteremo della vc Normale standardizzata, per semplicità.

Proposition 4.4.3. *Se $X_i \sim \mathrm{N}\left(\mu_i, \sigma_i^2\right)$, allora:*

$$\sum_{i=1}^{n} a_i X_i \sim \mathrm{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

Remark 177. La famiglia delle vc normali è chiusa rispetto ad ogni combinazione lineare: in particolare la combinazione lineare di vc normali e indipendenti è ancora una vc normale che ha per valore medio la combinazione lineare dei valori medi e per varianza la combinazione lineare delle varianze con i quadrati dei coefficienti (proprietà riproduttiva della vc normale).

Example 4.4.1 (Esame vecchio viroli)**.** A random variable X is distributed according to $\mathrm{N}(0,2)$ where 2 is the variance. What is the distribution of $Y = 2X$? Il risultato è $Y \sim \mathrm{N}(0,8)$ (come confermato dal Bigo).

Example 4.4.2 (Esame vecchio viroli)**.** A random variable X is distributed according to $\mathrm{N}(-1,1)$. What is the distribution of $Y = -2X+1$.
Correct answer is $Y \sim \mathrm{N}(3,4)$

Example 4.4.3 (Esame vecchio viroli)**.** Let $X \sim \mathrm{N}(0,2)$ and $Y \sim \mathrm{N}(1,1)$ be independent random variables where the parameters in the bracket are the expectation and the variace. What is the distribution of $Z = 2X + Y$

1. $Z \sim \mathrm{N}(1,9)$
2. $Z \sim \mathrm{N}(1,5)$
3. not possible to determine
4. $Z \sim \mathrm{N}(1,2)$

should be the first

4.5 Gamma

Remark 178. Viene utilizzata quando si deve verificare la lunghezza dell'intervallo di tempo fino all'istante in cui si verifica la n-esima manifestazione di un evento aleatorio di interesse.
Similmente alla Beta è chiamata cosi perché coinvolge l'omonima funzione matematica.

Remark 179 (Supporto e spazio parametrico)*.*

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \{n, \lambda \in \mathbb{R} : n, \lambda > 0\}$$

Definition 4.5.1 (Funzione di densità)**.** Una vc continua X si distribuisce come una Gamma con parametri $n > 0, \lambda > 0$, indicata con $X \sim \mathrm{Gamma}(n, \lambda)$, se presenta una funzione di densità come la:

$$f_X(x) = \frac{\lambda^n}{\Gamma(n)} \cdot x^{n-1} \exp(-\lambda x) \cdot \mathbb{1}_{R_X}(x) \tag{4.29}$$

Definition 4.5.2 (Funzione Gamma)**.** È definita come

$$\Gamma(n) = \int_0^{+\infty} x^{n-1} e^{-x} \, \mathrm{d}x \tag{4.30}$$

e presenta le seguenti proprietà: se $n \in \mathbb{R}, n > 1$, $\Gamma(n) = (n-1)\Gamma(n-1)$ (ossia è ricorsiva); se $n \in \mathbb{N} \setminus \{0\}$, $\Gamma(n) = (n-1)!$; ha valore notevole $\Gamma(1/2) = \sqrt{\pi}$.

Remark 180 (Funzione di ripartizione)*.* Non si può definire una funzione di ripartizione perché questa dipende dalla funzione Γ (a meno che n sia intero).

Proposition 4.5.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{n}{\lambda} \tag{4.31}$$

$$\mathrm{Var}[X] = \frac{n}{\lambda^2} \tag{4.32}$$

$$\mathrm{Asym}(X) = \frac{2}{\sqrt{n}} \tag{4.33}$$

$$\mathrm{Kurt}(X) = 3 + \frac{6}{n} \tag{4.34}$$

Remark 181 (Forma della distribuzione). λ è un parametro di scala mentre n determina la forma della distribuzione. All'aumentare del parametro λ la distribuzione si concentra sui valori più piccoli. Quando $n \to \infty$ la distribuzione diviene simmetrica e di forma campanulare (curtosi pari a 3). (figura 4.5)

```
params1 <- list(c('n' = 1, 'lambda' = 1),
                c('n' = 2, 'lambda' = 1),
                c('n' = 3, 'lambda' = 1))

params2 <- list(c('n' = 2, 'lambda' = 1),
                c('n' = 2, 'lambda' = 2),
                c('n' = 2, 'lambda' = 3))

par(mfrow = c(1,2))
tmp <- Map(function(p, add, col) {
    ## browser()
    plot_fun(function(x) dgamma(x, shape = p["n"], rate = p['lambda']),
             from = 0, to = 6,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 1),
             ylab = 'Density', las = 1, main = 'Gamma(n, 1)')
}, params1, as.list(c(F, T, T)), as.list(1:3))
leg <- unlist(lapply(params1, function(x)
    sprintf('n = %d', x['n'])))
legend('topright', legend = leg,  col = 1:3, lty = 'solid' )

tmp <- Map(function(p, add, col) {
    ## browser()
    plot_fun(function(x) dgamma(x, shape = p["n"], rate = p['lambda']),
             from = 0, to = 6,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 1.2),
             ylab = 'Density', las = 1, main = 'Gamma(2, lambda)')
}, params2, as.list(c(F, T, T)), as.list(1:3))
leg <- unlist(lapply(params2, function(x)
    sprintf('lambda = %d', x['lambda'])))
legend('topright', legend = leg,  col = 1:3, lty = 'solid' )
```

Remark 182. Si nota che se $n = 1$, la distribuzione gamma diviene una esponenziale, ovvero $\text{Gamma}(1, \lambda) \sim \text{Exp}(\lambda)$; pertanto la gamma è una generalizzazione della esponenziale.

Remark 183. Altro caso particolare, se $n = \frac{\nu}{2}$ (con $\nu \in \mathbb{N} \setminus \{0\}$, numero dei gradi di libertà) e $\lambda = \frac{1}{2}$ la distribuzione Gamma coincide con la Chi-quadrato.

Proposition 4.5.2. *La gamma gode della proprietà riproduttiva nel senso che la somma di gamma indipendenti ancora una gamma:*

$$\sum \text{Gamma}(n_i, \lambda) \sim \text{Gamma}\left(\sum_i n_i, \lambda\right) \tag{4.35}$$

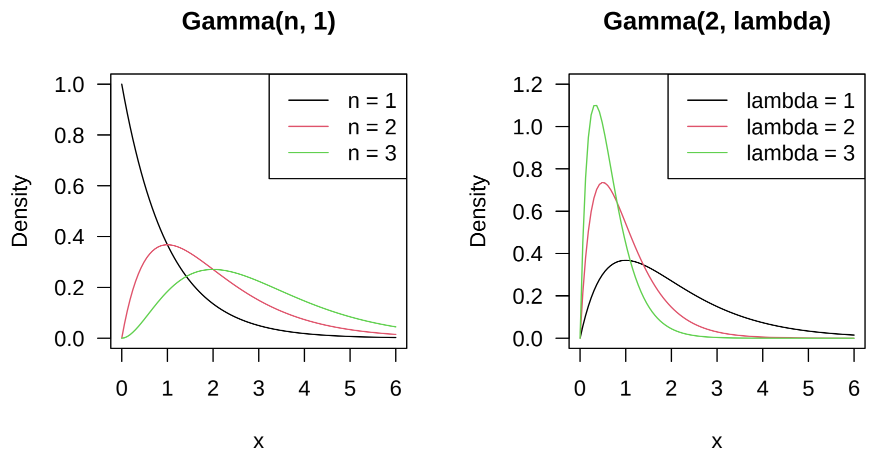


Figure 4.5: Distribuzione gamma

4.6 Chi-quadrato

Remark 184. La somma di ν vc normali standardizzate indipendenti ed elevate al quadrato è una vc continua sul supporto $(0, +\infty)$ che si distribuisce come una vc Chi-quadrato con ν gradi di libertà

$$\sum_{i=1}^{\nu} Z_i^2 \sim \chi^2_\nu \tag{4.36}$$

Remark 185 (Supporto e spazio parametrico).

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \{\nu \in \mathbb{N} \setminus \{0\}\}$$

Definition 4.6.1 (Funzione di densità)**.**

$$f_X(x) = \frac{1}{2^{\left(\frac{\nu}{2}\right)}\Gamma\left(\frac{\nu}{2}\right)} x^{\left(\frac{\nu}{2}-1\right)} e^{\left(-\frac{x}{2}\right)} \cdot \mathbb{1}_{R_X}(x) \tag{4.37}$$

con $x > 0$

Remark 186. Anche se ν può esser qualsiasi numero reale positivo, in pratica le applicazioni hanno tipicamente ν intero positivo.

Remark 187 (Forma della distribuzione). La vc Chi-quadrato è asimmetrica positiva e, al crescere di $\nu \to \infty$, tende ad assumere una forma sempre più vicina alla Normale. La forma della funzione di densità è monotona decrescente a zero se $\nu \leq 2$; se $\nu > 2$, presenta un picco intermedio in corrispondenza della moda (pari a $\nu - 2$). (figura 4.6)

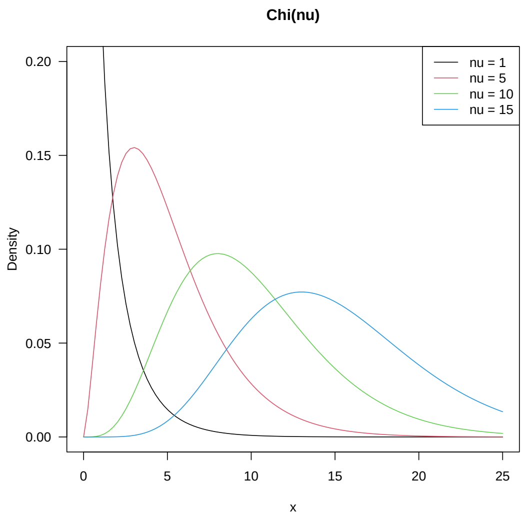


Figure 4.6: Distribuzione χ^2

```
nu <- c(1, 5, 10, 15)
tmp <- Map(function(p, add, col) {
    ## browser()
    plot_fun(function(x) dchisq(x, df = p),
             from = 0, to = 25,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 0.2),
             ylab = 'Density', las = 1, main = 'Chi(nu)')
}, nu, as.list(c(F, T, T, T)), as.list(1:4))
leg <- unlist(lapply(nu, function(x) sprintf('nu = %d', x)))
legend('topright', legend = leg,  col = 1:4, lty = 'solid')
```

Proposition 4.6.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \nu \tag{4.38}$$

$$\text{Var}[X] = 2\nu \tag{4.39}$$

$$\text{Asym}(X) = \sqrt{\frac{8}{\nu}} \tag{4.40}$$

$$\text{Kurt}(X) = 3 + \frac{12}{\nu} \tag{4.41}$$

Proposition 4.6.2. *Anche la distribuzione Chi-quadrato gode della proprietà riproduttiva:*

$$\sum_{i=1}^{n} \chi^2_{\nu_i} \sim \chi^2_{\sum_i \nu_i}$$

4.7 Beta

Remark 188. Viene utilizzata quando si vogliono definire a priori i valori possibili delle probabilità di successo per variabili Bernoulliane.

Remark 189 (Supporto e spazio parametrico).

$$R_X = [0, 1]$$
$$\boldsymbol{\Theta} = \{\alpha, \beta \in \mathbb{R} : \alpha, \beta > 0\}$$

Definition 4.7.1 (Funzione di densità)**.** Una vc continua X si definisce Beta con due parametri $\alpha > 0, \beta > 0$, e la indichiamo con $X \sim \text{Beta}(\alpha, \beta)$ se la sua funzione di densità è:

$$f_X(x, \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \cdot \mathbb{1}_{R_X}(x) \tag{4.42}$$

Definition 4.7.2 (Funzione Beta)**.** Definita come

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \tag{4.43}$$

Presenta le seguenti proprietà

$$B(\alpha, \beta) = B(\beta, \alpha)$$
$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$$

Remark 190. Una vc Beta è definita nell'intervallo $[0, 1]$, ma effettuando la trasformazione $Y = X(b-a) + a$, la si può ricondurre all'intervallo $[a, b]$.

Proposition 4.7.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \tag{4.44}$$

$$\text{Var}[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{4.45}$$

Remark 191 (Forma della distribuzione). La forma (figura 4.7) dipende dai parametri α, β:

- se $\alpha = \beta$ la distribuzione è simmetrica rispetto al valore centrale $x = 1/2$; nel caso particolare $\alpha = \beta = 1$, la distribuzione coincide con l'uniforme: $\text{Beta}(1, 1) \sim \text{Unif}(0, 1)$;
- altrimenti il segno di $\beta - \alpha$ denota l'asimmetria (es se negativo, perché $\alpha > \beta$, la coda è a sinistra, se positivo la coda a destra); scambiando α con β si inverte l'asse di simmetria.

```
p <- c(0.3, 0.7, 1, 4)
alphas <- p
betas <- p
```

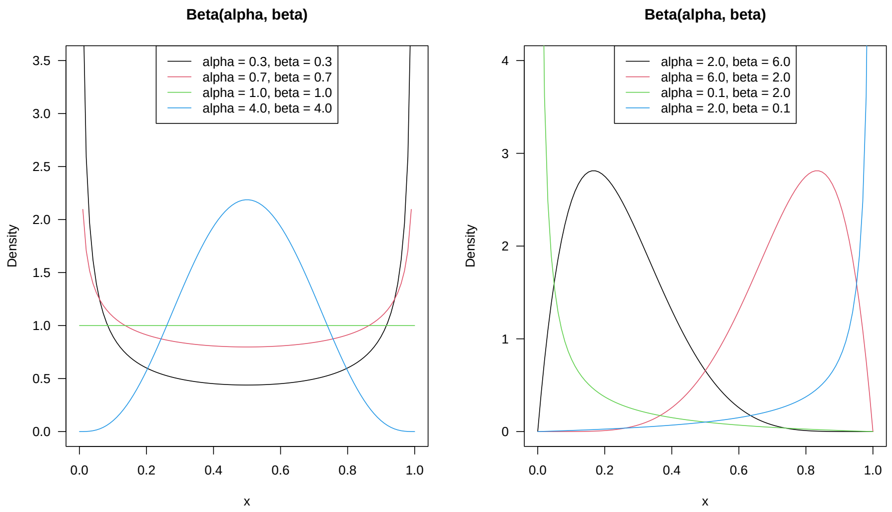


Figure 4.7: Distribuzione beta

```
par(mfrow = c(1,2))
tmp <- Map(function(a, b, add, col) {
    ## browser()
    plot_fun(function(x) dbeta(x, shape1 = a, shape2 = b),
             from = 0, to = 1,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 3.5),
             ylab = 'Density', las = 1, main = 'Beta(alpha, beta)')
}, as.list(alphas), as.list(betas), as.list(c(F, T, T, T)), as.list(1:4))
leg <- unlist(lapply(p, function(x) sprintf('alpha = %.1f, beta = %.1f', x, x)))
legend('top', legend = leg,  col = 1:4, lty = 'solid')

alphas <- c(2, 6, 0.1, 2)
betas  <- c(6, 2, 2  , 0.1)
tmp <- Map(function(a, b, add, col) {
    ## browser()
    plot_fun(function(x) dbeta(x, shape1 = a, shape2 = b),
             from = 0, to = 1,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 4),
             ylab = 'Density', las = 1, main = 'Beta(alpha, beta)')
}, as.list(alphas), as.list(betas), as.list(c(F, T, T,T)), as.list(1:4))
leg <- unlist(Map(function(a, b) sprintf('alpha = %.1f, beta = %.1f', a, b),
                  as.list(alphas), as.list(betas)))
legend('top', legend = leg,  col = 1:4, lty = 'solid')
```

4.8 T di Student

Remark 192. Il suo uso è prettamente teorico, in quanto è la risultante di una trasformazione su due variabili, una normale e una chi quadrato.

Remark 193 (Supporto e spazio parametrico).

$$
\begin{aligned}
R_X &= \mathbb{R} \\
\mathbf{\Theta} &= \{g \in \mathbb{N} \setminus \{0\}\}
\end{aligned}
$$

Definition 4.8.1 (Distribuzione T)**.** Se $Z \sim \mathrm{N}(0,1)$ ed C è una distribuzione indipendente tale che $C \sim \chi_g^2$ allora si definisce vc di Student la seguente X:

$$
X = \frac{Z}{\sqrt{C/g}} \sim \mathrm{T}(g) \tag{4.46}
$$

Definition 4.8.2 (Funzione di densità)**.**

$$
f_X(x) = \frac{\Gamma\left(\frac{g+1}{2}\right)}{\Gamma\left(\frac{g}{2}\right)\sqrt{\pi g}} \left(1 + \frac{x^2}{g}\right)^{-\frac{g+1}{2}} \cdot \mathbb{1}_{R_X}(x) \tag{4.47}
$$

Proposition 4.8.1 (Momenti caratteristici)**.**

$$
\begin{aligned}
\mathbb{E}[X] &= 0 \quad \textit{se } g > 1 \\
\mathrm{Var}[X] &= \frac{g}{g-2} \quad \textit{seg} > 2 \\
\mathrm{Kurt}(X) &= 3 + \frac{6}{g-4} \quad \textit{se } g > 4
\end{aligned}
$$

Remark 194 (Forma della distribuzione). Per $g \to \infty$ si nota la convergenza alla normale standardizzata. Verso $g = 30$, l'approssimazione è già buona; per g via via inferiore permane qualche differenza (code più alte rispetto alla normale, moda e media più basse). (figura 4.8)

```
g <- c(1, 10, 40, NA)
tmp <- Map(function(g, add, col) {
    plot_fun(function(x) if (!is.na(g))  dt(x, df = g)
                         else dnorm(x),
             from = -4, to = 4,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 0.4),
             ylab = 'Density', las = 1, main = 'T(g)')
}, as.list(g), as.list(c(FALSE, TRUE, TRUE, TRUE)), as.list(1:4))
leg <- unlist(lapply(g, function(x)
    if (!is.na(x)) sprintf('T(%d)', x)
    else 'N(0, 1)'))
legend('topright', legend = leg,  col = 1:4, lty = 'solid')
```

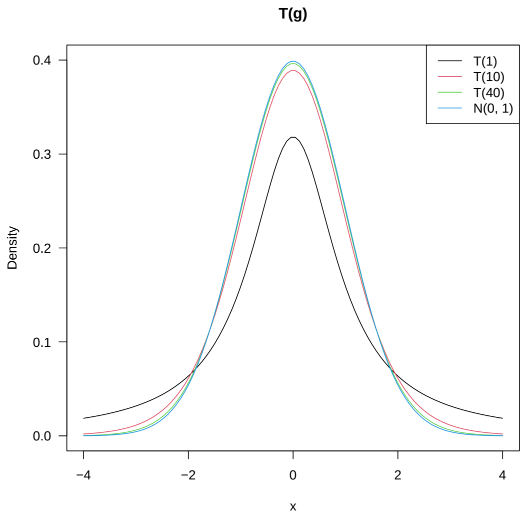


Figure 4.8: Distribuzione t

4.9 F di Fisher

Remark 195. Il suo uso è prettamente teorico, in quanto è risultate di una trasformazione. È la distribuzione che deriva dal rapporto tra due vc Chi quadrato indipendenti tra loro e divise per i rispettivi gradi di libertà.

Remark 196. Se $X_1 \sim \chi^2_{g_1}$ e $X_2 \sim \chi^2_{g_2}$, allora

$$X = \frac{X_1/g_1}{X_2/g_2} \sim \mathrm{F}\left(g_1, g_2\right) \tag{4.48}$$

ovvero X si distribuisce come una F con g_1 e g_2 gradi di libertà.

Remark 197 (Supporto e spazio parametrico).

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \{g_1, g_2 \in \mathbb{N} \setminus \{0\}\}$$

Definition 4.9.1 (Funzione di densità)**.**

$$f_X(x) = \frac{\Gamma\left(\frac{g_1+g_2}{2}\right)}{\Gamma\left(\frac{g_1}{2}\right)\Gamma\left(\frac{g_2}{2}\right)} \cdot \left(\frac{g_1}{g_2}\right)^{\frac{g_1}{2}} \cdot \frac{x^{(g_1-2)/2}}{\left(1+\frac{g_1}{g_2}x\right)^{\frac{g_1+g_2}{2}}} \cdot \mathbb{1}_{R_X}(x) \tag{4.49}$$

Remark 198 (Funzione di ripartizione). Anche per la F non vi è una forma chiusa della ripartizione e ci si affida alle tavole.

Proposition 4.9.1 (Momenti caratteristici)**.**

$$\mathbb{E}\left[X\right] = \frac{g_2}{g_2-2} \quad \textit{se } g_2 > 2$$
$$\mathrm{Var}\left[X\right] = \frac{2g_2^2(g_1+g_2-2)}{g_1(g_2-2)^2(g_2-4)} \quad \textit{se } g_2 > 4$$

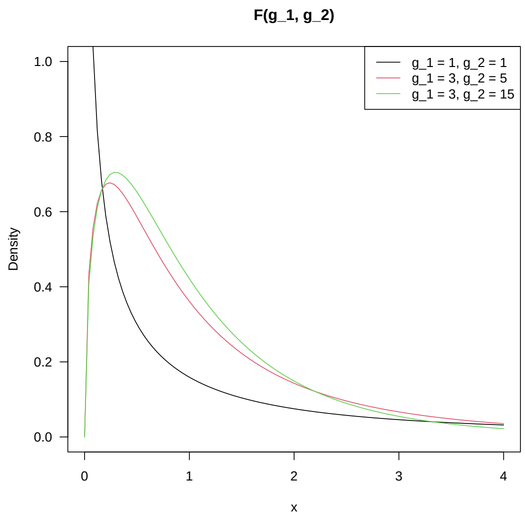


Figure 4.9: Distribuzione F

```
g1 <- c(1, 3,  3)
g2 <- c(1, 5, 15)
tmp <- Map(function(g1, g2, add, col) {
    plot_fun(function(x) df(x, df1 = g1, df2 = g2),
             from = 0, to = 4,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 1),
             ylab = 'Density', las = 1, main = 'F(g_1, g_2)')
}, as.list(g1), as.list(g2), as.list(c(F, T, T)), as.list(1:3))
leg <- unlist(Map(function(g1, g2) sprintf('g_1 = %d, g_2 = %d', g1, g2),
                  g1, g2))
legend('topright', legend = leg,  col = 1:3, lty = 'solid')
```

Remark 199 (Forma della distribuzione). Si nota che se $g_1 = g_2 = 1$ la funzione è monotona decrescente, se $g_1, g_2 \neq 1$ la funzione è asimmetrica positiva. (figura 4.9)
La distribuzione converge a quella di una normale solo se contemporaneamente $g_1 \to \infty$ e $g_2 \to \infty$.

4.10 Lognormale

Remark 200. Viene utilizzata quando la grandezza oggetto di studio è il risultato del prodotto di n fattori indipendenti

Remark 201 (Supporto e spazio parametrico).

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \left\{\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R} : \sigma^2 > 0\right\}$$

Definition 4.10.1 (Funzione di densità)**.**

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2} \cdot \mathbb{1}_{R_X}(x) \tag{4.50}$$

Proposition 4.10.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = e^{\mu + \frac{\sigma^2}{2}}$$
$$\mathrm{Var}[X] = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2}$$

Remark 202. Si ha che se $X \sim \mathrm{LogN}(mu, \sigma)$ allora $\log X \sim \mathrm{N}(\mu, \sigma^2)$, mentre se $Y \sim \mathrm{N}(\mu, \sigma^2)$, $e^Y \sim \mathrm{LogN}(\mu, \sigma^2)$

Remark 203 (Forma della distribuzione)*.* Con μ fisso all'aumentare di σ l'asimmetria si incrementa (figura 4.10)

```
mu <- rep(0, 6)
s <- c(0.125, 0.25, 0.5, 1, 1.5, 10)
tmp <- Map(function(mu, s, add, col) {
    plot_fun(function(x) dlnorm(x, meanlog = mu, sdlog = s),
             from = 0, to = 3,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 3),
             ylab = 'Density', las = 1,
             main = 'logN(mu, s)')
}, as.list(mu), as.list(s), as.list(c(F, T, T, T, T, T)), as.list(1:6))
leg <- unlist(Map(function(mu, s)
    sprintf('mu = %d, s = %.2f', mu, s), mu, s))
legend('topright', legend = leg,  col = 1:6, lty = 'solid')
```

4.11 Weibull

Remark 204. Viene utilizzata per studiare l'affidabilità dei sistemi di produzione nei processi industriali, in particolare per valutare i tassi di rottura

Remark 205. La Weibull presenta la caratteristica di avere una funzione di rischio variabile in funzione di un ulteriore parametro a: se la vc $(X/b)^a \sim \mathrm{Exp}(1)$, allora diremo che la vc continua X, definita sula semiretta positiva è una vc di Weibull con parametri $a > 0, b > 0$.

Remark 206 (Supporto e spazio parametrico)*.*

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \{a, b \in \mathbb{R} : a, b > 0\}$$

Remark 207 (Forma della funzione)*.* Il parametro a determina la forma (figura 4.11):

- se $a < 1$ il tasso di rottura è decrescente nel tempo, ci sono componenti difettose che si rompono subito e, una volta sostituito, il tasso diminuisce

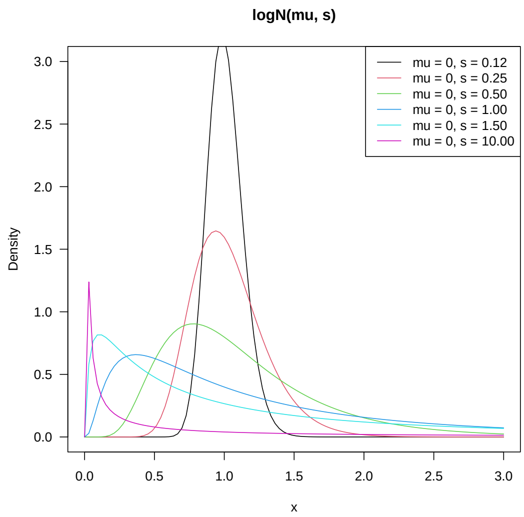


Figure 4.10: Distribuzione lognormale

- se $a = 1$ il tasso di rottura è costante nel tempo: le cause dei difetti sono casuali (e la distribuzione coincide con una esponenziale di parametro $1/b$, ossia $\text{Weibull}\,(1, b) \sim \text{Exp}\left(\frac{1}{b}\right)$)
- se $a > 1$ il tasso di rottura è crescente nel tempo, le cause della rottura dei componenti derivano dall'usura

Definition 4.11.1 (Funzione di densità).

$$f_X(x) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} e^{\left[-\left(\frac{x}{b}\right)^a\right]} \cdot \mathbb{1}_{R_X}(x) \tag{4.51}$$

Proposition 4.11.1 (Momenti caratteristici).

$$\mathbb{E}[X] = \frac{\Gamma\left(1 + \frac{1}{b}\right)}{a^{1/b}}$$
$$\text{Var}[X] = \frac{\Gamma\left(1 + \frac{2}{b}\right) - \Gamma^2\left(1 + \frac{1}{b}\right)}{a^{2/b}}$$

```r
b <- c(2, 2, 3, 4)
a <- c(0.5, 1, 1.5, 3)
tmp <- Map(function(mu, s, add, col) {
    plot_fun(function(x) dweibull(x, scale = mu, shape = s),
             from = 0, to = 5,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 2),
             ylab = 'Density', las = 1,
             main = 'Wei(a, b)')
}, as.list(a), as.list(b), as.list(c(F,T, T, T)), as.list(1:4))
leg <- unlist(Map(function(mu, s)
```

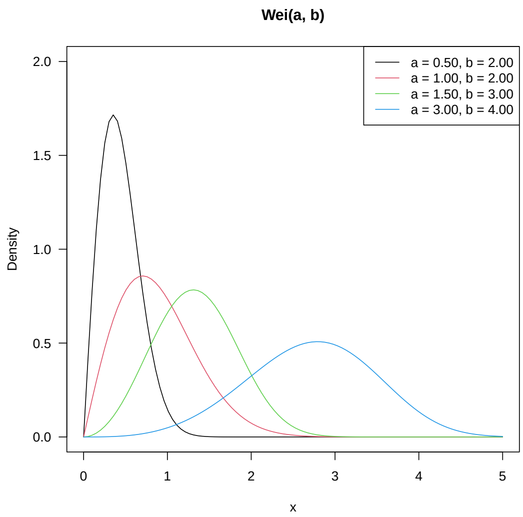


Figure 4.11: Distribuzione Weibull

```
    sprintf('a = %.2f, b = %.2f', mu, s), a, b))
legend('topright', legend = leg,  col = 1:4, lty = 'solid')
```

4.12 Pareto

Remark 208. Viene utilizzata quando si studiano distribuzioni di variabili che hanno un minimo (ad esempio come, con x_m = reddito minimo)

Remark 209 (Supporto e spazio parametrico).

$$R_X = (x_m, +\infty)$$
$$\boldsymbol{\Theta} = \{x_m, k \in \mathbb{R} : x_m, k > 0\}$$

Definition 4.12.1 (Funzione di densità)**.**

$$f_X(x) = k\frac{x_m^k}{x^{k+1}} \cdot \mathbb{1}_{R_X}(x) \tag{4.52}$$

Proposition 4.12.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{kx_m}{k-1} \qquad \textit{per } k > 1$$
$$\mathrm{Var}[X] = \left(\frac{x_m}{k-1}\right)^2 \frac{k}{k-2} \qquad \textit{per } k > 2$$

Remark 210 (Forma della distribuzione). Al crescere di k la distribuzione è disuguale, ed è molto probabile trovare valori vicini al limite inferiore x_m, poco proabile trovare valori molto grandi.

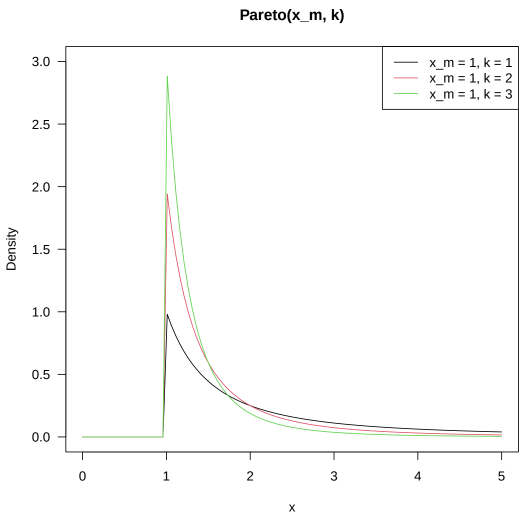


Figure 4.12: Distribuzione di Pareto

```
mu <- 1
k <- 1:3
tmp <- Map(function(mu, s, add, col) {
    plot_fun(function(x) VGAM::dpareto(x, scale = mu, shape = s),
             from = 0, to = 5,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 3),
             ylab = 'Density', las = 1,
             main = 'Pareto(x_m, k)')
}, as.list(mu), as.list(k), as.list(c(F, T, T)), as.list(1:3))
leg <- unlist(Map(function(mu, s)
    sprintf('x_m = %d, k = %d', mu, s), mu, k))
legend('topright', legend = leg,  col = 1:3, lty = 'solid')
```

Chapter 5

Misc topics

5.1 Quantili

Definition 5.1.1. Sia X una va univariata e sia $\alpha \in (0,1)$. Un numero $b \in \mathbb{R}$ si rice quantile di ordine α di X se:

NB: Rigo, dalla triennale

$$\mathbb{P}(X \leq b) \geq \alpha \geq \mathbb{P}(X < b)$$

Equivalentemente. detta F la funzione di ripartizione di X

$$F(b) \geq \alpha \geq F(b^-)$$

Dove $F(b^-) = \lim_{x \to b^-} F(x)$

Remark 211. Se

- F è continua, la condizione precedente diviene $F(b) = \alpha$
- F è continua e strettamente crscente $\forall \alpha \in (0,1)$ esiste uno e un solo quantile di ordine α, ovvero $\exists! b \in \mathbb{R}$ tale che $F(b) = \alpha$. In particolare, ciò è vero se X è assolutamente continua con densità strettamente positiva

Remark 212. Se

- $\alpha = 1/2$ un quantile di ordine $\frac{1}{2}$ si chiama mediana
- $\alpha = 1/4$ un quantile di ordine $\frac{1}{4}$ si chiama primo quartile
- $\alpha = 3/4$ un quantile di ordine $\frac{3}{4}$ si chiama terzo quartile

Remark 213. In ogni caso, al di la della definizione formale, detto in parole povere, un quantile di ordine α è un qualsiasi valore b che lasci alla propria sinistra un'area (della densità) pari a α. Questo si vede bene se X è assolutamente continua con densità > 0

Example 5.1.1. Sia $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ e sia b il quantile di ordine α, allora

$$\alpha = \mathbb{P}(X \leq b) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right)$$

Dove Φ è la funzione di ripartizione della $\mathrm{N}(0,1)$. Due commenti:

- se b è quantile di ordine α per una $\mathrm{N}\left(\mu, \sigma^2\right)$ allora $\frac{b-\mu}{\sigma}$ è quantile di ordine α per una $\mathrm{N}(0,1)$. Naturalmente vale anche il viceversa, ovvero se c è quantile di ordine α per una $\mathrm{N}(0,1)$, allora $\mu+\sigma c$ è quantile di ordine α per una $\mathrm{N}\left(\mu, \sigma^2\right)$. Infatti se $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$

$$\mathbb{P}(X \leq \mu+\sigma c)=\mathbb{P}\left(\frac{X-\mu}{\sigma} \leq c\right)=\Phi(c)=\alpha$$

- in queste considerazioni abbiamo usato il fatto che

$$X \sim \mathrm{N}\left(\mu, \sigma^2\right) \implies \alpha+\beta X \sim \mathrm{N}(\cdot, \cdot), \quad forall \alpha in \mathbb{R}, \forall \beta \neq 0$$

Ovvero ad eccezione del caso banale $\beta=0$, una qualsiasi trasformazione lineare di X è ancora normale.

5.2 Order statistics

Remark 214. Together with *rank* statistics, *order* statistics are fundamental tools in non-parametric statistics and inference.

Definition 5.2.1 (Order statistics)**.** Let $X=(X_1, \ldots, X_n)^\top$ be any n-variate random variable. The corresponding order statistics are the element of the vector $Y=(X_{(1)}, \ldots, X_{(n)})^\top$ where $X_{(1)} \leq \ldots \leq X_{(n)}$ are the elements of X arranged in non decreasing order, that is the following random variables

$$\begin{aligned}
X_{(1)} &= \min\{X_1, \ldots, X_n\} \\
X_{(2)} &= \min\left\{\{X_1, \ldots, X_n\} \setminus \{X_{(1)}\}\right\} \\
&\ldots \\
X_{(n)} &= \max\{X_1, \ldots, X_n\}
\end{aligned}$$

Example 5.2.1. Se $n=4$, $X_1=2, X_2=0, X_3=5, X_4=1$ allora $X_{(1)}=-1, X_{(2)}=0, X_{(3)}=2, X_{(4)}=5$.
Con $n=3$, $X_1=1, X_2=-2, X_3=1$, $X_{(1)}=-2, X_{(2)}=X_{(3)}=1$.

Remark 215. La terminologia deriva dalla statistica: basta pensare ad X come a un campione e ad Y come al campione ordinato.

Remark 216. Here the random vector can be conceptualized as measurement on the same variable for different units (not several measurement within one unit).

Definition 5.2.2 (k-th order statistic)**.** The k-th order statistic of the sample is equal to its k-th smallest value.

Example 5.2.2 (Minimum and maximum)**.** Important special cases of the order statistics are the *minimum* $X_{(1)}$, the *maximum* $X_{(n)}$, the sample *median* and other sample *quantiles*.

Example 5.2.3. Throwing a dice 6 times, having the sequence $X_1, \ldots, X_6$. To study the distribution of the minimum $X_{(1)}$, we can say that

$$\begin{aligned}
\mathbb{P}\left(X_{(1)}=6\right) &= \frac{1}{6} \cdot \ldots \cdot \frac{1}{6}=\left(\frac{1}{6}\right)^6 \\
\mathbb{P}\left(X_{(1)}=1\right) &= 1-\left(\frac{5}{6}\right)^6
\end{aligned}$$

Important remark 35 (Our focus). We:

- are interested in studying distribution/properties of these newly defined random variables, or in general, given the distribution of X, find the distribution of Y (note this another example of the general problem of transforming variable)
- deal with the simplest case where $X_1, \dots, X_n$ are iid

5.2.1 Minimum

Proposition 5.2.1 (Distribution function). *We have that*

$$F_{(1)}(x) = 1 - [1 - F_X(x)]^n \tag{5.1}$$

NB: Direi sia roba di Viroli, Rigo l'ha fatto come caso particolare di $X_{(i)}$

Proof.

$$\begin{aligned}
F_{(1)}(x) &= \mathbb{P}\left(X_{(1)} \leq x\right) = 1 - \mathbb{P}\left(X_{(1)} > x\right) \\
&= 1 - \mathbb{P}\left(X_1 > x, X_2 > x, \dots, X_n > x\right) \overset{(1)}{=} 1 - \prod_{i=1} \mathbb{P}\left(X_i > x\right) \\
&\overset{(2)}{=} 1 - \prod_{i=1} \mathbb{P}\left(X > x\right) = 1 - \left[\mathbb{P}\left(X > x\right)\right]^n = 1 - \left[1 - \mathbb{P}\left(X \leq x\right)\right]^n \\
&= 1 - [1 - F_X(x)]^n
\end{aligned}$$

with (1) we considered independent rvs and (2) identically distributed. □

Remark 217. Interpretazione affinché il minimo sia al più x si fa il complemento in cui si guarda la probabilità che siano tutte contemporaneamente $> x$

Proposition 5.2.2 (Density function).

$$f_{(1)}(x) = n f_X(x) \cdot [1 - F_X(x)]^{n-1}$$

Proof.

$$\begin{aligned}
f_1(x) &= \frac{\partial F_{(1)}(x)}{\partial x} = -n\left[1 - F_X(x)\right]^{n-1}\left(-f_X(x)\right) \\
&= n f_X(x) \cdot [1 - F_X(x)]^{n-1}
\end{aligned}$$

□

Example 5.2.4. A room is lit by 5 light bulbs, each bulb lifetime has a distribution $X \sim \operatorname{Exp}\left(\lambda = \frac{1}{100}\right)$. What is the probability that after 200 days *all the bulbs are still working*?
We can setup this as $\mathbb{P}\left(X_{(1)} > 200\right)$, therefore:

$$\mathbb{P}\left(X_{(1)} > 200\right) = 1 - \mathbb{P}\left(X_{(1)} \leq 200\right) = 1 - F_{(1)}(200)$$

we have that, being X an exponential

$$F_{(1)}(200) = 1 - (1 - F_X(200))^5 = 1 - \left(1 - 1 + e^{-200/100}\right)^5 = 1 - \frac{1}{e^{10}}$$

Therefore

$$\mathbb{P}\left(X_{(1)} > 200\right) = 1 - 1 + \frac{1}{e^{10}} = \frac{1}{e^{10}}$$

Example 5.2.5 (Virols eserciziario 1, es 6)**.** Let $X_1, \dots, X_n$ be a random sample from a Weibull (α, β) distribution, that is

$$f(x) = \alpha\beta x^{\beta 1} e^{-\alpha x^\beta}, \quad x > 0, \alpha, \beta > 0$$

Derive the probability density function of $X_{(1)}$ and recognize it.
The distribution function of a Weibull rv is

$$F_X(x) = 1 - e^{-\alpha x^\beta}$$

therefore

$$F_{(1)}(x) = 1 - [1 - F_X(x)]^n = 1 - \left[e^{-\alpha x^\beta}\right]^n = 1 - e^{-\alpha n x^\beta}$$

which is a weibull with parameters $n\alpha$ and β

5.2.2 Maximum

NB: Direi sia roba di Viroli, Rigo l'ha fatto come caso particolare di $X_{(i)}$

Proposition 5.2.3 (Distribution function)**.**

$$F_{(n)}(x) = [F_X(x)]^n \tag{5.2}$$

Proof.

$$\begin{aligned} F_{(n)}(x) &= \mathbb{P}\left(X_{(n)} \le x\right) = \mathbb{P}\left(X_1 \le x, \dots, X_N \le x\right) \\ &\overset{(iid)}{=} [\mathbb{P}(X \le x)]^n = [F_X(x)]^n \end{aligned}$$

□

Remark 218. Il massimo sia $\le x$ se tutte le vc sono $\le x$

Proposition 5.2.4 (Density function)**.**

$$f_{(n)}(x) = n [F_X(x)]^{n-1} f_X(x) \tag{5.3}$$

Proof.

$$f_{(n)}(x) = \frac{\partial}{\partial x} F_{(n)}(x) = n [F_X(x)]^{n-1} f_X(x)$$

□

Example 5.2.6. Considering again a room lit by 5 light bulbs, each bulb lifetime has a distribution $X \sim \text{Exp}\left(\lambda = \frac{1}{100}\right)$. What is the probability that after 200 days *at least a bulb will be working*?
This can be setup with

$$\begin{aligned} \mathbb{P}\left(X_{(n)} > 200\right) &= 1 - \mathbb{P}\left(X_{(n)} \le 200\right) = 1 - F_{(n)}(200) \\ &= 1 - [F_X(200)]^5 = 1 - \left(1 - e^{-2}\right)^5 \simeq 0.52 \end{aligned}$$

Example 5.2.7. Draw randomly 12 numbers between from $X \sim \text{Unif}(0, 1)$. What is the probability that at least a number > 0.9?
If $X \sim \text{Unif}(0, 1)$, $F_X(x) = x$. We have

$$\mathbb{P}\left(X_{(n)} > 0.9\right) = 1 - \mathbb{P}\left(X_{(n)} \le 0.9\right) = 1 - [F_X(0.9)]^{12} = 1 - 0.9^{12} = 0.718$$

Example 5.2.8 (Esame vecchio viroli)**.** A random variable X has density function

$$f(x, \theta) = \frac{3x^2}{\theta^3}$$

with $X \in [0, \theta]$. Compute the cumulative distribution function of the maximum $X_{(n)}$.
Per ottenerla occorre sviluppare la cumulata della funzione di partenza

$$F_X(x) = \int \frac{3x^2}{\theta^3} = \frac{3}{\theta^3} \int x^2 = \frac{3}{\theta} \frac{x^3}{3} = \frac{x^3}{\theta^3}$$

Da cui

$$F_{X_{(i)}}(x) = [F_X(x)]^n = \left(\frac{x}{\theta}\right)^{3n}$$

come confermato da taluni

Example 5.2.9 (Esame vecchio viroli)**.** A random variable X has density function

$$f(x, \theta) = \frac{2x}{\theta^2}$$

with $X \in [0, \theta]$. Compute the probability distribution function of the maximum $X_{(n)}$

1. $F_n(x) = \frac{x^{2n}}{\theta^n}$
2. $F_n(x) = \frac{x^{n-1}}{\theta^n}$
3. $F_n(x) = \frac{x^{3n-1}}{\theta^{3n}}$
4. $F_n(x) = \frac{x^{2n}}{\theta^{2n}}$; taluni suggeriscono questa

Analogamente

$$F_X(x) = \int \frac{2x^2}{\theta^2} = \frac{2}{\theta^2} \int x^2 = \frac{2}{\theta} \frac{x^2}{2} = \frac{x^2}{\theta^2}$$

da cui

$$F_{X_{(i)}}(x) = [F_X(x)]^n = \left(\frac{x}{\theta}\right)^{2n}$$

5.2.3 Generalized $X_{(i)}$

Important remark 36. If we write $X_{(i)} \sim F_{(i)}(x)$, with $i = 1, \ldots, n$ we mean that $X_{(i)}$ is distributed following the i-th ordered statistic.

Proposition 5.2.5 (Distribution function of i-th ordered statistics)**.** *We have*

$$F_{(i)}(x) = \mathbb{P}\left(X_{(i)} \leq x\right) = \sum_{j=i}^{n} \binom{n}{j} F_X(x)^j \cdot (1 - F_X(x))^{n-j} \tag{5.4}$$

Proof. To find the distribution function of $X_{(i)}$, it is convenient to think that a success occurs at trial i if $X_i \leq x$ (here is just comparing the unsorted sequence of realization with a treshold x of interest). One obtains

$$\begin{aligned}
F_{(i)}(x) &= \mathbb{P}\left(X_{(i)} \leq x\right) \\
&\overset{(1)}{=} \mathbb{P}\left(\text{at } \textit{least } i \text{ successes/observation below } x\right) \\
&= \sum_{j=i}^{n} \mathbb{P}\left(\text{exactly } j \text{ successes occur}\right) \\
&\overset{(2)}{=} \sum_{j=i}^{n} \binom{n}{j} p^j (1-p)^{n-j} \\
&= \sum_{j=1}^{n} \binom{n}{j} F(x)^j (1-F(x))^{n-j}
\end{aligned}$$

where in

- (1) to have the i-th ordered observation under a certain treshold x means that we need *at least* i observations under that treshold (could be more as well, no problem, we're just focusing on the first i);
- in (2) where p is the probability of a success in a single trial, $\mathbb{P}(X \leq x)$ and it coincides with the distribution function F common to $X_1, \ldots, X_n$, that is $p = F(x)$

□

Example 5.2.10. Imagine $n = 3$ with $x_{(1)} = 3$, $x_{(2)} = 5$, $x_{(3)} = 7$. We have that $\mathbb{P}\left(X_{(2)} \leq x\right)$ is the probability that 2 rvs are $\leq x$ *OR* the probability that 3 random variables are $\leq x$.

Example 5.2.11 (Maximum)**.** As a sepecial case, for $i = n$, one obtains

$$\mathbb{P}\left(X_{(n)} \leq x\right) = \binom{n}{n} F(x)^n (1-F(x))^{n-n} = F(x)^n$$

The above result may be also obtained arguing as follows

$$\mathbb{P}\left(X_{(n)} \leq x\right) = \mathbb{P}\left(X_i \leq x, \forall i\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i \leq x\right) = \mathbb{P}\left(X_1 \leq x\right)^n = F(x)^n$$

Example 5.2.12 (Minimum)**.** For $i = 1$

$$\begin{aligned}
\mathbb{P}\left(X_{(1)} \leq x\right) &= \sum_{j=1}^{n} \binom{n}{j} F(x)^j (1-F(x))^{n-j} \\
&= \left[\sum_{j=0}^{n} \binom{n}{j} F(x)^j (1-F(x))^{n-j}\right] - \binom{n}{0} F(x)^0 (1-F(x))^{n-0} \\
&= (F(x) + 1 - F(x))^n - (1-F(x))^n \\
&= 1 - (1-F(x))^n
\end{aligned}$$

Once again this result can also be obtained as follows

$$\mathbb{P}\left(X_{(1)} \leq x\right) = 1 - \mathbb{P}\left(X_{(1)} > x\right) = 1 - \mathbb{P}\left(X_i > x, \forall i\right) = 1 - \prod_{i=1}^{n} \mathbb{P}\left(X_i > x\right)$$
$$= 1 - \mathbb{P}\left(X_1 > x\right)^n = 1 - [1 - F(x)]^n$$

Remark 219. Next we want more: we want the distribution of the ordered vector $Y = \left\{X_{(1)}, \ldots, X_{(n)}\right\}$.
To this end, it is convenient to make a further assumption: not only the element of X are iid, but their common distribution is absolutely continuous

Theorem 5.2.6. *If $X_1, \ldots, X_n$ are iid and their common distribution is absolutely continuous, then Y is still absolutely continuous and the joint density of Y is*

$$g(x_1, x_2, \ldots, x_n) = \begin{cases} n! \prod_{i=1}^{n} f(x_i) & \textit{if } x_1 < x_2 < \ldots < x_n \\ 0 & \textit{otherwise} \end{cases}$$

where f denotes the density common to $X_1, \ldots X_n$

Remark 220. So if we have a random vector composed by iid absolutely continuous random variables, the vector of order statistics is still absolutely continuous and the joint density is described above.
Intuitively the productory of f is due to the orginal vector components (iid), then we have $n!$ permutations to produce the same arrangement.

Example 5.2.13. For instance if $n = 2$ then $\mathbb{P}\left(X_1 = X_2\right) = 0$ since X_1, X_2 are absolutely continuous and $\mathbb{P}\left(X_{(1)} < X_{(2)}\right) = 1$.
The density g of $(X_{(1)}, X_{(2)})^\top$ is null on the part under main bisector, ($\left\{(x, y) \in \mathbb{R}^2 : x \geq y\right\}$, we have that the higher ordered element is y while lowest is x).
On the set $\left\{(x, y) \in \mathbb{R}^2 : y > x\right\}$ (above bisettrice) we have that the density is given by

$$g(x, y) = 2f(x)f(y)$$

Example 5.2.14. Let X_1 and X_2 be iid with $X_1 \sim \text{Unif}\,(0, 1)$. Then $Y = (X_{(1)}, X_{(2)})^\top$ is absolutely continuous with density

$$g(x, y) = \begin{cases} 2! f(x) f(y) & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}$$

Since

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

one finally obtains

$$g(x, y) = \begin{cases} 2 \cdot 1 \cdot 1 = 2 & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Intuition (area below have to sum always to 1 so):

- the density of $X = (X_1, X_2)^\top$ is 1 on the square between $(0,0)$ and $(1,1)$;
- the density of $X = (X_{(1)}, X_{(2)})^\top$ is 2 on half of the above square, that is on triangle between $(0,0)$, $(1,1)$ and $(1,0)$

Proposition 5.2.7 (Density function for i-th order statistic)**.**

NB: Da qui in poi della Virols direi

$$f_{(i)}(x) = \mathbb{P}\left(X_{(i)} = x\right) = \binom{n}{i} \cdot i \cdot F_X(x)^{i-1} \cdot f_X(x)(1 - F_X(x))^{n-i} \tag{5.5}$$

Important remark 37. Eg when $i = 1$ we obtain the formula for minimum

$$\begin{aligned} f_{(i)}(x) &= \binom{n}{1} 1 F_X(x)^0 \cdot f_X(x)(1 - F_X(x))^{n-1} \\ &= n f_X(x) \cdot [1 - F_X(x)]^{n-1} \end{aligned}$$

while for $i = n$ the maximum

$$f_{(n)}(x) = \binom{n}{n} n F_X(x)^{n-1} \cdot f_X(x)(1 - F_X(x))^0 = n\left[F_X(x)\right]^{n-1} f_X(x)$$

Example 5.2.15. Let $X_1, \ldots, X_n \sim \text{Unif}\,(0,1)$ be n iid uniforms, therefore having

$$f_X(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1 \\ 0 & \text{elsewhere} \end{cases}, \quad F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 < x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

The k-th ordered statistic is distributed as a beta. Let's see it:

$$f_{(k)}(x) = k \binom{n}{k} x^{k-1} (1-x)^{n-k}$$

Now we have that

$$k\binom{n}{k} = \frac{n!}{(k-1)!(n-k)!} = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} = \frac{1}{B(k, n-k+1)}$$

Therefore

$$f_{(k)}(x) = \frac{1}{B(k, n-k+1)} x^{k-1}(1-x)^{n-k}$$

or $X_{(k)} \sim \text{Beta}\,(k, n-k+1)$. As special cases

$$\begin{aligned} X_{(1)} &\sim \text{Beta}\,(1, n) \\ X_{(n)} &\sim \text{Beta}\,(n, 1) \end{aligned}$$

5.3 Inequalities

5.3.1 Tchebychev (Rigo)

Remark 221. One reason for Tchebychev inequality is so useful is that it holds for any rv X without any further assumption.
However, just for this reason it usually does not provide a precise estimate of $\mathbb{P}(|X| > c)$, just an upper margin.

Theorem 5.3.1. *For any real random variable X, $\forall c > 0, \forall \alpha > 0$ (eg $\alpha = 1, 2, \ldots$)*

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}[|X|^\alpha]}{c^\alpha} \tag{5.6}$$

Proof. In general, given an event A in $\mathscr{A}$ we have the indicator random variable

$$I_A = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

Then we have that

$$\mathbb{E}[I_A] = 0 \cdot \mathbb{P}(I_A = 0) + 1\,\mathbb{P}(I_A = 1) = \mathbb{P}(I_A = 1) = \mathbb{P}(A)$$

To prove Tchebychev lets define

$$A = \{\omega : |X(\omega)| \geq c\} = \{|X| \geq c\}$$

then

$$\mathbb{E}[|X|^\alpha] \overset{(1)}{\geq} \mathbb{E}[I_A \cdot |X|^\alpha] \overset{(2)}{\geq} \mathbb{E}[I_A \cdot c^\alpha] = c^\alpha\,\mathbb{E}[I_A] = c^\alpha\,\mathbb{P}(A)$$

where:

- (1) because $|X|^\alpha \geq I_A \cdot |X|^\alpha$
- (2) we have $|X|^\alpha \geq c^\alpha$ since $|X| \geq c$ when we select with the indicator I_A (otherwise inside parenthesis is 0)

Therefore we conclude that

$$\mathbb{P}(A) = \mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}[X]^\alpha}{c^\alpha}$$

□

Remark 222. An important special case is when $X = Y - \mathbb{E}[Y]$ and $\alpha = 2$, in this case the inequaliy goes to

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq c) \leq \frac{\mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right]}{c^2} = \frac{\mathrm{Var}[Y]}{c^2}$$

But to apply Tchebychev in this form we need to know that the variance exists. Some books call this special case Cebicev inequality and call the general case Markov inequality.

5.3.2 Jensen (Rigo)

Definition 5.3.1 (Convex function (conca tipo $y = x^2$))**.** $f : I \to \mathbb{R}$ is a convex function if

$$\begin{cases} f[\alpha x + (1-\alpha)y] \leq \alpha f(x) + (1-\alpha)f(y) \\ \forall \alpha \in [0,1], x, y \in I \end{cases}$$

where:

- $f[\alpha x + (1-\alpha)y]$ can be seen as a the value given by the function at the mean point between x and y
- $\alpha f(x) + (1-\alpha)f(y)$ the mean of the value assumed by the function in the two extremes

Important remark 38. If f is twice differentiable:

$$f \text{ is convex } \iff f'' \geq 0$$

Example 5.3.1. For instance $f(x) = x^2$, $f(x) = e^x$, $f(x) = |x|$ are convex.

Definition 5.3.2 (Strictly convex function)**.** Same definition as above but instead of $\leq$ we have $<$: $f : I \to \mathbb{R}$ is strictly convex

$$\begin{cases} f[\alpha x + (1-\alpha)y] < \alpha f(x) + (1-\alpha)f(y) \\ \forall \alpha \in [0,1], x, y \in I \end{cases}$$

Important remark 39. If f is twice differentiable:

$$f \text{ is strictly convex } \iff f'' > 0$$

Example 5.3.2. For instance $f(x) = x^2$, $f(x) = e^x$ are strictly convex. Similarly if $I = (0, \infty)$, $f(x) = \frac{1}{x}$ is strictly convex. In fact $f''(x) = 2x^{-3} > 0$, $\forall x > 0$

Proposition 5.3.2 (Jensen inequality)**.** *Let X be a real random variable and $f : I \to \mathbb{R}$ a function defined on interval I. If*

1. *f is a convex function*
2. $\mathbb{P}(X \in I) = 1$
3. $\mathbb{E}[|X|] < +\infty$, $\mathbb{E}[|f(X)|] < +\infty$

Then:

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Moreover, if f is strictly convex and X is not degenerate, then

$$\mathbb{E}[f(X)] > f(\mathbb{E}[X])$$

NB: quest'anno non fatte?

Example 5.3.3. Let's see some application of Jensen inequality.

- $f(x) = x^2$ is strictly convex (second derivative $= 2 > 0$). If we apply Jensen we find out that

  $$\mathbb{E}\left[X^2\right] > [\mathbb{E}[X]]^2 \tag{5.7}$$

  This was already known sice variance (for non degenerate variables as per the theorem) is ≥ 0 (by computational formula of variance).

- absolute value $f(x) = |x|$ (second derivative $= 0$); applying Jensen we discover something new

  $$\mathbb{E}[|X|] \geq |\mathbb{E}[X]| \tag{5.8}$$

- $f(x) = x^{b/a}$ for any $x >= 0$ with $(0 < a < b)$. Applying Jensen

$$\mathbb{E}\left[|X|^b\right] = \mathbb{E}\left[(|X|^a)^{\frac{b}{a}}\right] \geq [\mathbb{E}\left[|X|^a\right]]^{\frac{b}{a}} \tag{5.9}$$

thus Jensen implies that

$$\mathbb{E}\left[(|X|^a)^{\frac{1}{a}}\right] \leq \mathbb{E}\left[\left(|X|^b\right)^{\frac{1}{b}}\right]$$

Remark 223. Now we use Jensen to prove that the rv is degenerate iff its variance is 0.

Proposition 5.3.3.

$$X \sim \delta_a \iff \mathrm{Var}\left[X\right] = 0$$

Proof. Respectively:

- supposing $X = a$ almost surely ($\mathbb{P}(X = a) = 1$), then $\mathbb{E}[X] = a$ and also $\mathbb{E}\left[X^2\right] = a^2$, thus

$$\mathrm{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}\left[X\right])^2 = a^2 - a^2 = 0$$

- otherwise suppose $\mathrm{Var}[X] = 0$: we prove that by contradition. By applying Jensen inequality with $f(x) = x^2$, strictly convex, if X *is non degenerate* we get:

$$\mathbb{E}\left[X^2\right] = \mathbb{E}\left[f(X)\right] > f(\mathbb{E}\left[X\right]) = (\mathbb{E}\left[X\right])^2$$

this happens if and only if $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 > 0$: but we assumed $\mathrm{Var}[X] = 0$ so we found a contradition (thus X must be degenerate).

□

5.4 Characteristic and moment generating function

5.4.1 Characteristic function

Definition 5.4.1 (Characteristic function)**.** If $\mathbf{X} = \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix}$ is a n-variate random vector, the characteristic function $\varphi : \mathbb{R}^n \to \mathbb{C}$ of $\mathbf{X}$ is

$$\begin{aligned} \varphi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}\left[e^{i\mathbf{t}^\top \mathbf{X}}\right] = \mathbb{E}\left[e^{i\sum_i t_i X_i}\right] \qquad \forall \mathbf{t} = \begin{bmatrix} t_1 \\ \dots \\ t_n \end{bmatrix} \in \mathbb{R}^n \\ &= \mathbb{E}\left[\cos \mathbf{t}^\top \mathbf{X} + i \sin \mathbf{t}^\top \mathbf{X}\right], \end{aligned}$$

where

- $i \in \mathbb{I} : i^2 = -1$

- being both $\mathbf{t}$ and $\mathbf{X}$ vectors the $\mathbf{t}^\top \mathbf{X} = \sum_{i=1}^n t_i X_i$ is a scalar and Euler's formula ($e^{ix} = \cos(x) + i\sin(x)$) applies.

Remark 224. However, from now on we assume single variable (because it's more convenient) not n-variate random vector. The definition above simplifies to the following

Definition 5.4.2 (Characteristic function)**.** Let X be a random variable, the characteristic function $\varphi_X(t) : \mathbb{R} \to \mathbb{C}$, existing $\forall t \in \mathbb{R}$ is defined as

$$\begin{aligned}\varphi_X(t) &= \mathbb{E}\left[e^{itX}\right] = \int_{-\infty}^{+\infty} e^{itx} f(x)\, \mathrm{d}x \\ &= \int_{-\infty}^{+\infty} \cos(tx) f(x)\, \mathrm{d}x + i \int_{-\infty}^{+\infty} \sin(tx) f(x)\, \mathrm{d}x\end{aligned}$$

NB: esempio viroliano credo

Example 5.4.1 (Characteristic function of a binomial)**.** Let $X \sim \mathrm{Bin}(n, p)$, $D_x = \{0, 1, \dots, n\}$, the characteristic function is

$$\begin{aligned}\varphi_X(t) &= \mathbb{E}\left[e^{itX}\right] = \sum_{x=0}^{n} e^{itx} \cdot \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (\underbrace{pe^{it}}_{a})^x (\underbrace{1-p}_{b})^{n-x} \\ &\overset{(1)}{=} (1 - p + pe^{it})^n\end{aligned}$$

where in (1) we applied binomial formula $(a+b)^n = \sum_{x=0}^n \binom{n}{x} a^x b^{n-x}$

Example 5.4.2 (Characteristic function of $X \sim \mathrm{N}(0, 1)$)**.**

$$\begin{aligned}\varphi_X(t) &= \mathbb{E}\left[e^{itX}\right] = \int_{-\infty}^{+\infty} e^{itx} f(x)\, \mathrm{d}X = \int_{-\infty}^{+\infty} e^{itx} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\, \mathrm{d}X \\ &= e^{-\frac{t^2}{2}}\end{aligned}$$

l'ultimo passaggio l'ha giusto detto a-la "trust me" presumo (integrale di funzione complessa)

Important remark 40 (Usefulness)*.* Despite being complicated/complex functions, they are useful for several reasons (both theoretical and practical):

1. they *determine the distribution* of the random variable: this is the reason this stuff is so important to statistic (important for Rigo);
2. they provide a *link with the moment* of order k of the variable via *differentiation* (with respect to t evaluated at $t = 0$);
3. they provide a *link with the distribution function* via the *inversion formula*.

Theorem 5.4.1 (Link with distribution)**.** *Supposing we have two random vectors* $\mathbf{X}, \mathbf{Y}$*, these have the same distribution iff they share the characteristic function:*

$$X \sim Y \iff \varphi_{\mathbf{X}}(\mathbf{t}) = \varphi_{\mathbf{Y}}(\mathbf{t}), \quad \forall t \in \mathbb{R}^n \tag{5.10}$$

Important remark 41 (Important properties (Rigo))*.* We have:

1. **link with random variables independence**: if $X \perp\!\!\!\perp Y$, the characteristic function of the sum is equal to the product of the single characteristic functions

$$\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t), \quad \forall t \in \mathbb{R} \tag{5.11}$$

This because

$$\begin{aligned}\varphi_{X+Y}(t) &= \mathbb{E}\left[e^{it(X+Y)}\right] = \mathbb{E}\left[e^{itX}e^{itY}\right] \stackrel{(1)}{=} \mathbb{E}\left[e^{itX}\right]\mathbb{E}\left[e^{itY}\right] \\ &= \varphi_X(t) \cdot \varphi_Y(t), \quad \forall t \in \mathbb{R}\end{aligned}$$

where in (1), since $X \perp\!\!\!\perp Y$, any combination is independent as well, and so we apply the expected value property for product of independent variables $Z \perp\!\!\!\perp W \implies \mathbb{E}[ZW] = \mathbb{E}[Z]\,\mathbb{E}[W]$.
Because of 5.11, characteristic function becomes *very handy* when working with sums of independent rvs.

2. **characteristic function and moments**: if the random variable has the moment of order j, that is $\mathbb{E}\left[|X|^j\right] < +\infty$, then:

   (a) the characteristic function $\varphi_X(t) \in C^j$, where C^j is the collection of functions which have the derivative of order j and such derivative is continuous;

   (b) the derivative of order $r \leq j$ is:

   $$\begin{aligned}\varphi_X(t)^{(r)} &= \frac{\partial^r}{\partial t^r}\varphi_X(t) = \frac{\partial^r}{\partial t^r}\mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[\frac{\partial^r}{\partial t^r}e^{itX}\right] \\ &= \mathbb{E}\left[(iX)^r e^{itX}\right]\end{aligned}$$

   This latter means that in each derivative up to order j we can interchange the operator of derivative and the operator of expectation.
   For instance, suppose we want to calculate the first derivative; by setting $r = 1$

   $$\varphi_X(t)' = \mathbb{E}\left[iXe^{itX}\right]$$

   Actually the **interesting fact** is that if we evaluate the r-th derivative for $t = 0$ we have a direct interpretation/link with the r-th moment

   $$\varphi_X(0)^{(r)} = \mathbb{E}\left[(iX)^r e^{i0X}\right] = i^r\,\mathbb{E}[X^r]$$

Before we said that $\mathbb{E}\left[|X|^j\right] < +\infty \implies \varphi_X(t) \in C^j$. The converse implication does not generally holds (only if j is even). We have

$$\begin{cases}\text{If } j \text{ is odd,} & \mathbb{E}\left[|X|^j\right] < +\infty \implies \varphi_X(t) \in C^j \\ \text{If } j \text{ is even,} & \mathbb{E}\left[|X|^j\right] < +\infty \iff \varphi_X(t) \in C^j\end{cases}$$

As counterexample of the first missing counterimplication, we will see a case where (with $j = 1$ odd) where $\varphi_X(t) \in C^1$ ma $\mathbb{E}[|X|] = +\infty$;

3. **inversion theorem** gives a closed formula for determining the distribution function starting from characteristic function.
Let F be the distribution function of X If/forall $a < b$ such that $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$ then:

$$F(b) - F(a) = \mathbb{P}(a < X \leq b) = \frac{1}{2\pi i} \lim_{c \to +\infty} \int_{-c}^{c} \frac{e^{-ita} - e^{-itb}}{t} \varphi_X(t) \, \mathrm{d}t$$

NB: For inversion the important fact t call for the exam is characteristic functio be inverted: if you l the characteristic f tion, there exists a for that allows to write o the distribution fun (no need to memorize the exam).

4. **continuity theorem**: if X_n and X are real rvs then

$$X_n \xrightarrow{d} X \iff \varphi_X(t) = \lim_{n \to +\infty} \varphi_{X_n}(t), \quad \forall t \in \mathbb{R}$$

We'll see later $\xrightarrow{d}$ means convergence in distribution (an important type of convergence): point is that any time we want to prove convergence in distribution we can, if convenient, prove the limit of characteristic function.

Proposition 5.4.2 (Altre proprietà utili trovate su wikipedia)**.** *Si ha:*

1. *If $X_1, \ldots, X_n$ are independent random variables, and $a_1, \ldots a_n \in \mathbb{R}$, the characteristic function of the linear combination*

$$\varphi_{a_1 X_1 + \ldots + a_n X_n}(t) = \varphi_{X_1}(a_1 t) \cdot \ldots \varphi_{X_n}(a_n t).$$

2. *Let the random variable $Y = aX + b$ be the linear transformation of a random variable X. The characteristic function of Y is $\varphi_Y(t) = e^{itb}\varphi_X(at)$.*

3. *For random vectors $\mathbf{X}$ and $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{B}$ (where $\mathbf{A}$ is a constant matrix and $\mathbf{B}$ a constant vector), we have*

$$\varphi_{\mathbf{Y}}(\mathbf{t}) = e^{i\mathbf{t}^\top \mathbf{B}} \varphi_{\mathbf{X}}\left(\mathbf{A}^\top \mathbf{t}\right)$$

Example 5.4.3 (Characteristic function of $N(\mu, \sigma^2)$)**.** As example of the second Rigo considered that if $X \sim \mu + \sigma Z$ where $Z \sim \mathrm{N}(0,1)$, then $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$. For its characteristic function we have

$$\begin{aligned}\varphi_X(t) &= \mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[e^{it(\mu+\sigma Z)}\right] = \mathbb{E}\left[e^{it\mu} e^{it\sigma Z}\right] = e^{it\mu}\,\mathbb{E}\left[e^{i(t\sigma)Z}\right] = e^{it\mu}\varphi_Z(t\sigma) \\ &= e^{it\mu} e^{-\frac{t^2\sigma^2}{2}}\end{aligned}$$

where in the last passage we used results from example 5.4.2.

Example 5.4.4 (Example by Rigo, fatto solo il mio anno? forse da postporre alle convergenze: weak law of large number)**.** In this example we show that if X_n is iid and the characteristic function has the first derivative at 0, $\exists \varphi_X(0)'$, then the sample mean converges (in distribution and probability) to a constant/degenerate rv.
Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of iid rvs; we define the sample mean as:

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

The characteristic function of the sample mean is

$$\varphi_{\bar{X}_n}(t) = \varphi_{\sum_i X_i}\left(\frac{t}{n}\right) \overset{(\perp\!\!\!\perp)}{=} \prod_{i=1}^{n} \varphi_{X_i}\left(\frac{t}{n}\right) \overset{(id)}{=} \left[\varphi_{X_i}\left(\frac{t}{n}\right)\right]^n$$

Suppose now that the first derivative of the characteristic function of X_i exists in 0, that is $\exists \varphi_{X_i}(0)'$; then by Taylor expansion formula

$$\varphi_{\bar{X}_n}(t) = \left[\varphi_{X_i}\left(\frac{t}{n}\right)\right]^n = \left[\varphi_{X_i}(0) + \frac{t}{n}\varphi_{X_i}(0)' + o\left(\frac{t}{n}\right)\right]^n = \left[1 + \frac{t\varphi_{X_i}(0)' + no\left(\frac{t}{n}\right)}{n}\right]^n$$

where $o\left(\frac{t}{n}\right)$ is the Peano rest. In general $g = o(f)$ if $\lim_{x \to x_0} \frac{g(x)}{f(x)} = 0$.
Now, what is the limit of the formula above for $n \to +\infty$? Using the fact that

$$\text{if } a_n \to a \implies \left(1 + \frac{a_n}{n}\right)^n \to e^a$$

we have (with $a_n = t\varphi_{X_i}(0)' + no\left(\frac{t}{n}\right)$ and noted that $a_n \to t\varphi_{X_i}(0)' + 0$)

$$\varphi_{\bar{X}_n}(t) \to e^{t\varphi_{X_i}(0)'}$$

Now it can be shown (we won't) that the first derivative in 0 is

$$\varphi_{X_i}(0)' = i\alpha, \quad \alpha \in \mathbb{R}$$

and thus we our characteristic function converges to

$$\varphi_{\bar{X}_n}(t) \to e^{it\alpha}, \forall t \in \mathbb{R}$$

Is $e^{it\alpha}$ a characteristic function? Yes the δ_α has this characteristic function since if $X \sim \delta_\alpha$

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[e^{it\alpha}\right] = e^{it\alpha}$$

Hence $\bar{X}_n \xrightarrow{d} \alpha$, by continuity theorem, and since the limit is a degenerate rv, we have not only convergence in distribution byt also convergence in probability $\bar{X}_n \xrightarrow{p} \alpha$.

Important remark 42. The above should be *weak law* of large number (convergence not a.s. but only in probability, check with Virols).
Furthermore, if the sequence is not only iid, but also the mean exists, $\mathbb{E}[|X_i|] < +\infty$, then $\bar{X}_n \xrightarrow{a.s.} \mathbb{E}[X_i]$ then the sample mean converges almost surely to the mean (this is the *strong law of large number*).
But as noted above, it may be that $\exists \varphi_{X_i}(0)'$ even if $\mathbb{E}[|X_i|] = +\infty$.

5.4.2 Moment generating function

Definition 5.4.3 (Moment generating function (mgf))**.** It's obtained from the characteristic function by evaluating it at $-it$, $\varphi_X(-it)$, so that there are no complex number:

$$\varphi_X(-it) = \mathbb{E}\left[e^{-iitX}\right] = \mathbb{E}\left[e^{tX}\right] = M_X(t), \quad \forall t \in \mathbb{R}$$

so

$$M_X(t) = \mathbb{E}\left[e^{tX}\right], \quad \forall t \in \mathbb{R}$$

Poiché $e^{tX} > 0$ si ha che $M_X(t) > 0$

Important remark 43. It's simpler than characteristic function (no i here) but has its drawbacks. Differently from characteristic function (always exists, we have inversion thm):

- MGF always exists for $t = 0$

  $$M_X(0) = 1 < +\infty$$

  MGF may *fail to exist* for $t \neq 0$. If for some $t \neq 0$ (or for $\forall t \in \mathbb{R}$) it is $M_X(t) = +\infty$, in those case MGF is not useful/does not exists;

- we don't have an inversion theorem, so it's useful only for the moments (not distribution).

Important remark 44 (Random variable with MGF). If we know that the moment generating function is finite in a neighborhood of $t = 0$, that is

$$M_X(t) < +\infty, \quad \forall t \in (-\varepsilon, \varepsilon)$$

we say that X *has moment generating function.* In that case it may be convenient to use it instead of the characteristic function, since it can be proven that:

- the random variable has *moments of every order*: $\mathbb{E}\left[|X|^n\right] < +\infty, \forall n$
- the probability distribution of X is *determined* by its moments that is $X \sim Y$ for any rv Y such that $\mathbb{E}\left[X^n\right] = \mathbb{E}\left[Y^n\right], \forall n$.
  The sequence of moments $\mathbb{E}\left[X^n\right]$, with $n = 1, 2, \ldots$, determines the distribution, in the sense that if X and Y does *not* have the same distribution then *either* one of them have some moments not finite or moments both are finite but different for some n:

  $$X \nsim Y \implies \begin{cases} \mathbb{E}\left[|X|^n\right] = +\infty, \text{ for some } n, \text{ OR} \\ \mathbb{E}\left[X^n\right] \neq \mathbb{E}\left[Y^n\right], \text{ for some } n \end{cases}$$

Important remark 45. Consider two random variables X, Y, with moments of every order (mean, variance, third moment etc) *existing* and *coinciding*:

$$\begin{cases} \mathbb{E}\left[|X|^n\right] < +\infty, \ \mathbb{E}\left[|Y|^n\right] < +\infty \\ \mathbb{E}\left[X^n\right] = \mathbb{E}\left[Y^n\right] \end{cases} \quad \forall n$$

Can we conclude that the two random variables have the same distribution? **No** we cannot conclude that (this is contrary to intuition).
Eg if X is lognormal, one can build a suitable rv Y such that X and Y have the same moments of every order and yet $X \nsim Y$.
However this annoying fact doesn't occur if one between X and Y has moment generating function. In that case we can say they have the same distribution.

$$\begin{cases} \mathbb{E}\left[|X|^n\right] < +\infty, \ \mathbb{E}\left[|Y|^n\right] < +\infty \\ \mathbb{E}\left[X^n\right] = \mathbb{E}\left[Y^n\right] \\ X \text{ or } Y \text{ has finite moment generating function} \end{cases} \quad \forall n \implies X \sim Y$$

Example 5.4.5. An important special case where $M_X(t) < +\infty$, $\forall t \in \mathbb{R}$ is

$$|X| \leq c \text{ a.s. for some constant } c$$

In fact

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] \leq \mathbb{E}\left[e^{|tX|}\right] \leq \mathbb{E}\left[e^{|t|c}\right] = e^{|t|c} < +\infty, \quad \forall t \in \mathbb{R}$$

Proposition 5.4.3 (Properties)**.**

NB: da qui in poi roba della viroli direi

$$\left[\frac{\partial^k}{\partial t^k} M_X(t)\right]_{t=0} = \mathbb{E}\left[X^k\right] \tag{5.12}$$

$$M_X(0) = \mathbb{E}\left[e^{0X}\right] = \mathbb{E}\left[1\right] = 1 \tag{5.13}$$

$$M_X(t) = M_Y(t), \forall t \iff F_X(x) = F_Y(y) \qquad \textit{(uniqueness)} \tag{5.14}$$

$$M_{aX+b}(t) = e^{tb} M_X(at), \quad a, b \in \mathbb{R} \tag{5.15}$$

$$X \perp\!\!\!\perp Y \implies M_{X+Y}(t) = M_X(t) \cdot M_Y(t) \tag{5.16}$$

TODO: l'implicazione per l'indipendenza è anche coimplicazione?

Proof. For 5.15

$$M_{aX+b}(t) = \mathbb{E}\left[e^{t(aX+b)}\right] = \mathbb{E}\left[e^{taX} \cdot \underbrace{e^{tb}}_{\text{constant}}\right] = e^{tb} \cdot \mathbb{E}\left[e^{taX}\right] = e^{tb} M_X(at)$$

For 5.16

$$M_{X+Y}(t) = \mathbb{E}\left[e^{t(X+Y)}\right] = \mathbb{E}\left[e^{tX} e^{tY}\right]$$

Now note that:

- first

TODO: questo andrebbe portato più in vista nella sezione indipendenza o prop v. atteso

$$\begin{aligned}
\mathbb{E}\left[g(X)h(Y)\right] &= \int_{D_x} \int_{D_y} g(x)h(y)f(x,y) \, \mathrm{d}x \, \mathrm{d}y \overset{(1)}{=} \int_{D_x} \int_{D_y} g(x)h(y)f_X(x)f_Y(y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_{D_x} g(x)f_X(x) \, \mathrm{d}x \int_{D_y} h(y)f_Y(y) \, \mathrm{d}y = \mathbb{E}\left[g(X)\right] \mathbb{E}\left[h(Y)\right]
\end{aligned}$$

  where (1) due to be $X \perp\!\!\!\perp Y$.

- furthermore

$$\begin{aligned}
\mathbb{E}\left[g(X) + h(Y)\right] &= \int_{D_x} \int_{D_y} (g(x) + h(y)) f(x,y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_{D_x} \int_{D_y} g(x)f(x,y) \, \mathrm{d}x \, \mathrm{d}y + \int_{D_x} \int_{D_y} h(y)f(x,y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_{D_x} g(x) \underbrace{\int_{D_y} f(x,y) \, \mathrm{d}y}_{f(x)} \, \mathrm{d}x + \int_{D_x} \int_{D_y} h(y)f(x,y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_{D_x} g(x)f(x) \, \mathrm{d}x + \int_{D_y} h(y)f(y) \, \mathrm{d}y = \mathbb{E}\left[g(X)\right] + \mathbb{E}\left[h(Y)\right]
\end{aligned}$$

Therefore coming back to our focus, under independence and using the first one

$$M_{X+Y}(t) = \mathbb{E}\left[e^{tX}e^{tY}\right] \overset{(1)}{=} \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] = M_X(t)M_y(t)$$

in (1) because of $\perp\!\!\!\perp$ $\square$

Example 5.4.6 (Esempio rigo triennale)**.** Se $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ allora X possiede la MGF che è (non dimostrato)

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = e^{t\mu + \frac{\sigma^2 t^2}{2}}, \quad \forall t \in \mathbb{R}$$

Inoltre si ha che la derivata prima

$$M_X(t)' = e^{t\mu + \frac{\sigma^2 t^2}{2}}(\mu + \sigma^2 t)$$

e valutandola per $t = 0$

$$M_X(0)' = e^0(\mu + \sigma^2 \cdot 0) = \mu \implies \mathbb{E}\left[X^1\right] = \mathbb{E}\left[X\right] = \mu$$

Proseguendo con la derivata seconda

$$M_X(t)'' = e^{t\mu + \frac{\sigma^2 t^2}{2}}(\mu + \sigma^2 t)^2 + e^{t\mu + \frac{\sigma^2 t^2}{2}}\sigma^2$$

e valutandola per $t = 0$ si ha.

$$M_X(0)'' = e^0(\mu + \sigma^2 \cdot 0)^2 + e^0\sigma^2 = \mu^2 + \sigma^2 \implies \mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2$$

Da cui possiamo ricavare la varianza considerando

$$\mathrm{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$$

Example 5.4.7 (Mgf of bernoulli and binomial)**.** If $X \sim \mathrm{Bern}\left(p\right)$, $p(x) = p^x(1-p)^{1-x}$, $D_x = \{0, 1\}$. Its mgf is:

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = e^{t\cdot 0} \cdot (1-p)p^0 0 + e^{t\cdot 1}p^1(1-p)^0 = 1 - p + pe^t$$

Being the binomial $Y = X_1 + \ldots + X_n$ with X_i iid, by properties of mgfs, the mgf of a binomial is

$$M_Y(t) = \prod_{i=1}^{n}(1 - p + pe^t) = (1 - p + pe^t)^n$$

Example 5.4.8 (Mgf of poisson)**.** Let $X \sim \mathrm{Pois}\left(\lambda\right)$, let's determine $M_X(t)$

$$\begin{aligned} M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \sum_{x=0}^{\infty} e^{tx}\frac{1}{x!}e^{-\lambda}\lambda^x = \sum_{x=0}^{\infty}(e^t\lambda)^x\frac{1}{x!}e^{-\lambda} \\ &\overset{(1)}{=} e^{-\lambda}\cdot e^{\lambda e^t} = e^{-\lambda(1-e^t)} = e^{\lambda(e^t-1)} \end{aligned}$$

where in (1) we used $\sum_{x=0}^{\infty}\frac{c^x}{x!} = e^c$.

Example 5.4.9 (Esercizio richiesto Viroli)**.** By using 5.16 find $M_Y(t)$, with $Y = \sum_{i=1}^n X_i$, $X_i \sim \text{Pois}(\lambda_i)$, and $X_i \perp\!\!\!\perp X_j$.
La mgf di una poisson con parametro λ è $M_X(t) = e^{\lambda(e^t-1)}$, da cui per l'indipendenza possiamo applica la produttoria

$$M_Y(t) = \prod_{i=1}^n e^{\lambda_i(e^t-1)} = e^{\sum_{i=1}^n \lambda_i(e^t-1)} = e^{(e^t-1)\cdot\sum_{i=1}^n \lambda_i}$$

che è la mgf di una poisson con parametro lambda la somma delle lambda componenti (come atteso).
Therefore $\implies Y \sim \text{Pois}(\sum_{i=1}^n \lambda_i)$.

Example 5.4.10 (Esame vecchio viroli)**.** Let X be a bernoulli rv with parameter $\frac{1}{2}$. Find the moment generating functions of $Y = \frac{1}{2} + \frac{X}{2}$.
We have that for the bernoulli

$$M_X(t) = 1 - p + pe^t$$

and consider

$$M_{aX+b}(t) = e^{bt}M_X(at)$$

Now here we have $Y = \frac{1}{2} + \frac{X}{2}$ so $a = b = \frac{1}{2}$, therefore:

$$M_Y(t) = e^{t/2}M_X\left(\frac{t}{2}\right) = e^{\frac{t}{2}}(1 - p + pe^{t/2})$$

Finally, if $p = \frac{1}{2}$

$$M_Y(t) = e^{t/2}\left(\frac{1}{2} + \frac{e^{t/2}}{2}\right) = \frac{1}{2}\left(e^{t/2} + e^t\right)$$

Therefore we have that $M_Y(t) = \frac{1}{2}(e^t + e^{\frac{t}{2}})$

Example 5.4.11 (Esame vecchio viroli)**.** Let X_1 and X_2 be two independent Bernoulli rv with parameters 1/2. find the moment generating function of $Z = X_1 - X_2$.
If $X \sim \text{Bern}(p)$, its $M_X(t) = (1 - p + pe^t)$. Here for the difference of two bernoulli we apply the mgf properties

$$M_{X_1-X_2}(t) = M_{X_1+(-X_2)}(t) \overset{(1)}{=} M_{X_1}(t) \cdot M_{-X_2}(t) \overset{(2)}{=} M_{X_1}(t) + M_{X_2}(-t)$$

with 1 for independence and 2 for linear transformation properties. So considering both as bernoulli with $p = 1/2$

$$\begin{aligned} M_{X_1-X_2}(t) &= (1 - p + pe^t)(1 - p + pe^{-t}) = \left(\frac{1}{2} + \frac{1}{2}e^t\right)\left(\frac{1}{2} + \frac{1}{2}e^{-t}\right) \\ &= \frac{1}{4} + \frac{1}{4}e^{-t} + \frac{1}{4}e^t + \frac{1}{4} = \frac{1}{2} + \frac{1}{4}\left(e^t + e^{-t}\right) \end{aligned}$$

so $M_{X_1-X_2}(t) = 1/2 + 1/4(e^t + e^{-t})$. And Bigo confirms.

Remark 225. The following is a result that become useful sometimes (eg clt)

Proposition 5.4.4 (Mc Laurin expansion of mgf)**.**

$$M_X(t) = 1 + t\,\mathbb{E}\left[X\right] + \frac{t^2}{2!}\,\mathbb{E}\left[X^2\right] + \frac{t^3}{3!}\,\mathbb{E}\left[X^3\right] + \dots \tag{5.17}$$

Proof. In general decomposition of $M_X(t)$ is like the following. Considered that mclaurin expansion of e^{tx}

$$e^{tx} = 1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \dots$$

then

$$\begin{aligned} M_X(t) &= \mathbb{E}\left[e^{tX}\right] = \int_{D_X} e^{tx} f(x)\,\mathrm{d}x \\ &= \underbrace{\int_{D_X} 1 f(x)\,\mathrm{d}x}_{=1} + \int_{D_X} tx f(x)\,\mathrm{d}x + \int_{D_X} \frac{(tx)^2}{2!} f(x)\,\mathrm{d}x + \int_{D_X} \frac{(tx)^3}{3!} f(x)\,\mathrm{d}x + \dots \\ &= 1 + t\int_{D_X} x f(x)\,\mathrm{d}x + \frac{t^2}{2!}\int_{D_X} x^2 f(x)\,\mathrm{d}x + \frac{t^2}{3!}\int_{D_X} x^3 f(x)\,\mathrm{d}x + \dots \\ &= 1 + t\,\mathbb{E}\left[X\right] + \frac{t^2}{2!}\,\mathbb{E}\left[X^2\right] + \frac{t^3}{3!}\,\mathbb{E}\left[X^3\right] + \dots \end{aligned}$$

□

Remark 226. Now we see an example where mgf does not always exists

Example 5.4.12 (Mgf of Gamma)**.** Let $X \sim \text{Gamma}\left(\alpha, \beta\right)$, $\alpha, \beta > 0$

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

with $D_x = [0, +\infty)$ and

$$\begin{aligned} &\Gamma(x) = \int_0^{+\infty} x^{\alpha-1} e^{-x}\,\mathrm{d}x, \quad \forall \alpha > 0 \\ &\alpha \in \mathbb{N} \implies \Gamma(\alpha) = (\alpha - 1)! \end{aligned}$$

Let's evaluate $M_X(t)$

$$\begin{aligned} M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \int_0^{+\infty} e^{tx} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} e^{-(\beta-t)x} \cdot x^{\alpha-1}\,\mathrm{d}x \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} e^{-(\beta-t)x} \cdot x^{\alpha-1} \frac{(\beta-t)^\alpha}{(\beta-t)^\alpha}\,\mathrm{d}x \\ &= \frac{\beta^\alpha}{(\beta-t)^\alpha} \underbrace{\int_0^{+\infty} \frac{(\beta-t)^\alpha}{\Gamma(\alpha)} \cdot e^{-(\beta-t)x} x^{\alpha-1}\,\mathrm{d}x}_{=1,\text{ since } f(x) \text{ of a Gamma}\,(\alpha,\beta-t)} \end{aligned}$$

Therefore

$$M_X(t) = \frac{\beta^\alpha}{(\beta - t)^\alpha} = \left(\frac{\beta}{(\beta - t)}\right)^\alpha = \left(\frac{\beta - t}{(\beta)}\right)^{-\alpha} = \left(1 - \frac{t}{\beta}\right)^{-\alpha}$$

where, since $\alpha > 0$ (and it's an exponent), $M_X(t)$ is well defined only if the base is positive

$$1 - \frac{t}{\beta} > 0 \iff t < \beta$$

Example 5.4.13 (Esercizio richiesto Viroli)**.** For this exercise:

1. compute the second moment $\mathbb{E}\left[X^2\right]$ of the binomial distribution using the second derivative of mgf evaluated in 0;
2. for the binomial, verify property 2 of mgf, that is $M_X(0) = 1$;
3. eval $\mathbb{E}\left[X\right]$ where X is Gamma by using first derivative of mgf

We have

1. per la prima deriviamo due volte e valutiamo in 0 la mgf della binomiale che è $(1 - p + pe^t)^n$. Si ha

   $$\begin{aligned}&\left[(1 - p + pe^t)^n\right]' = n(1 - p + pe^t)^{n-1}(pe^t)\\&\left[(1 - p + pe^t)^n\right]'' = n\left[(n-1)(1 - p + pe^t)^{n-2}(pe^t)^2 + (pe^t)(1 - p + pe^t)^{n-1}\right]\end{aligned}$$

   che valutata per $t = 0$ da

   $$n(n-1)p^2 + np = n^2p^2 - np^2 + np$$

   Possiamo verificare il risultato applicando la formula di calcolo della varianza (dato che della binomiale si conoscono varianza e valore atteso)

   $$\begin{aligned}\text{Var}\left[X\right] &= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2\\np(1-p) &= \mathbb{E}\left[X^2\right] - n^2p^2\\\mathbb{E}\left[X^2\right] &= np(1-p) + n^2p^2 = np - np^2 + n^2p^2\end{aligned}$$

2. per $t = 0$, si ha $(1 - p + pe^t)^n = (1 - p + p)^n = 1$
3. la mgf della gamma è $\left(\frac{\lambda}{\lambda - t}\right)^\alpha$ la sua derivata prima

   $$\alpha\left(\frac{\lambda}{\lambda - t}\right)^{\alpha-1}\left(-\frac{\lambda(-1)}{(\lambda - t)^2}\right) = \alpha\left(\frac{\lambda}{\lambda - t}\right)^{\alpha-1}\left(\frac{\lambda}{(\lambda - t)^2}\right)$$

   che valutata in $t = 0$ da α/λ, il valore atteso della gamma

Example 5.4.14 (Normal distributions)**.** Let $X \sim \mathrm{N}(0,1)$, then lets derive the mgf

$$\begin{aligned} M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \int_{-\infty}^{+\infty} e^{tx} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \, \mathrm{d}x \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{tx-\frac{1}{2}x^2} \, \mathrm{d}x \end{aligned}$$

Now we apply this substitution trick

$$tx - \frac{1}{2}x^2 = \frac{t^2 - (x-t)^2}{2}$$

because of the expansion

$$\frac{t^2 - (x-t)^2}{2} = \frac{t^2 - x^2 - t^2 + 2xt}{2} = tx - \frac{x^2}{2}$$

So

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{t^2-(x-t)^2}{2}} \, \mathrm{d}x \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}} e^{\frac{-(x-t)^2}{2}} \, \mathrm{d}x \\ &= e^{\frac{t^2}{2}} \underbrace{\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-t)^2}{2}} \, \mathrm{d}x}_{=1 \text{ since integral of } \mathrm{N}(t,1)} \\ &= e^{\frac{t^2}{2}} \end{aligned}$$

Therefore

$$X \sim \mathrm{N}(0,1) \iff M_X(t) = e^{t^2/2}$$

while applying properties of mgf it turns out that, if $X \sim \mathrm{N}(0,1)$

$$\sigma X + \mu \sim \mathrm{N}\left(\mu, \sigma^2\right) \iff M_{\sigma X+\mu}(t) = e^{\mu t} M_X(\sigma t) = e^{\mu t} e^{\frac{\sigma^2 t^2}{2}}$$

Example 5.4.15 (Esercizio richiesto Viroli)**.** Regarding the normal (consider $X \sim \mathrm{N}(0,1)$):

- prove $\frac{\partial M_{\sigma X+\mu}(t)}{\partial t} = \mu$
- derive $\mathbb{E}\left[X^2\right]$ by mgf
- check that $\mathrm{Var}\left[\sigma X + \mu\right] = \sigma^2$ (applying $\mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$)

If the mgf of the general normal is $e^{\mu t} e^{\frac{\sigma^2 t^2}{2}}$

1. we derive it one time and evaluate for $t = 0$ to find μ

$$e^{\mu t} \cdot \mu \cdot e^{\frac{1}{2}\sigma^2 t^2} + e^{\frac{1}{2}\sigma^2 t^2} \cdot \left(\frac{1}{2}\sigma^2 2t\right) \cdot e^{\mu t} = e^{\mu t + \frac{1}{2}\sigma^2 t^2}\left(\mu + \frac{1}{2}\sigma^2 2t\right)$$

   che valutata per $t = 0$ restituisce $e^0(\mu + 0) = 1 \cdot \mu = \mu$

2. la derivata seconda è

$$e^{\mu t+\frac{1}{2}\sigma^2 t^2}\left(\mu+\frac{1}{2}\sigma^2 2t\right)^2+\left(\frac{1}{2}\sigma^2 2\right)e^{\mu t+\frac{1}{2}\sigma^2 t^2}$$
$$e^{\mu t+\frac{1}{2}\sigma^2 t^2}\left[\left(\mu+\frac{1}{2}\sigma^2 2t\right)^2+\sigma^2\right]$$

se $t=0$

$$e^0\left[(\mu+0)^2+\sigma^2\right]=\mu^2+\sigma^2$$

3. abbiamo

$$\operatorname{Var}[Y]=\mathbb{E}\left[Y^2\right]-\mathbb{E}[Y]^2=\mu^2+\sigma^2-\mu^2=\sigma^2$$

Example 5.4.16 (Esercizio viroli, primo set)**.** Let $\mathbf{X}=\begin{bmatrix}X_1\\X_2\end{bmatrix}$ be a bivariate vector with joint density $f_{\mathbf{X}}(x_1,x_2)=2e^{-(x_1+x_2)}$ where $X_1>X_2>0$

1. find $M_{\mathbf{X}}(t)$
2. compute $\mathbb{E}[X_1]$ by $M_{\mathbf{X}}(t)$
3. compute $\mathbb{E}[X_1]$ by definition
4. are $X_1 \perp\!\!\!\perp X_2$, both by density and by moment generating function

We have:

1.

$$\begin{aligned}M_{\mathbf{X}}(t)&=2\int_0^{+\infty}\int_{x_2}^{\infty}e^{tx_1}e^{tx_2}e^{-(x_1+x_2)}\,\mathrm{d}x_1\,\mathrm{d}x_2\\&=2\int_0^{+\infty}e^{-x_2(1-t_2)}\cdot\int_{x_2}^{+\infty}\mathrm{d}x_1\,\mathrm{d}x_2\\&=2\int_0^{\infty}e^{-x_2(1-t_2)}\cdot\left[-\frac{e^{x_1(1-t_1)}}{1-t_1}\right]_{x_2}^{\infty}\mathrm{d}x_2\\&=2\frac{1}{1-t_1}\int_0^{+\infty}e^{-x_2(2-t_1-t_2)}\,\mathrm{d}x_2\\&=\frac{2}{(1-t_1)(2-t_1-t_2)}\end{aligned}$$

2.

$$\begin{aligned}\frac{\partial M_{\mathbf{X}}(\mathbf{t})}{\partial t_1}\Big|_{\mathbf{t=0}}&=2(1-t_1)^{-2}(2-t_1-t_2)^{-1}+2(1-t_1)^{-1}(2-t_1-t_2)^{-2}\Big|_{\mathbf{t=0}}\\&=\frac{2}{2}+\frac{2}{4}=\frac{3}{2}\end{aligned}$$

3. it's longer, we have:

$$\mathbb{E}[X_1] = \int_{D_{X_1}} x_1 f_{X_1}(x_1)\, \mathrm{d}x_1$$

where

$$\begin{aligned} f_{X_1}(x_1) &= \int_{D_{X_2}} f_{\mathbf{X}}(x_1, x_2)\, \mathrm{d}x_2 \\ &= \int_0^{x_1} 2e^{-(x_1+x_2)}\, \mathrm{d}x_2 = \int_0^{x_1} 2e^{-x_1} e^{-x_2}\, \mathrm{d}x_2 \\ &= 2e^{-x_1} \cdot \int_0^{x_1} e^{-x_2}\, \mathrm{d}x_2 = 2e^{-x_1} \left[-e^{-x_2}\right]_0^{x_1} \\ &= 2e^{-x_1}(1 - e^{-x_1}) = 2e^{-x_1} - 2e^{-2x_1} \end{aligned}$$

therefore

$$\begin{aligned} \mathbb{E}[X_1] &= \int_0^{+\infty} x_1 \big(2e^{-x_1} - 2e^{-2x_1}\big)\, \mathrm{d}x_1 \\ &= 2 \underbrace{\int_0^{+\infty} x_1 e^{-x_1}\, \mathrm{d}x_1}_{\text{expected value of Exp}\,(1)} - \underbrace{\int_0^{+\infty} x_1 2e^{-2x_1}\, \mathrm{d}x_1}_{\text{expected value of Exp}\,(2)} \\ &= 2 \cdot 1 - \frac{1}{2} = \frac{3}{2} \end{aligned}$$

4. by the density

$$f_{X_2}(x_2) = \int_{x_2}^{+\infty} 2e^{-(x_1+x_2)}\, \mathrm{d}x_1 = 2e^{-x_1} \cdot \left[-e^{-x_1}\right]_{x_2}^{+\infty} = e^{-x_2} e^{-x_2} = e^{-2x_2}$$

Now we check if $f_{X_1}(x_1) \cdot f_{X_2}(x_2) = f_{\mathbf{X}}(x_1, x_2)$:

$$2e^{-x_1}(1 - e^{-x_1}) e^{-2x_2} 2 \neq 2e^{-(x_1+x_2}$$

therefore they are not independent.
Now let's check according to the moment generating function; we observe that:

$$M_{X_1}(t_1) = M_{\mathbf{X}}(t_1, 0) = \frac{2}{(1-t_2)\frac{1}{2-t_1}} \qquad M_{X_2}(t_2) = M_{\mathbf{X}}(0, t_2) = \frac{2}{2-t_2}$$

Since $M_{\mathbf{X}}(\mathbf{t}) \neq M_{X_1}(t_1) M_{X_2}(t_2)$ are not independent.
Note: in case of mutually independent rvs:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{p} f_{X_i}(x_i) \\ F_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{p} F_{X_i}(x_i) \\ M_{\mathbf{X}}(\mathbf{t}) &= \prod_{i=1}^{p} M_{X_i}(t_i) \end{aligned}$$

Example 5.4.17 (Mgf of Geometric and Negative binomial)**.** Let $X_1, \dots, X_n \sim$ Geom(p) iid rvs. Find $M_Y(t)$ where $Y = \sum_{i=1}^n X_i$. What can you say about the distribution of Y?
For a geometric rv we have

$$\mathbb{P}(X = x) = p(1-p)^{x-1}, \qquad D_X = \{1, 2, \dots\}$$

so

$$\begin{aligned} M_X(t) &= \sum_{x=1}^{\infty} e^{tx} p(1-p)^{x-1} = \sum_{x=1}^{\infty} e^{tx} p \frac{1-p}{1-p}(1-p)^{x-1} \\ &= \frac{p}{1-p} \cdot \sum_{x=1}^{\infty} \left[e^t(1-p)\right]^x = \frac{p}{1-p} \cdot \left(\sum_{x=0}^{\infty} \left[e^t(1-p)\right]^x - 1\right) \end{aligned}$$

Now we define $q = 1-p$; if $|e^t(1-p)| < 1$ the previous series converges to $\frac{1}{1-qe^t}$. Therefore the $M_X(t)$ exists only for $e^t < \frac{1}{1-p}$, that is $t < -\log(1-p))$. For such values we have

$$M_X(t) = \frac{p}{q}\left(\frac{1}{1-qe^t} - 1\right) = \frac{p}{q}\left(\frac{qe^t}{1-qe^t}\right) = \frac{pe^t}{1-qe^t}$$

Now

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t) = \left[\frac{pe^t}{1-q^{e^t}}\right]^n$$

with the last being the moment generating function of a negative binomial distribution with parameters n and p

5.5 Conditional distribution

5.5.1 Definition and examples

Remark 227. Roughly speaking the problem is: given 2 real random variable Y, X we aim to evaluate the distribution of Y given that $X = x$.
Ad esempio se $X =$ altezza e $Y =$ peso vorrei conoscere la distribuzione del peso nell'ipotesi che l'altezza sia 1.70

Definition 5.5.1 (Conditional distribution)**.** Let $\begin{bmatrix} X \\ Y \end{bmatrix}$ be a bivariate rv; the conditional distribution of Y given X is any function

$$\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)$$

defined $\forall x \in \mathbb{R}$, $\forall C \in \beta(\mathbb{R}^n)$ satisfying the following properties:

NB: direi gli input siano due, C e x

1. for each fixed $x \in \mathbb{R}$, the map

$$C \to \mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)$$

   is a probability measure on $\beta(\mathbb{R}^2)$

2. for each fixed $C \in \beta(\mathbb{R}^2)$, the map
$$x \to \mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)$$
is Borel measurable and satisfies
$$\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C\right) = E_X \left\{\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)\right\}$$
where E_X means expectation with respect to X.

Important remark 46 (Important remarks). Some important remarks:

1. it can be shown that the conditional distribution of Y given X (namely a function satisfying definition) *always exists* and is *almost surely unique*. This remark is important because looking at the defn it's not sure that any function such as that defined exists.
   Here a.s uniqueness is meant with respect to the probability distribution of X
2. the notation $\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)$ is very useful but also quite dangereous; it should be regarded as "the conditional probability that $\begin{bmatrix} X \\ Y \end{bmatrix} \in C$ given $X = x$. Be careful however: by definition $\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)$ is only a function satisfying the two properties of the definition.
   In particular it's *not necessarily equal* to ratio between probability intersection divided by probability $P(X = x)$:
$$\text{it's not necessarily } \frac{\mathbb{P}\left(X = x, \begin{bmatrix} X \\ Y \end{bmatrix} \in C\right)}{\mathbb{P}\left(X = x\right)}$$
   This should be obvious since it may be that $P(X = x) = 0, \forall x \in \mathbb{R}$ (which is possible eg in continuous distribution).
   For instance suppose $P(X = x) = 0, \forall x \in \mathbb{R}$ (or equivalently the distribution function is continuous) then by the previous remark $\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)$ exists, but it certainly does not coincide with the ratio above (not defined)
3. if X is **discrete** the operator $E_X(\cdot)$ means
$$E_X \left\{\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)\right\} = \sum_{x \in B} \mathbb{P}\left(X = x\right) \mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)$$
   where B is any set, finite or countable such that $\mathbb{P}\left(X \in B\right) = 1$.
   Similarly if X is **absolutely continuous** the operator $E_X(\cdot)$ means
$$E_X \left\{\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)\right\} = \int_{-\infty}^{+\infty} f(x) \, \mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)$$
   where f is the density of X

Important remark 47 (Some useful properties of conditional distribution). Some properties:

- any time we aim to evaluate the conditional probability we can substitute as follows:

$$\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right) = \mathbb{P}\left(\begin{bmatrix} x \\ Y \end{bmatrix} \in C | X = x\right) \tag{5.18}$$

  This is intuitively obvious, since we're conditioning on $X = x$ we know that $X = x$ and can substitute it within parenthesis;

- if $X \perp\!\!\!\perp Y$, then

$$\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right) = \mathbb{P}\left(\begin{bmatrix} x \\ Y \end{bmatrix} \in C | X = x\right) = \mathbb{P}\left(\begin{bmatrix} x \\ Y \end{bmatrix} \in C\right)$$

  where in the last passage i can drop the conditioning because X and Y are independent

- if we know that $\mathbb{P}(X \in A) = 1$ for some $A \in \beta(\mathbb{R}^n)$, then is enough to evaluate

$$\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right), \quad \forall x \in A$$

  and not necessarily $\forall x \in \mathbb{R}$.
  For example if $X \sim \text{Unif}(0,1)$, its enough to evaluate

$$\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right), \quad \forall x \in (0,1)$$

Remark 228. Unfortunately, in general there is not an intuitive formula to evaluate conditional distributions (there is in some cases as we'll see later).

Remark 229. Let's see some examples, the first of which is fundamental

Example 5.5.1 ($\mathbb{P}(Y = X)$: a usual question at the Rigo exam)**.** Suppose $X \perp\!\!\!\perp Y$ and Y has a continuous distribution function. What is the $\mathbb{P}(X = Y)$? This should be 0. Let's show it.
To answer let's define $C = \{(x,y) \in \mathbb{R}^2 : x = y\}$ which is the set of points constituting the diagonal

NB: qui ho solo invertito $\mathbb{P}(Y = X)$ per facilitare la memorizzazione

$$\begin{aligned}
\mathbb{P}(Y = X) &= \mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C\right) \overset{(1)}{=} E_X\left\{\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)\right\} \\
&\overset{(2)}{=} E_X\left\{\mathbb{P}\left(\begin{bmatrix} x \\ Y \end{bmatrix} \in C | X = x\right)\right\} \overset{(3)}{=} E_X\left\{\mathbb{P}\left(\begin{bmatrix} x \\ Y \end{bmatrix} \in C\right)\right\} \\
&= E_X(\underbrace{\mathbb{P}(Y = x)}_{0}) \overset{(4)}{=} E_X(0) = 0
\end{aligned}$$

with:

- (1) by property 2 of defn,

- (2) by 5.18 (since we're conditioning I can write x instead of X)
- (3) since thery are independent I can drop the conditioning
- (4) since being Y continuous, the probability that $Y = x$ (aka a single value) is zero

NB: fatto solo il mio anno?

Remark 230. Note that: in statistical inference the elements of the sample are often assumed to be iid. Under this assumption, if the distribution of the character in the population is *continuous* what is the prob of having the sample with all different observation?
It's 1 (almost sure event). This because $\mathbb{P}(X_i = X_j) = 0$, $\forall i \neq j$, so that the probability that $\mathbb{P}(X_1, \ldots, X_n$ are all distinct$) = 1$

NB: in questo ho invertito i ruoli di X e Y se no mi incartavo. Notare che nel caso precedente Y è continua, qui è X ad esserlo

Example 5.5.2 ($\mathbb{P}(Y = \sin(X))$)**.** Suppose $X \perp\!\!\!\perp Y$, and X has a continuous distribution function. Let's prove that $\mathbb{P}(Y = \sin(X)) = 0$.
A quick way to do it is the following: since Y is independent of X then is still indipendent of any trasformation (and thus $\sin(X)$).
Hence by exercise 5.5.1 it suffices province either/equivalently that:

- $\sin(X)$ has a continuous distribution function (because if it's continuous we can repeat the argument of the previous exercise)
- $\mathbb{P}(\sin(X) = a)) = 0$, $\forall a \in \mathbb{R}$ (this is trivially true if $a \notin [-1, 1]$)

We follow the second way, supposing $a \in [-1, 1]$ and define:

$$I_a = \{x \in \mathbb{R} : \sin x = a\}$$

we have that I_a is countable (pensa sull'asse delle x, ci sono infiniti punti di seno che hanno altezza a se questo è tra -1 e 1). Thus the probability:

$$\mathbb{P}(\sin(X) = a) = \mathbb{P}(X \in I_a) \overset{(1)}{=} \sum_{x \in I_a} \mathbb{P}(X = x) \overset{(2)}{=} \sum_{x \in I_a} 0 = 0$$

with (1) since I_a is countable and (2) because X is a continuous distribution function

Example 5.5.3 ($\mathbb{P}(Y = X)$ with independent discrete distribution)**.** What can be said about $\mathbb{P}(Y = X)$ if $X \perp\!\!\!\perp Y$ and they are both discrete?
Since X is discrete, there's a set $B \subset \mathbb{R}$ such that $\mathbb{P}(X \in B) = 1$. Hence the $P(X = Y)$ can be written as

$$\begin{aligned}\mathbb{P}(X = Y) &= \sum_{x \in B} \mathbb{P}(X = x, Y = X) = \sum_{x \in B} \mathbb{P}(X = x, Y = x) \\ &\overset{(\perp\!\!\!\perp)}{=} \sum_{x \in B} \mathbb{P}(X = x)\,\mathbb{P}(Y = x)\end{aligned}$$

and this may be > 0.

Example 5.5.4. Let A, B, C be iid with all $\sim \operatorname{Exp}(1)$[1]. Let's define the random parabola (named like this because coefficiente a, b, c are rvs):

$$f(x) = Ax^2 + Bx + C, \qquad \forall x \in \mathbb{R}$$

[1] Always positive rv with density function e^{-x} and distribution $1 - e^{-x}$

What is the probability that f has real roots?
It is the probability $\mathbb{P}\left(B^2 - 4AC \geq 0\right)$. To evaluate it, we have to choose one of the three variable and condition on it; eg let's condition on C:

$$\begin{aligned}
\mathbb{P}\left(B^2 - 4AC \geq 0\right) &= E_C\left\{\mathbb{P}\left(B^2 \geq 4AC | C = c\right)\right\} \\
&= E_C\left\{\mathbb{P}\left(B^2 \geq 4Ac | C = c\right)\right\} \\
&\overset{(\perp\!\!\!\perp)}{=} E_C\left\{\mathbb{P}\left(B^2 \geq 4Ac\right)\right\} \\
&= \int_{-\infty}^{+\infty} \mathbb{P}\left(B^2 \geq 4Ac\right) f(c)\,\mathrm{d}c \\
&\overset{(1)}{=} \int_{0}^{+\infty} \mathbb{P}\left(B^2 \geq 4Ac\right) e^{-c}\,\mathrm{d}c \\
&\overset{(2)}{=} \int_{0}^{+\infty} E_A\left\{\mathbb{P}\left(B^2 \geq 4Ac\right) | A = a\right\} e^{-c}\,\mathrm{d}c \\
&= \int_{0}^{+\infty} E_A\left\{\mathbb{P}\left(B^2 \geq 4ac\right)\right\} e^{-c}\,\mathrm{d}c \\
&= \int_{0}^{+\infty}\int_{0}^{+\infty} \mathbb{P}\left(B^2 \geq 4ac\right) e^{-a} e^{-c}\,\mathrm{d}a\,\mathrm{d}c \\
&= \int_{0}^{+\infty}\int_{0}^{+\infty} \mathbb{P}\left(B \geq 2\sqrt{ac}\right) e^{-a} e^{-c}\,\mathrm{d}a\,\mathrm{d}c \\
&\overset{(3)}{=} \int_{0}^{+\infty}\int_{0}^{+\infty} e^{-2\sqrt{ac}} e^{-a} e^{-c}\,\mathrm{d}a\,\mathrm{d}c
\end{aligned}$$

where in

- (1) since C is exponential (and with density only on positive side)
- (2) we have to evaluate $\mathbb{P}\left(B^2 \geq 4Ac\right)$ and its convenient to do it conditioning further on A
- (3) considered that B is exponential (if $Z \sim \operatorname{Exp}(\lambda)$ then $\mathbb{P}(Z > z) = 1 - \mathbb{P}(Z \leq z) = 1 - (1 - e^{-\lambda z}) = e^{-\lambda z}$.

Example 5.5.5. Suppose $X \perp\!\!\!\perp Y$, $X \sim \operatorname{Unif}(0,1)$ and $Y \sim \mathrm{N}(0,1)$. We want to evaluate the distribution function of the product XY.
Here conditional distribution become handy. For all $a \in \mathbb{R}$, by definition the distribution function is

NB: fatto solo il mio anno?

$$\begin{aligned}
\mathbb{P}(XY \leq a) &\overset{(1)}{=} E_X(\mathbb{P}(XY \leq a | X = x)) = E_X(\mathbb{P}(xY \leq a | X = x)) \\
&\overset{(\perp\!\!\!\perp)}{=} E_X(\mathbb{P}(xY \leq a)) \overset{(2)}{=} \int_{-\infty}^{+\infty} \mathbb{P}(xY \leq a) f(x)\,\mathrm{d}x \\
&= \int_{0}^{1} \mathbb{P}(xY \leq a) 1\,\mathrm{d}x \overset{(3)}{=} \int_{0}^{1} \mathbb{P}\left(\mathrm{N}(0,1) \leq \frac{a}{x}\right)\mathrm{d}x
\end{aligned}$$

with (1) by definition, (2) since X is uniform (absolutely continuous), (3) because Y is normal.
After this we go to our friend mathematician asking for help.

5.5.2 Formula to calculate it?

Important remark 48. How to calculate $\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right)$? We know this object exists and in many problem its enough to know it.
Unfortunately *there is not* a general formula which allows to calculate the probability above in every situation. Such a formula exists in *two special cases*:

1. X discrete
2. $\begin{bmatrix} X \\ Y \end{bmatrix}$ absolutely continuous

Definition 5.5.2 (Discrete case)**.** If X is discrete, there is a set $B \subset \mathbb{R}$, B finite or countable, $\mathbb{P}(X \in B) = 1$, and $\mathbb{P}(X = x) > 0$, $\forall x \in B$ (true by definition of discreteness). Hence it suffices to let

$$\mathbb{P}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in C | X = x\right) = \frac{\mathbb{P}\left(X = x, \begin{bmatrix} X \\ Y \end{bmatrix} \in C\right)}{\mathbb{P}(X = x)}, \forall x \in B, \forall C \in \beta(\mathbb{R}^2)$$

This is just the base definition of conditional probability with positive denominator, being the distribution discrete and focusing on $x \in B$.

Definition 5.5.3 (Continuous case)**.** If $(X, Y)^\top$ is absolutely continuous with joint density $f(x, y)$, then the conditional distribution of Y given $X = x$ is still absolutely continuous with *conditional density*:

$$h(y|x) = \frac{f(x, y)}{f_1(x)}, \quad \text{where } f_1(x) > 0$$

where $f(x, y)$ is the joint density of $(X, Y)^\top$ and f_1 is the marginal density of X (namely the integral $f(x, y)$ in $\,\mathrm{d}y$)
In this case we have an explicit formula for *conditional distribution* function of Y given $X = x$ as:

$$\mathbb{P}(Y \leq y | X = x) = \int_{-\infty}^{y} \frac{f(x, t)}{f_1(x)} \,\mathrm{d}t, \qquad \forall x, y \in \mathbb{R} : f_1(x) > 0$$

TODO: esempio fatto solo quest'anno, non nel mio

Example 5.5.6. Suppose $X \sim \mathrm{Unif}(0, 1)$ and $Y | X = x \sim \mathrm{Bin}(n, x)$, $\forall x \in (0, 1)$. Find the conditional distribution of X given Y.
To this end we first note that, being Y discrete

$$\mathbb{P}(Y \in \{0, 1, \dots, n\}) = E_X\left\{\underbrace{\mathbb{P}(Y \in \{0, 1, \dots, n\} | X = x)}_{=1}\right\} = E_X(1) = 1$$

Again Y is discrete so that

$$\begin{aligned}
\mathbb{P}(X \in A|Y = y) &= \frac{\mathbb{P}(X \in A, Y = y)}{\mathbb{P}(Y = y)} \\
&= \frac{E_x\{\mathbb{P}(X \in A, Y = y|X = x)\}}{E_x\{\mathbb{P}(Y = y|X = x)\}} \\
&= \frac{E_x\{\mathbb{P}(X \in A, Y = y|X = x)\}}{E_x\left\{\binom{n}{y}x^y(1-x)^{n-y}\right\}} \\
&= \frac{E_x\left\{1_A(x)\binom{n}{y}x^y(1-x)^{n-y}\right\}}{E_x\left\{\binom{n}{y}x^y(1-x)^{n-y}\right\}} \\
&= \frac{\int_0^1 1_A(x)\binom{n}{y}x^y(1-x)^{n-y}\,\mathrm{d}x}{\int_0^1 x^y(1-x)^{n-y}\,\mathrm{d}x}
\end{aligned}$$

Example 5.5.7. If $n = 2$, $X_{(1)} = min(X_1, X_2)$ and $X_{(2)} = max(X_1, X_2)$

Example 5.5.8 (Example with order statistics)**.** Let S and T be iid with $S \sim \text{Unif}(0,1)$. Define $X = \min(S,T)$ and $Y = \max(S,T)$. We want to write the conditional distribution of Y given $X = x$.
To this end we first note that for theorem 5.2.6 $\begin{bmatrix} X \\ Y \end{bmatrix}$ is absolutely continuous being the order statistic of corresponding to $\begin{bmatrix} S \\ T \end{bmatrix}$, and S, T are iid and absolutely continuous.
In addition the joint density f of $\begin{bmatrix} X \\ Y \end{bmatrix}$ is

$$f(x,y) = \begin{cases} 2! \cdot g(x)g(y) & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}$$

where g is density common to S, T, namely $\text{Unif}(0,1)$

$$g_1(x) = \begin{cases} 1 & \text{if } x \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

Hence

$$f(x,y) = \begin{cases} 2 & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

the marginals are (eg)

$$g(x) = \int f(x,y)\,\mathrm{d}y = \int_x^1 2\,\mathrm{d}y = 2(1-x), \quad \forall x \in (0,1)$$

Hence since $\begin{bmatrix} X \\ Y \end{bmatrix}$ is absolutely continuous, we have the formula for the *conditional distribution* of Y given X, which is still absolutely continuous with density:

$$h(y|x) = \begin{cases} \dfrac{f(x,y)}{g(x)} = \dfrac{2}{2(1-x)} = \dfrac{1}{1-x}, & \text{if } 0 < x < y < 1 \\ 0, & \text{elsewhere} \end{cases}$$

So $\forall x \in (0,1)$ (select first the min in 0,1) $Y|X = x \sim \text{Unif}\,(x,1)$ (the max is uniformely distributed between the min and 1, distribution max). A bits of *interpretation* on the results: Since S and T are iid Unif $(0,1)$ observing the pair (S,T) is like to select "at random" a point form the unit square.
Suppose now that $X = \min(S,T)$; what can be said about $Y = \max(S,T)$? Certainly $Y > X$ so if we fix a point $x \in [0,1]$, y is above the diagonal $y = x$, that is it is in the $[x,1]$. In fact the distribution of $Y|X \sim \text{Unif}\,(x,1)$: and this is why we obtained $1/(1-x)$ as density (coming from that distribution)

5.6 Multivariate normal

NB: quest'anno è partito secco dalla multivariate e pace

Remark 231. Let's start from univariate and see that multivariate formula are univariate generalization

Proposition 5.6.1 (Characteristic functions of univariate normal)**.** *If* $Z \sim \text{N}\,(0,1)$ *and* $X \sim \text{N}\left(\mu, \sigma^2\right)$ *then* $\forall t \in \mathbb{R}$*:*

$$\varphi_Z(t) = e^{-t^2/2} \tag{5.19}$$

$$\varphi_X(t) = e^{it\mu - \frac{1}{2}(t\sigma)^2} \tag{5.20}$$

Proof. If $Z \sim \text{N}\,(0,1)$: then its characteristic function is

$$\varphi_Z(t) = \mathbb{E}\left[e^{itZ}\right] = \int_{-\infty}^{+\infty} e^{itx} \frac{\exp\left(-\frac{1}{2}x^2\right)}{\sqrt{2\pi}}\, \mathrm{d}x \overset{(1)}{=} \ldots = e^{-t^2/2}, \quad \forall t \in \mathbb{R}$$

(in (1) after doing calculation). If $X \sim \text{N}\left(\mu, \sigma^2\right)$ then X can be written as $X = \mu + \sigma Z$ with $Z \sim \text{N}\,(0,1)$ and thus we can derive the formula given the definition (in the univariate case) as:

$$\begin{aligned} \varphi_X(t) &= \mathbb{E}\left[e^{it(\mu+\sigma Z)}\right] = \mathbb{E}\left[\underbrace{e^{it\mu}}_{\text{constant}} \; e^{it\sigma Z}\right] = e^{it\mu}\,\mathbb{E}\left[e^{i(t\sigma)Z}\right] \\ &= e^{it\mu} \cdot \varphi_Z(t\sigma) = e^{it\mu - \frac{1}{2}(t\sigma)^2} \qquad \forall t \in \mathbb{R} \end{aligned}$$

□

Remark 232. MVN is not so important for this course: it's very important for statistician, but from point of view of probability it's just a special distribution among the others.

Definition 5.6.1. An n-dimensional random vector $X = (X_1, \ldots, X_n)^\top$ is normally distributed with parameters $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}$ (a $n \times n$ symmetric non-negative definite/ positive semi-definite (≥ 0) matrix [2]), if the characteristic

[2] Recappino da wikipedia: For any real invertible matrix X the product $X^\top X$ is a positive definite matrix (if the means of the columns of X are 0, then this is also called the covariance matrix). A simple proof is that for any non-zero vector $\mathbf{z}$, the condition $\mathbf{z}^\top X^\top X \mathbf{z} = (X\mathbf{z})^\top (X\mathbf{z}) = \|X\mathbf{z}\|^2 > 0$, since the invertibility of matrix X means that $X\mathbf{z} \neq 0$.

function of X is given by:

$$\varphi_X(t) = \mathbb{E}\left[e^{i\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top \boldsymbol{\Sigma}\mathbf{t}}\right], \quad \forall \mathbf{t} \in \mathbb{R}^n$$

Important remark 49 (Meaning of parameters). $\boldsymbol{\mu}$ is the vector of the mean, $\boldsymbol{\Sigma}$ is the so called covariance matrix which have variances on the diagonal, covariance out of main diagonal

$$\boldsymbol{\mu} = \begin{bmatrix} \mathbb{E}[X_1] \\ \dots \\ \mathbb{E}[X_n] \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}[X_2] & \dots & \mathrm{Cov}(X_2, X_n) \\ \dots & \dots & \dots & \dots \\ \mathrm{Cov}(X_n, X_1) & \dots & \dots & \mathrm{Var}[X_n] \end{bmatrix}$$

Remark 233. Our definition includes not only absolutely continuous normal vector, but also other (eg degenerate in some cases)

Important remark 50 (Main properties). Some remarks:

- a **linear trasformation** of a normal random variable is still normal: if $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

  $$\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X} \sim \mathrm{N}\left(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top\right)$$

  where $\mathbf{a} \in \mathbb{R}^m$ the matrix $\mathbf{B}$ is $m \times n$.
  In particular, if X is normal, all marginals are still normal being the marginal obtained via a linear transformation (therefore we get a normal) that merely extract the marginal/subset. Eg

  $$\mathbf{Y} = \begin{bmatrix} X_1 \\ X_2 \\ X_4 \end{bmatrix} = \mathbf{0} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \mathbf{0} + \mathbf{B}\mathbf{X}$$

- if $X \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then independence of componence amounts to null covariances: if $\boldsymbol{\Sigma}$ is diagonal, then $X_1, \dots, X_n$ are independent;

- regarding $\det \boldsymbol{\Sigma}$:

  - if $\det(\boldsymbol{\Sigma}) > 0$, then $\boldsymbol{\Sigma}$ is also positive-definite (> 0, not only ≥ 0) and $\mathbf{X}$ is absolutely continuous with density

    $$f(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}, \qquad \forall \mathbf{x} \in \mathbb{R}^n$$

    The univariate density we know is a special case where the matrix $\boldsymbol{\Sigma}$ is positive definite (otherwise matrix can be inverted).
    For $n = 1$ the density $\boldsymbol{\Sigma}$ reduce to a scalar ($\boldsymbol{\Sigma} = \sigma^2$, variance of the variable) and $\boldsymbol{\mu}$ to a single number

    $$f(x) = \frac{\exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)}{\sigma\sqrt{2\pi}}$$

– if otherwise $\det \mathbf{\Sigma} = 0$ then X is still normal, but the distribution of X is no longer absolutely continuous.
For instance if $n = 1$ and $\sigma^2 = \mathbf{\Sigma} = 0$ then

$$\varphi_X(t) = e^{-it\mu}$$

and $X = \mu$ is degenerate. In other terms if $n = 1$, the above definition implies that the degenerate random variable are considered normal.

NB: l'ha fatta il mio anno, non questo

Linear transformation proof. In order to prove that if $\mathbf{X} \sim \mathrm{MVN}(\boldsymbol{\mu}, \mathbf{\Sigma})$ then $\mathbf{Y} = \mathbf{a} + \mathbf{BX} \sim \mathrm{MVN}\left(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B\Sigma B}^\top\right)$ we write the characteristic function of $\mathbf{Y}$ according to the definition above. Let's evaluate it:

$$\mathbb{E}\left[e^{i\mathbf{t}^\top \mathbf{Y}}\right] = \mathbb{E}\left[e^{i\mathbf{t}^\top(\mathbf{a}+\mathbf{BX})}\right] = \mathbb{E}\left[\underbrace{e^{i\mathbf{t}^\top \mathbf{a}}}_{\text{constant}} \cdot e^{i\mathbf{t}^\top \mathbf{BX}}\right] = e^{i\mathbf{t}^\top \mathbf{a}} \underbrace{\mathbb{E}\left[e^{i\mathbf{t}^\top \mathbf{BX}}\right]}_{\varphi_X(\mathbf{B}^\top \mathbf{t})}$$
$$= e^{i\mathbf{t}^\top \mathbf{a}} \cdot e^{i\mathbf{t}^\top \mathbf{B}\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top \mathbf{B\Sigma B}^\top \mathbf{t}} = \exp\left(i\mathbf{t}^\top(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}) - \frac{1}{2}\mathbf{t}^\top(\mathbf{B\Sigma B}^\top)\mathbf{t}\right)$$
$$\iff Y \sim \mathrm{N}\left(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B\Sigma B}^\top\right)$$

□

Example 5.6.1 (Assignment 1 Viroli, Exercise 3)**.** Suppose that $\mathbf{X}$ is a bivariate Gaussian vector with components (X_1, X_2) which are marginally standard normally distributed and with correlations $1/2$:

1. What is the distribution of $Y_1 = 2X_1 - X_2$ and $Y_2 = X_1 - X_2/2$
2. find the linear transformation from $\mathbf{X}$ to $\mathbf{Y}$ and ask what is the distribution of $\mathbf{Y}$

Since $X_1, X_2 \sim \mathrm{N}(0, 1)$ and considered that

$$\frac{1}{2} = \mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}[X_1]}\sqrt{\mathrm{Var}[X_1]}} = \frac{\mathrm{Cov}(X_1, X_2)}{1 \cdot 1}$$
$$= \mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$$

1. if $Y_1 = 2X_1 - X_2$ and $Y_2 = X_1 - X_2/2$, then Y_1, Y_2 will be linear combinations of correlated normals; the distributions of Y_1, Y_2 will be normals with mean the linear combinations of means:

$$\mathbb{E}[Y_1] = \mathbb{E}[2X_1 - X_2] = 2\,\mathbb{E}[X_1] - \mathbb{E}[X_2] = 0$$
$$\mathbb{E}[Y_2] = \mathbb{E}\left[X_1 - \frac{1}{2}X_2\right] = \mathbb{E}[X_1] - \frac{1}{2}\,\mathbb{E}[X_2] = 0$$

Applying $\mathrm{Var}[aX + bY] = a^2\,\mathrm{Var}[X] + b^2\,\mathrm{Var}[Y] + 2ab\,\mathrm{Cov}(X, Y)$ we have:

$$\mathrm{Var}[Y_1] = \mathrm{Var}[2X_1 - X_2] = 4\,\mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2 \cdot 2(-1)\,\mathrm{Cov}(X_1, X_2)$$
$$= 4 + 1 - 4\frac{1}{2} = 5 - 2 = 3$$
$$\mathrm{Var}[Y_2] = \mathrm{Var}\left[X_1 - \frac{1}{2}X_2\right] = \mathrm{Var}[X_1] + \frac{1}{4}\,\mathrm{Var}[X_2] + 2\left(-\frac{1}{2}\right)\mathrm{Cov}(X_1, X_2)$$
$$= 1 + \frac{1}{4} - \frac{1}{2} = \frac{3}{4}$$

Therefore: $Y_1 \sim \mathrm{N}(0,3)$, $Y_2 \sim \mathrm{N}\left(0, \frac{3}{4}\right)$

2. in general, a linear trasformation of a multivariate normal is still normal; if $\mathbf{X} \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a n-dimensional random vector and $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$, with $\mathbf{A}$ an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^m$, then $\mathbf{Y}$ is a m-dimensional random vector and specifically $\mathbf{Y} \sim \mathrm{MVN}\left(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top\right)$.
   In our case $m = n = 2$ and we have:
$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}\right), \qquad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 2X_1 - X_2 \\ X_1 - \frac{1}{2}X_2 \end{bmatrix}$$
   so $\mathbf{Y} = \mathbf{AX} + \mathbf{0}$, represent the linear transformation needed to obtain $\mathbf{Y}$, where
$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 1 & -1/2 \end{bmatrix}$$
   Therefore to evaluate the parameters of $\mathbf{Y}$:
$$\begin{aligned} \mathbf{A}\boldsymbol{\mu} + \mathbf{b} &= \begin{bmatrix} 2 & -1 \\ 1 & -1/2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top &= \begin{bmatrix} 2 & -1 \\ 1 & -1/2 \end{bmatrix} \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ -1 & -1/2 \end{bmatrix} = \begin{bmatrix} 3 & 3/2 \\ 3/2 & 3/4 \end{bmatrix} \end{aligned}$$
   Finally:
$$\mathbf{Y} \sim \mathrm{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 & 3/2 \\ 3/2 & 3/4 \end{bmatrix}\right)$$

Chapter 6

Convergences and related topics

6.1 Convergence

Remark 234. Given a sequence $X_1, \ldots, X_n, \ldots$ of real random variables and another real random variable X:

- we are interested in checking wether or not X_n converges to X as n goes to $+\infty$, written $X_n \to X$.
- there are 4 types/modes of convergence: in each case as n become larger, X_n get "closer" to X; but *the way this happen is different* so one convergence does not necessarely imply others (we will see relationship between them in the following.

Remark 235. In the following all the standard calculus limits involved are meant for $n \to +\infty$.

Definition 6.1.1 (Almost sure convergence)**.** X_n converge almost surely to X and we write $X_n \xrightarrow{a.s.} X$ if and only if

$$\begin{aligned}\mathbb{P}\left(\omega \in \Omega : X_n(\omega) \to X(\omega)\right) &= \mathbb{P}\left(\omega \in \Omega : \lim_{n \to +\infty} X_n(\omega) = X(\omega)\right) \\ &= 1\end{aligned}$$

Remark 236 (Interpretation)*.* The idea s simple: for a fixed $\omega \in \Omega$, with varying n $X_n(\omega)$ is a sequence of real number (not random variables); this sequence can converge to the real number $X(\omega)$ or not.
If this is going to happen for all the elements of Ω then we met the condition.

Definition 6.1.2 (L_p convergence)**.** Considered $p > 0$, X_n converges to X in L_p, written $X_n \xrightarrow{L_p} X$ if and only if:

1. all the X_n have moment of order p: $\mathbb{E}\left[|X_n|^p\right] < +\infty$;
2. X has moment of order p as well: $\mathbb{E}\left[|X|^p\right] < +\infty$;

3. finally $\mathbb{E}\left[|X_n - X|^p\right] \to 0$ that is

$$\lim_{n\to+\infty} \mathbb{E}\left[|X_n - X|^p\right] = 0$$

Definition 6.1.3 (Convergence in probability)**.** X_n converges to X in probability, written $X_n \xrightarrow{p} X$, if and only if

$$\lim_{n\to+\infty} \mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0, \qquad \forall\varepsilon > 0$$

Definition 6.1.4 (Convergence in distribution)**.** X_n converges to X in distribution, written $X_n \xrightarrow{d} X$, if and only if

$$\lim_{n\to+\infty} F_{X_n}(x) = F_X(x), \qquad \forall x \in \mathbb{R} \text{ where } F \text{ is continuous}$$

where F_{X_n} and F_X are the distribution functions of X_n and X.

Remark 237 (Requirement of convergence in distribution)*.* For this last case, intuitively, it would be more natural to require the convergence to hold on all the domain ($\forall x \in \mathbb{R}$, not only where F is continuous) that is

$$\lim_{n\to+\infty} F_{X_n}(x) = F_X(x), \qquad \forall x \in \mathbb{R}$$

but this would be a too much severe requirement.
To understand why, suppose we have both degenerate $X_n = \frac{1}{n}$ and $X = 0$. Intuitively we would like X_n to converge to X. Here, for these degenerate, the distribution functions are:

$$F_{X_n}(x) = \begin{cases} 0 & \text{if } x < \frac{1}{n} \\ 1 & \text{if } x \geq \frac{1}{n} \end{cases}, \quad F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

The unique discontinuity point of F is $x = 0$ where $F_X(0) = 1$ and $F_{X_n}(0) = 0$ but, if we exclude it

$$\begin{cases} \lim_{n\to+\infty} F_{X_n}(x) = 0 = F(x) & \text{if } x < 0 \\ \lim_{n\to+\infty} F_{X_n}(x) = 1 = F(x) & \text{if } x > 0 \end{cases}$$

so we can say $X_n \xrightarrow{d} X$.
However if we would require the stronger condition (convergence $\forall x \in \mathbb{R}$), it is no longer true that $\lim_{n\to\infty} F_{X_n}(x) = F(x)$ since for $x = 0$ we have

$$\lim_{n\to+\infty} F_{X_n}(0) = 0 \neq 1 = \neq F_X(0)$$

and thus $X_n \not\xrightarrow{d} X$.
Thus if we would require $\lim_{n\to+\infty} F_{X_n}(x) = F_X(x), \forall x \in \mathbb{R}$ we would get the disturbing consequence that $X_n = \frac{1}{n}$ does not converge in distribution to $X = 0$ and this is a consequence we don't like.

Proposition 6.1.1 (Convergence in distribution of transformation)**.** *If $X_n \xrightarrow{d} X$ and $f : \mathbb{R} \to \mathbb{R}$ is a continuous function then $f(X_n) \xrightarrow{d} f(X)$*

Important remark 51 (Connection among 4 modes of convergence). Summarized by the following schema (to be read as "if $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{p}$ to the same X"):

$$
\begin{array}{ccccc}
\xrightarrow{L_k} & \underset{k>s}{\Longrightarrow} & \xrightarrow{L_s} & & \\
 & & \Downarrow & & \\
\xrightarrow{a.s.} & \Longrightarrow & \xrightarrow{p} & \Longrightarrow & \xrightarrow{d}
\end{array}
$$

Finally, there's only a special case of double implication between $\xrightarrow{p}$ and $\xrightarrow{d}$ in case of degenerate distribution. If X is degenerate ($X = a$ almost surely):

$$X_n \xrightarrow{p} X \iff X_n \xrightarrow{d} X$$

Important remark 52 (Some remarks on $\xrightarrow{p}$). We have that

1. if $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{p} Y$, then $X = Y$ almost surely ($\mathbb{P}(X = Y) = 1$). In other terms the limit in probability is unique (provided it exists).
   This fact has a useful consequence: suppose that we know $X_n \xrightarrow{p} X$ and we aim to prove $X_n \xrightarrow{a.s.} X$ or $X_n \xrightarrow{L_p} X$ (stronger types). In this case, the only possible limit can be X: suppose for example that $X_n \xrightarrow{a.s.} Z$ then, we have $X_n \xrightarrow{p} Z$ by the diagram, and by the previous result we have that $Z = X$ almost surely;

2. in general if $X_n \xrightarrow{p} X$ it may be that X_n fails to converge to X a.s.. However if $X_n \xrightarrow{p} X$, there is a subsequence $1 \leq n_1 < n_2 < n_3 < \dots$ such that $X_{n_j} \xrightarrow{a.s.} X$ as $j \to +\infty$.
   In other terms, convergence in probability implies a.s. convergence along a suitable subsequence

Remark 238. Now some counterexamples to show that some double implications don't work (as stated in the graph of convergence implications).

Example 6.1.1 ($X_n \xrightarrow{p} X \not\Longrightarrow X_n \xrightarrow{L_p} X$ counterexample)**.** Let X_n be such that $\mathbb{P}(X_n = 0) = \frac{n-1}{n}$ and $\mathbb{P}(X_n = n) = \frac{1}{n}$. Let $X = 0$. In this case $X_n \xrightarrow{p} X$ but $X_n \not\xrightarrow{L_1} X$, there's convergence in probability but not in L_p (with $p = 1$):

- to prove convergence in probability we fix $\varepsilon > 0$. Then

$$
\begin{aligned}
\mathbb{P}(|X_n - X| > \varepsilon) &= \mathbb{P}(|X_n| > \varepsilon) \\
&\overset{(1)}{=} \mathbb{P}(|X_n| > \varepsilon, X_n = 0) + \mathbb{P}(|X_n| > \varepsilon, X_n = n) \\
&= 0 + \mathbb{P}(X_n = n) = \frac{1}{n}
\end{aligned}
$$

  where in (1) $\mathbb{P}(A) = \sum_i \mathbb{P}(B_i) \cdot \mathbb{P}(A|B_i) = \sum_i \mathbb{P}(A \cap B_i)$ was applied (considered that X_n by assumption takes 2 values, 0 and n). Hence since $\frac{1}{n} \to 0$ we can state $X_n \xrightarrow{p} X$

- to show that $X_n \not\xrightarrow{L_1} X$ we note that

$$\begin{aligned}\mathbb{E}\left[|X_n - X|\right] &= \mathbb{E}\left[|X_n|\right] = |0|\,\mathbb{P}\left(X_n = 0\right) + |n|\,\mathbb{P}\left(X_n = n\right)\\ &= 0 + n\,\mathbb{P}\left(X_n = n\right) = n\frac{1}{n} = 1, \quad \forall n\end{aligned}$$

  Therefore $\mathbb{E}\left[|X_n - X|\right] \not\to 0$ and thus $X_n \not\xrightarrow{L_1} X$

Proof. To prove the implication $X_n \xrightarrow{L_p} X \implies X_n \xrightarrow{p} X$ it suffices to use Tchebychev inequality. Suppose infact that $X_n \xrightarrow{L_p} X$: then given $\varepsilon > 0$, to have convergence in probability $\mathbb{P}\left(|X_n - X| > \varepsilon\right)$ must go to 0. Now we have that an upper bound for $\mathbb{P}\left(|X_n - X| > \varepsilon\right)$ is

$$\mathbb{P}\left(|X_n - X| > \varepsilon\right) \overset{(1)}{\leq} \frac{\mathbb{E}\left[|X_n - X|^p\right]}{\varepsilon^p} \overset{(2)}{\to} 0$$

where

- (1) due to Tchebychev
- (2) since by definition/assumption on $X_n \xrightarrow{L_p} X$

So given that the right part goes to 0, even the left part goes to 0 and saying that means that $X_n \xrightarrow{p} X$. ☐

NB: new entry questo anno?

Example 6.1.2. Let $(X_n)_{n\in\mathbb{N}}$ be iid Unif $(0, 1)$ and define

$$Y_n = n \min(X_1, \ldots, X_n)$$

Does Y_n converge in distribution and where?
To see it, fix x

- $\forall x > 0$ and consider distribution function

$$\mathbb{P}\left(Y_n \leq x\right) = \mathbb{P}\left(\min(X_1, \ldots, X_n) \leq \frac{x}{n}\right) = 1 - \left(1 - F\left(\frac{x}{n}\right)\right)^n$$

  where F is the distribution function of $X_1 \sim$ Unif $(0, 1)$.
  Hence $\mathbb{P}\left(Y_n \leq x\right) = 1 - \left(1 - \frac{x}{n}\right)^n$ converges to $1 - e^{-x}$ for $n \to \infty$.

- for $x \leq 0$ $\mathbb{P}\left(X_n \leq x\right) = 0$, $\forall n$

To summarize

$$\lim_{n\to+\infty} \mathbb{P}\left(X_n \leq x\right) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-x} & \text{if } x > 0 \end{cases}$$

The latter is the distribution function of Exp (1).
Hence $Y_n \xrightarrow{d}$ Exp (1)

NB: Non fatto st'anno?

Example 6.1.3 (A counterexample for $X_n \xrightarrow{a.s.} X \not\Longrightarrow X_n \xrightarrow{L_p} X$)**.** As counterexample where a.s. convergence does not imply L_1 convergence considering the space:

$$(\Omega, \mathscr{A}, \mathbb{P}) = \left([0, 1], \mathscr{B}([0, 1]), m\right)$$

with m the Lebesgue measure. In general the Lebesuge meassure is *not* a probability measure, because on the real line it gives $+\infty$; but if defined on 0-1 its max is 1 so can be a probability measure.
We define also

$$X_n = n \cdot \mathbb{1}_{[0,\frac{1}{n}]}(\omega)$$
$$X = 0$$

Here by construction we have that $\omega \in [0,1]$; if

- $\omega \in (0,1]$ then $\omega > \frac{1}{n}$ for large n. Therefore for large n we have that $X_n(\omega) = 0$.

- $\omega = 0$, we have that $X(0) = n\mathbb{1}_{[0,\frac{1}{n})}(0) = n$ that goes to $+\infty$ as $n \to +\infty$.

Hence

$$\begin{aligned}\mathbb{P}(\omega \in \Omega : X_n(\omega) \to X(\omega)) &= \mathbb{P}(\omega \in \Omega : X(\omega) = 0) = \mathbb{P}(0,1] \\ &= m(0,1] = 1 - 0 = 1\end{aligned}$$

That is $X_n \xrightarrow{a.s.} X$.
However:

$$\begin{aligned}\mathbb{E}[|X_n - X|] = \mathbb{E}[|X_n|] &= \mathbb{E}\left[n \cdot \mathbb{1}_{[0,1/n)}(\omega)\right] = n \cdot \mathbb{E}\left[\mathbb{1}_{[0,1/n)}(\omega)\right] \\ &= n \cdot \mathbb{P}([0,1/n)) = n \cdot m[0,1/n) = n \cdot \frac{1}{n} \\ &= 1\end{aligned}$$

Hence here $X_n \xrightarrow{a.s.} X$ but $X_n \not\xrightarrow{L_1} X$.

Example 6.1.4. An example where convergence in distribution does not imply convergence in probability. Considering the same space: **NB**: Non fatto st'anno?

$$(\Omega, \mathscr{A}, \mathbb{P}) = ([0,1], \mathscr{B}([0,1]), m)$$

now define $X_n = \mathbb{1}_{[0,1/2]}(\omega)$ and $X = \mathbb{1}_{(1/2,1]}(\omega)$. In this case we have that

$$|X_n - X| = 1, \quad \forall n$$

so X_n fails to converge to X in probability: $X_n \not\xrightarrow{p} X$. However the distribution functions are:

$$F(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

and the other is the same:

$$F_n(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Hence since $F_n = F$, $\forall n$, we have that $X_n \xrightarrow{d} X$.

Example 6.1.5 (Esercizio dalla triennale)**.** Si verifichi che $\frac{1}{n} \xrightarrow{d} 0$ con

- $\frac{1}{n}$ la variabile degenere X_n tale che $\mathbb{P}\left(X_n = \frac{1}{n}\right) = 1$
- 0 è la degenere X tale che $\mathbb{P}(X = 0) = 1$

Le funzioni di ripartizione sono

$$F_{X_n}(x) = \begin{cases} 1 & \text{se } x \geq \frac{1}{n} \\ 0 & \text{se } x < \frac{1}{n} \end{cases}, \quad F_X(x) = \begin{cases} 1 & \text{se } x \geq 0 \\ 0 & \text{se } x < 0 \end{cases}$$

Se

- $x < 0$ non ci sono problemi, infatti
$$F_{X_n}(x) = F(x) = 0, \quad \forall n \geq 1, \forall x < 0$$
- $x > 0$ si ha $\frac{1}{n} < x$ $\forall n$ sufficientemente grande, quindi si ha
$$F_{X_n}(x) = 1 = F(x), \quad \forall n \text{sufficientemente grande}, \forall x < 0$$

In definitiva $F(x) = \lim_{n\to+\infty} F_{X_n}(x)$, $\forall x \neq 0$.

Questo basta per concludere che $1/n \xrightarrow{d} 0$. Infatti la funzione di ripartizione di X è discontinua in 0, quindi, per avere convergenza in distribuzione, non occorre che $F_X(0) = \lim_n F_{X_n}(0)$. Ed infatti in questo esempio non è vero che $F_X(0) = \lim_n F_{X_n}(0)$ ($F_X(0) = 1$ ma $F_{X_n}(0) = 0$, $\forall n$)

6.2 Laws of large numbers

Remark 239. Laws (plural) because there are many of them (some of which are more famous/attractive).

Definition 6.2.1 (Sequence satisfying LLN)**.** Let $(X_n)_{n\in\mathbb{N}} = X_1, X_2, \dots$ be a sequence of real rvs. We say it satisfies the law of large number if the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_i X_i$$

converges to V for some random variable V.

Important remark 53 (Types of LLN). If

- convergence of sample mean is almost sure, $\overline{X}_n \xrightarrow{a.s.} V$, we speak of *strong low of large numbers*;
- convergence of sample mean is in probability, $\overline{X}_n \xrightarrow{p} V$, we speak of *weak law of large number*

Remark 240. Roughly speaking, any time we prove sample mean converges to a limit we have a law of large number. There are research papers that discover new large of large numbers frequently: they simply prove that a sample mean of certain sequences $X_1, \dots, X_n$ converges to something.
In the following we'll see 3 strong laws of large numbers (SLLN) and one weak law of large numbers (WLLN).

Important remark 54. The limit V can be an arbitrary real rv; however the most important/famous (what most people think is LLN) is when the rvs are iid with existing mean, $\mathbb{E}[X_i], \forall i \in \mathbb{N}$, and V is degenerate $V = \delta_{\mathbb{E}[X_i]}$.

6.2.1 Strong laws

Theorem 6.2.1 (Kolmogorov strong law of large numbers (SLLN1))**.** *If $(X_n)_{n\in\mathbb{N}}$ is iid, then $\overline{X}_n$ converges a.s. $\iff$ $\mathbb{E}[|X_1|] < +\infty$.*
Moreover if $\mathbb{E}[|X_1|] < +\infty$, then $\overline{X}_n \xrightarrow{a.s.} \mathbb{E}[X_1]$ (that is $V = \mathbb{E}[X_1]$)

Anno mio, più sintetico: If $(X_n)_{n\in\mathbb{N}}$ is iid and $\mathbb{E}[|X_1|] < +\infty$, then $\overline{X}_n \xrightarrow{a.s.}$ $\mathbb{E}[X_1]$.

Remark 241. Another strong law of large number follows, where we drop the iid hypothesis and replace it with some other condition.

Theorem 6.2.2 (A second example of strong LLN (SLLN2))**.** *Given a sequence of rvs $(X_n)_{n\in\mathbb{N}}$, if:*

- $\sup_n \mathbb{E}\left[X_n^2\right] < +\infty$ *(higher second moment is still finite)*
- *random variables have common mean* $\mathbb{E}[X_1] = \mathbb{E}[X_n], \forall n$
- $\mathrm{Cov}(X_i, X_j) \leq 0, \forall i \neq j$

then again $\overline{X}_n \xrightarrow{a.s.} \mathbb{E}[X_1]$

Remark 242. Note that in this second version, the X_n are neither independent nor identically distributed; this version is not as popular as the first one, however it's practically very useful.

Proof. Let's prove SLLN2 (only convergence in probability, the almost sure is easier).
It suffices to apply Tchebychev inequality: given $\varepsilon > 0$ we have that

$$
\begin{aligned}
\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]\right| > \varepsilon\right) \overset{(1)}{=} \mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[X_1]\right| > \varepsilon\right) \leq \frac{\mathrm{Var}\left[\overline{X}_n\right]}{\varepsilon^2} &= \frac{\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n X_i\right]}{\varepsilon^2} = \frac{\mathrm{Var}\left[\sum_{i=1}^n X_i\right]}{n^2\varepsilon^2} \\
&= \frac{1}{n^2\varepsilon^2}\left\{\sum_{i=1}^n \mathrm{Var}[X_i] + 2\underbrace{\sum_{1\leq i<j\leq n} \mathrm{Cov}(X_i, X_j)}_{\leq 0}\right\} \\
&\leq \frac{1}{n^2\varepsilon^2}\sum_{i=1}^n \mathrm{Var}[X_i] \leq \frac{1}{n^2\varepsilon^2}\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \\
&\overset{(2)}{\leq} \frac{nc}{n^2\varepsilon^2} = \frac{c}{\varepsilon^2}\frac{1}{n} \to 0
\end{aligned}
$$

where

- in (1) before applying raw Tchebychev, we substituted applying equivalence of expected values

$$\mathbb{E}\left[\overline{X}_n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{n}{n}\mathbb{E}\left[X_i\right] = \mathbb{E}\left[X_i\right]$$

- c is such that $\mathbb{E}\left[X_n^2\right] \leq c$, $\forall n$.

This proves that $\overline{X}_n \xrightarrow{p} \mathbb{E}\left[X_1\right]$. Indeed, as claimed in the thorem, one also obtains $X_n \xrightarrow{a.s.} \mathbb{E}\left[X_1\right]$ but we will not prove almost sure convergence. □

Remark 243. In the next example we have a strong law but the limit is not the mean. To state it we recall a definition.

Definition 6.2.2 (Stationary sequence of rv)**.** A sequence $(X_n)_{n\in\mathbb{N}}$ is said to be *stationary* if the probability distribution of the sequence starting from two, is the same of the distribution of the unshifted sequence:

$$(X_2, X_3, X_4, \ldots) \sim (X_1, X_2, X_3, \ldots)$$

Hence the probability distribution of the sequence is invariant (doesn't change under shifts); in some framework this is the classical assumptions.

Theorem 6.2.3 (SLLN3)**.** *If X_n is stationary and $\mathbb{E}\left[|X_1|\right] < +\infty$ (mean of X_1 exists), then $\overline{X}_n \xrightarrow{a.s.} V$ where V is a suitable rv (not necessarily degenerate).*

Remark 244. Two reasons why we mention the result above:

1. stationarity is an important assumption (like iid)
2. this is an example where we have a strong lln (being the convergence as) but the limit is not the mean (this does not need to be the case).

6.2.2 Examples and consequences

Example 6.2.1 (A very classical example)**.** We have an urn containing black an white balls from which we draw *with* replacement. The proportion p of white balls is not known, we want to make inference on it. Let:

$$X_i = \begin{cases} 1 & \text{if } \textit{white} \text{ ball drawn at trial } i \\ 0 & \text{if } \textit{black} \text{ ball drawn at trial } i \end{cases}$$

Since the drawing are with replacement sequence (X_i) are iid, and $\mathbb{E}\left[X_1\right] = p$, so by Kolmogorov's strong law we obtain that the sample mean converges to p, that is $\overline{X}_n \xrightarrow{a.s.} p$:

$$\frac{\text{n. palline bianche nelle prime } n \text{ prove}}{n} = \overline{X}_n \xrightarrow{a.s.} \mathbb{E}\left[X_1\right] = p$$

In other words, in this example the laws ensure us that the procedure based on intuition on making inference on p (look at extracted proportion) is going to converge to the desidered quantity of interest and so if n is high, we can hope $\overline{X}_n$ to be near p.
A question unsolved (for the moment?) is: how much big must be n in order $\overline{X}_n$ be near to p.

Remark 245. A consequence of SLLN1 is the following

Theorem 6.2.4 (Glivenko-Cantelli thm)**.** *If $(X_n)_{nin\mathbb{N}}$ is iid, then*

$$\sup_t |F_n(t) - F_X(t)| \xrightarrow{a.s.} 0$$

where F_X is the distribution function common to the X_n and F_n is the so-called empirical distribution function

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq t), \quad \forall t \in \mathbb{R}$$

Remark 246. This result just states that, the empirical distributin function F_n, regarded as an estimate based on the data '$X_1, \ldots, X_n$ of the true/common distribution function F_X is consistent.
In fact F_n converges to F_X uniformly over t with probability 1.
The Glivenko-Cantelli thm is especially meaningful in nonparametric statistical inference

Example 6.2.2. The white black ball example 6.2.1 can be generalized as follows: let $(X_n)_{n\in\mathbb{N}}$ be iid but the distribution function F of X_1 is unknown. To make inference on F we fix a real number $x \in \mathbb{R}$ and we define the following random indicator variables

NB: esempio dal mio anno: questo è da mettere assieme a glivenko cantelli, è la dimostrazione

$$Y_i = \mathbb{1}.(X_i \leq x)$$

Then $\{Y_i\}$ are still iid and

$$\mathbb{E}[Y_1] = \mathbb{E}[\mathbb{1}.(X_1 \leq x)] = \mathbb{P}(X_1 \leq x) = F(x)$$

Hence

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}.(X_1 \leq x) \xrightarrow{a.s.} F(x)$$

In general:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}.(X_1 \leq x)$$

is called the *empirical distribution function*, and can be regarded as an estimate of F. Infact, in statistical terms, the empirical distribution function is a *consistent* estimator of the true distribution function (that is, as the sample size goes $n \to \infty$, the procedure converge to the true value).

6.2.3 A weak law

Remark 247. Finally we state a weak law of large numbers.

Proposition 6.2.5. *If $(X_n)_{n\in\mathbb{N}}$ is iid then the following conditions are in $\iff$ relation between them (if one is fullfilled all the others are):*

- $\overline{X}_n \xrightarrow{p} a$ *for some constant a;*

- $\varphi_{X_1}(t)$ *is differentiable at* 0 *and* $\varphi_{X_1}(0)' = ia$
- *we have*

$$\lim_{c\to+\infty} \mathbb{E}\left[X_1 \cdot \mathbb{1}.(|X_1| \le c)\right] = a$$
$$\lim_{c\to+\infty} c \cdot \mathbb{P}\left(|X_1| > c\right) = 0$$

Proof. Let's prove

$$\varphi_{X_1}(0)' = ia \implies \overline{X}_n \xrightarrow{p} a$$

Suppose infact $(X_n)_{n\in\mathbb{N}}$ is iid and exists the first derivative in point 0 of the characteristic function. Then the characteristic function of the sample mean is:

$$\begin{aligned}\varphi_{\overline{X}_n}(t) &= \mathbb{E}\left[e^{it\overline{X}_n}\right] = \mathbb{E}\left[e^{i\frac{t}{n}\sum_{i=1}^n X_i}\right] = \varphi_{\sum_{i=1}^n X_i}\left(\frac{t}{n}\right) \overset{(\perp\!\!\!\perp)}{=} \prod_{i=1}^n \varphi_{X_i}\left(\frac{t}{n}\right) \overset{(1)}{=} \left[\varphi_{X_1}\left(\frac{t}{n}\right)\right]^n \\ &\overset{(2)}{=} \left[\varphi_{X_1}(0) + \frac{t}{n}\varphi_{X_1}(0)' + o\left(\frac{t}{n}\right)\right]^n \overset{(3)}{=} \left[\varphi_{X_1}(0) + \frac{t}{n}ia + o\left(\frac{t}{n}\right)\right]^n \\ &= \left[1 + \frac{ita + n\cdot o\left(\frac{t}{n}\right)}{n}\right]^n\end{aligned}$$

where

- in (1) equally distributed
- in (2) we apply Taylor up to the first order
- in 3 we substituted $\varphi_{X_1}(0)' = ia$ by hypothesis

Now we use the general fact that if $z_n, z \in \mathbb{C}$ and $z_n \to z$ then

$$\left(1 + \frac{z_n}{n}\right)^n \to e^z$$

Especially considering the limit for final result

$$\lim_{n\to\infty}\left[1 + \frac{ita + n\cdot o\left(\frac{t}{n}\right)}{n}\right]^n = e^{ita}$$

In our case we used $z_n = ita + n\cdot o\left(\frac{t}{n}\right) \to iat$. We also recall that

$$o(x) \text{ as } x \to x_0 \iff \lim_{x\to x_0}\frac{o(x)}{x} = 0$$

Hence

$$no\left(\frac{t}{n}\right) = t\frac{o\left(\frac{t}{n}\right)}{\frac{t}{n}} \to 0 \text{ as } n \to +\infty$$

To summarize

$$\varphi_{\overline{X}_n}(t) \to e^{iat}, \quad \text{as } n \to +\infty$$

but e^{iat} os the characteristic function of the degenerate rv $X = a$ a.s.
By property 4 of characteristic functions we get that $\overline{X}_n \xrightarrow{d} a$, but since a degenerate rv, we also get that sample mean converges to a not only in distribution but also in probability $X_n \xrightarrow{p} a$. ☐

Example 6.2.3. This following example exhibits that in case $(X_n)_{n\in\mathbb{N}}$ is iid

- we could have $\mathbb{E}[|X_1|] = a < +\infty$ (X_1 has the mean) and convergence of sample mean to a is almost sure (not only in probability, by the strong law of Kolmogorov);

- if however $\mathbb{E}[|X_1|] = +\infty$, the strong law fails (and we don't have almost sure convergence). But it may be the case that characteristic function has first derivative at 0 (so $\exists\varphi_X(0)'$), or the other equivalent condition on limits holds, and thus weak law of large numbers holds givin at least convergence in probability to a of sample mean.

Suppose that X_1 is absolutely continuous with density:

$$f(x) = \begin{cases} \frac{c}{x^2 \log|x|} & \text{if } |x| > 2 \\ 0 & \text{if } |x| \le 2 \end{cases}$$

where c is the normalizing constant (to make integral=1). Then:

$$\begin{aligned}\mathbb{E}[|X|] &= \int_{-\infty}^{+\infty} |x|\, f(x)\, \mathrm{d}x \overset{(1)}{=} 2\int_{0}^{+\infty} |x|\, f(x) \overset{(2)}{=} 2c\int_2^{\infty} \frac{x}{x^2\log x}\, \mathrm{d}x \\ &= 2c\int_2^{+\infty} \frac{1}{x\log x}\, \mathrm{d}x = +\infty\end{aligned}$$

where (1) because it's an even function and in (2) integral lower limit changes due to density
So this random variable does not have mean; hence, since (X_n) is iid the sequence $\overline{X}_n$ does not converge almost surely and the SLLN1 does not hold.
However X_1 is symmetric (since f is an even function). The condition

$$\lim_{b\to+\infty} \mathbb{E}[X_1 \cdot \mathbb{1}.(|X_1| \le b)] = a$$

is certainly true with $a = 0$ if X_1 is symmetric (recalling that Y is symmetric if $Y \sim -Y$) in fact

$$X_1 \text{ symmetric} \implies X_1 \cdot \mathbb{1}.(|X_1| \le b) \text{ is symmetric}$$

and if a symmetric rv Y has the mean then $\mathbb{E}[Y] = 0$

$$Y \sim -Y \implies \mathbb{E}[Y] = \mathbb{E}[-Y] = -\mathbb{E}[Y] \implies \mathbb{E}[Y] = 0$$

Hence

$$\mathbb{E}[X_1 \cdot \mathbb{1}.(|X_1| \le b)] = 0, \quad \forall b > 0$$

Moreover $\forall b > 2$

$$\begin{aligned}
b \cdot \mathbb{P}\left(|X_1| > b\right) &= b2\,\mathbb{P}\left(X_1 > b\right) = 2b \int_b^{+\infty} f(x)\,\mathrm{d}x \\
&= 2bc \int_b^{+\infty} \frac{1}{x^2 \log x}\,\mathrm{d}x \\
&\leq \frac{2bc}{\log b} \int_b^{+\infty} \frac{1}{x^2}\,\mathrm{d}x \\
&= \frac{2bc}{\log b} \left[-\frac{1}{x}\right]_b^{+\infty} = \frac{2bc}{\log b}\frac{1}{b} \\
&= \frac{2c}{\log b} \to 0 \quad \text{as } b \to +\infty
\end{aligned}$$

Hence we concluded that $\overline{X}_n \xrightarrow{p} a = 0$.

6.3 Central limit theorem

6.3.1 CLT

Remark 248. Big topic of probability, one of the main findings together with law of large numbers.
Here as well, there are several CLTs (all fullfill the following general definition).

Definition 6.3.1 (Central limit theorem)**.** A sequence $(X_n)_{n\in\mathbb{N}} = X_1, X_2 \ldots$ of real random variable satisfies the CLT if there are two constants $a_n \in \mathbb{R}$ and $b_n > 0$ such that

$$\frac{\sum_{i=1}^n X_i - a_n}{b_n} \xrightarrow{d} \mathrm{N}\left(0,1\right)$$

Important remark 55 (CLT of sum)*.* The sequence $(X_n)_{n\in\mathbb{N}}$ is arbitrary; one can think of sequence $X_1, X_2, \ldots$ as the sequence of observation.
To prove a CLT we need to find a_n and b_n for the ratio above to go in distribution to the standard normal.
The main/most important/natural special case is when:

$$\begin{aligned}
a_n &= \mathbb{E}\left[\sum_{i=1}^n X_i\right] \quad \text{mean of the sum} \\
b_n &= \sqrt{\mathrm{Var}\left[\sum_{i=1}^n X_i\right]} \quad \text{sd of the sum}
\end{aligned}$$

provided of course that such moment exist and $\mathrm{Var}\left[\sum_{i=1}^n X_i\right] > 0$.
Under these choices we have that the standardization of the sum fullfill the CLT definition

Remark 249 (Natural/tipical application of CLT)*.* Why CLT is so important in applications? Suppose we are interested in the distribution of $\sum_{i=1}^n X_i$ but we don't know how to evaluate it.

In this case, if CLT holds, we brutally replace such unknown distribution with $\mathrm{N}\left(a_n, b_n^2\right)$. Obviously, making this replacement we make an error. Howevere, thanks to the CLT, the error is expected to be small if n is large. In fact CLT implies that the distribution of standardized sample mean is close to standard normal

$$\frac{\sum_{i=1}^n X_i - a_n}{b_n} \sim \mathrm{N}(0,1)$$

Hence the distribution of the sum is close to

$$\sum_{i=1}^n X_i \sim a_n + b_n \, \mathrm{N}(0,1) = \mathrm{N}\left(a_n, b_n^2\right)$$

If I adopt the normal for a fixed n we surely make an error, the distribution is not normal: but the distribution becomes normal as n gets larger, and the error smaller.

Remark 250. Now we start with some examples of CLT: in case LLN the most important is Kolmogorov one, similarly in CLT the main/most popular statement of this kind is the so-called CLT1.

Proposition 6.3.1 (CLT1). *If $(X_n)_{n\in\mathbb{N}}$ is sequence of iid rvs with $\mathbb{E}\left[X_i^2\right] < +\infty$ (finite second moments) and X_i is not degenerate, then the standardized sum converges in distribution to standard normal*

$$\frac{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\sqrt{\mathrm{Var}\left[\sum_{i=1}^n X_i\right]}} = \frac{\sum_{i=1}^n X_i - n\,\mathbb{E}[X_i]}{\sqrt{n\,\mathrm{Var}[X_i]}} \xrightarrow{d} \mathrm{N}(0,1)$$

Remark 251. Thus in CLT1, we have $a_n = n\,\mathbb{E}[X_i]$ and $b_n^2 = n\,\mathrm{Var}[X_i]$.
Not also that

$$\frac{\sum_{i=1}^n X_i - n\,\mathbb{E}[X_i]}{\sqrt{n\,\mathrm{Var}[X_i]}} = \frac{\sum_{i=1}^n X_i - n\,\mathbb{E}[X_i]}{\sqrt{n}\sqrt{\mathrm{Var}[X_i]}} = \frac{\sqrt{n}}{\sqrt{\mathrm{Var}[X_i]}}\left(\overline{X}_n - \mu\right)$$

Proof. Let ϕ denote the characteristic function of the standardized single variable $\frac{X_1-\mu}{\sigma}$ where $\mu = \mathbb{E}[X_1], \sigma = \sqrt{\mathrm{Var}[X_1]}$ (here I can divide for standard deviation cause looking at the assumption, the rv is not degenerate so the variance is positive). In the following we need the following facts where we use the fact that expected value and variance of standardized variables are 0 while second moment is 1:

$$\phi'(0) = i\,\mathbb{E}\left[\frac{X_1 - \mathbb{E}[X_1]}{\sigma(X_1)}\right] = i \cdot 0 = 0$$
$$\phi''(0) = i^2\,\mathbb{E}\left[\left(\frac{X_i - \mathbb{E}[X_i]}{\sigma(X_i)}\right)^2\right] = -1 \cdot 1 = -1$$

moment 1. The standardized sum is

$$Z_n = \frac{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\sigma(\sum_{i=1}^n X_i)} \overset{(iid)}{=} \frac{\sum_{i=1}^n X_i - n\,\mathbb{E}[X_i]}{\sqrt{n\,\mathrm{Var}[X_1]}} = \frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{\sqrt{n\,\mathrm{Var}[X_1]}}$$
$$= \frac{1}{\sqrt{n}} \frac{\sum_{i=1}^n X_i - \mathbb{E}[X_i]}{\sigma(X_i)}$$

Its characteristic function is

$$\begin{aligned}
\varphi_{Z_n}(t) &= \varphi_{\frac{1}{\sqrt{n}}\frac{\sum_{i=1}^n X_i - \mathbb{E}[X_i]}{\sigma(X_i)}}(t) = \varphi_{\frac{\sum_{i=1}^n X_i - \mathbb{E}[X_i]}{\sigma(X_i)}}\left(\frac{t}{\sqrt{n}}\right) \\
&\stackrel{(1)}{=} \left[\varphi_{\frac{X_i - \mathbb{E}[X_i]}{\sigma(X_i)}}\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&\stackrel{(2)}{=} \left[\phi(0) + \frac{t}{\sqrt{n}}\phi'(0) + \frac{t^2}{n}\frac{1}{2}\phi''(0) + o\left(\frac{t^2}{n}\right)\right]^n \\
&= \left[1 + 0 - \frac{t^2}{n} + o\left(\frac{t^2}{n}\right)\right]^n \\
&= \left[1 + \frac{-t^2/2 + n \cdot o\left(\frac{t^2}{n}\right)}{n}\right]^n
\end{aligned}$$

where

- in (1) by iid: by independence the characteristic function of the sum is the product of the char function, and being identically distributed we have the power.
- in (2) as in the weak LLN proof, we use that rv by assumption have second moment finite; so we can say that the its characteristic function is C^2 and we can apply Taylor expansion (up to the the second order). So by Taylor (with Peano remainder)
- in (3) we substituted using previous result regarding single standardized variable characteristic function. since second moment exists, it exist the first as well and in the previous step we did the substitution

Now, for $n \to +\infty$ the last term developed for $\varphi_{Z_n}(t)$ converges[1] to $e^{-t^2/2}$ which is the characteristic function of a standard normal

$$\left[1 + \frac{-t^2/2 + n \cdot o\left(\frac{t^2}{n}\right)}{n}\right]^n \to e^{-t^2/2}$$

and this concludes the proof. □

Remark 252. Let's see another version of CLT, among the several.

Proposition 6.3.2 (CLT2)**.** *If* $(X_n)_{n\in\mathbb{N}}$ *are independent, with* $\mathbb{E}\left[|X_i|^3\right] < +\infty$, $\mathbb{E}[X_i] = 0, \forall n$ *and the following strange expression holds:*

$$\frac{\sum_{i=1}^n \mathbb{E}\left[|X_i - \mathbb{E}[X_i]|^3\right]}{\left(\sum_{i=1}^n \mathbb{E}\left[(X_i - \mathbb{E}[X_i])^2\right]\right)^{\frac{3}{2}}} = \frac{\sum_{i=1}^n \mathbb{E}\left[|X_i|^3\right]}{\left(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right]\right)^{\frac{3}{2}}} \to 0$$

[1]Again using the fact that in general if $a_n \to a$ then $\left(1 + \frac{a_n}{n}\right)^n \to e^a$, and for us in it suffices to let $a_n = -\frac{t^2}{2} + n \cdot o\left(\frac{t^2}{n}\right) \to -\frac{t^2}{2}$

Then, as previously, the standardized sum converges to standard normal

$$\frac{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\sqrt{\operatorname{Var}\left[\sum_{i=1}^n X_i\right]}} = \frac{\sum_{i=1}^n X_i - n\,\mathbb{E}\left[X_i\right]}{\sqrt{n\operatorname{Var}\left[X_i\right]}} = \frac{\sum_{i=1}^n (X_i - \mathbb{E}\left[X_i\right])}{\sigma(\sum_{i=1}^n X_i)}$$
$$= \frac{\sum_{i=1}^n X_i}{\sigma(\sum_{i=1}^n X_i)} \xrightarrow{d} \mathrm{N}\left(0,1\right)$$

Remark 253. Here the conclusions is the same as the CLT1 but the main difference is in the assumption where we are not forced to assume that rvs are identically distributed, they're just independent.
In order for the thm to still works, we need to replace that assumption with the new strange condition (dont' try to attach a meaning to this condition: it's just a technical condition for the theorem to hold).
this second example is **useful because** it can be used when X_i are not identically distributed.

Important remark 56. Supponiamo che $(X_n)_{n\in\mathbb{N}}$ sia iid, con $\mathbb{E}\left[X_i^2\right] < +\infty$ e $\sigma = \operatorname{Var}\left[X_i\right] > 0$. Se vogliamo conoscere la distribuzione di $\sum_{i=1}^n X_i$, grazie al CLT1 possiamo comunque dire che

NB: Considerazione utile dalla triennale

$$\frac{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\sqrt{\operatorname{Var}\left[\sum_{i=1}^n X_i\right]}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathrm{N}\left(0,1\right)$$

con $\mu = \mathbb{E}\left[X_i\right]$ e $\sigma^2 = \operatorname{Var}\left[X_i\right]$. Quindi per n grande la distribuzione di mio interesse di $\sum_{i=1}^n X_i$

$$\sum_{i=1}^n X_i \xrightarrow{d} \mathrm{N}\left(n\mu, n\sigma^2\right)$$

Poi se ci interessa la media

NB: aggiuntina improvvisata mia

$$\frac{\sum_{i=1}^n X_i}{n} \xrightarrow{d} \frac{1}{n}\mathrm{N}\left(n\mu, n\sigma^2\right) = \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Per le proprietà di media e valore atteso

6.3.2 Examples

Example 6.3.1. Suppose $(X_n)_{n\in\mathbb{N}}$ are independent, all rvs with null mean ($\mathbb{E}\left[X_i\right] = 0, \forall i$), that variables are somehow bounded ($|X_i| \leq c, \forall i$) and $\operatorname{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n X_i^2 \to +\infty$.
We are interested in convergence in distribution of the standardized sum

$$Z_n = \frac{\sum_{i=1}^n X_i}{\sigma(\sum_{i=1}^n X_i)}$$

The tool to study convergence in distribution for sum/mean is clt; in these cases we use the second version because we didn't say they are identically distributed. The random variable are independent, their mean is zero, thus in order to conclude that $Z_n \xrightarrow{d} \mathrm{N}\left(0,1\right)$ is enough to verify that "strange condition" holds, that is

$$\frac{\sum_{i=1}^n \mathbb{E}\left[|X_i|^3\right]}{(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right])^{\frac{3}{2}}} \to 0$$

To answer we note that

$$|X_i|^3 = |X_i|\, X_i^2 \overset{(1)}{\leq} cX_i^2$$

with (1) by assumptions. Hence:

$$\begin{aligned}\frac{\sum_{i=1}^n \mathbb{E}\left[|X_i|^3\right]}{(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right])^{\frac{3}{2}}} &\leq \frac{\sum_{i=1}^n \mathbb{E}\left[c \cdot X_i^2\right]}{(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right])^{\frac{3}{2}}} = \frac{c(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right])}{(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right])^{\frac{3}{2}}} \\ &= \frac{c}{(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right])^{\frac{1}{2}}} \to \frac{c}{\infty} = 0\end{aligned}$$

where the denominator goes to $+\infty$ by assumption. So since the strange condition expression is upper bunded by 0, it goes to 0 as well. Hence $Z_n \xrightarrow{d} \mathrm{N}(0,1)$

Example 6.3.2. Suppose $(X_n)_{n\in\mathbb{N}}$ is iid with $\mathbb{E}[X_i] = 0$ and second moment $\mathbb{E}\left[X_i^2\right] = 1$, so variance $\mathrm{Var}[X_i] = 1$. We're interested in convergence in distribution of this ratio:

$$Z_n = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n X_i^2}}$$

We use CLT1 because of iid rvs. In fact Z_n can be written as (by dividing by $\sqrt{n}$ both numerator and denominator):

$$Z_n = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i}{\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}} = \frac{\sum_{i=1}^n X_i}{\sqrt{n}} \cdot \frac{1}{\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}}$$

Now

- by CLT1 $\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \xrightarrow{d} \mathrm{N}(0,1)$ (think mean 0 and variance 1)
- since $(X_n)_{n\in\mathbb{N}}$ are iid $(X_n^2)_{n\in\mathbb{N}}$ are iid as well. Moreover $\mathbb{E}\left[X_1^2\right] = 1 < \infty$ by assumption. Thus Kolmogorov's SLLN applied to (X_i^2) we have that

$$\frac{\sum_{i=1}^n X_i^2}{n} \xrightarrow{a.s.} \mathbb{E}\left[X_i^2\right] = 1$$

Hence

$$Z_n = \frac{\sum_{i=1}^n X_i}{\sqrt{n}} \cdot \frac{1}{\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}} \xrightarrow{d} \mathrm{N}(0,1) \cdot \frac{1}{\sqrt{1}} = \mathrm{N}(0,1)$$

NB: sta considerazione (mio anno) per ora la lascio, si sa mai

Remark 254. In the above example as in the proof of CLT1, among other things, we used that if X_n is iid

$$\frac{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\sigma(\sum_{i=1}^n X_i)} = \frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{\sqrt{n}\sigma(X_i)} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

where $\sigma = \sigma(X_i)$ and $\mu = \mathbb{E}[X_i]$. In many theorem we write the quantity in that way.
Now $\sqrt{n} \to +\infty$ while $\overline{X}_n - \mu \xrightarrow{a.s.} 0$ if X_n is iid (and the moment exists).

$$\underbrace{\frac{\sqrt{n}}{\sigma}}_{\to+\infty} \cdot \underbrace{(\overline{X}_n - \mu)}_{\xrightarrow{a.s.} 0} \xrightarrow{d} \mathrm{N}(0,1)$$

Example 6.3.3. Suppose $(X_n)_{n\in\mathbb{N}}$ independent, $X_i \in \{-1, 0, 1\}$

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

$\forall i$. Let's find conditions on the constant α_i under which

$$Z_n = \frac{\sum_{i=1}^n X_i}{\sigma(\sum_{i=1}^n X_i)} \xrightarrow{d} \mathrm{N}(0,1)$$

These rvs can take only three values. We have that:

$$\begin{aligned}
\mathbb{E}[X_i] &= 0 \cdot \mathbb{P}(X_i = 0) + 1 \cdot \mathbb{P}(X_i = 1) + (-1)\,\mathbb{P}(X_i = -1) = \frac{\alpha}{2} - \frac{\alpha}{2} = 0 \\
\mathbb{E}\left[X_i^2\right] &= 1 \cdot \mathbb{P}\left(X_i^2 = 1\right) + 0 \cdot \mathbb{P}\left(X_i^2 = 0\right) = \mathbb{P}(X_i = +1) + \mathbb{P}(X_i = -1) = \alpha_i \\
\operatorname{Var}[X_i] &= \mathbb{E}\left[X_i^2\right] - (\mathbb{E}[X_i])^2 = \alpha_i
\end{aligned}$$

Now since variables are independent and bounded ($|X_i| \leq c, \forall i$ if $c = 1$) by example 6.3.1 we can conclude that a sufficient condition for

$$Z_n = \frac{\sum_{i=1}^n X_i}{\sigma(\sum_{i=1}^n X_i)} \xrightarrow{d} \mathrm{N}(0,1)$$

provided that the sum $\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \to +\infty$. But since $\mathbb{E}\left[X_i^2\right] = \alpha_i$, we finally obtain

$$\sum_{i=1}^n \alpha_i \to +\infty \implies Z_n \xrightarrow{d} \mathrm{N}(0,1)$$

To prove that the converse holds, that is

$$Z_n \xrightarrow{d} \mathrm{N}(0,1) \implies \sum_{i=1}^n \alpha_i \to +\infty$$

Toward the contraddiction suppose that $\sum_{i=1}^n \alpha_i \not\to +\infty$ that is

$$\lim_{n\to\infty} \sum_{i=1}^n \alpha_i = \alpha < +\infty$$

Note that we are summing non negative constants α_i which are the variances of the random variables; given independence we can rewrite the sum as

TODO: qui ho iniziato a improvvisare perché le note non sono esplicite

$$\sum_{i=1}^n \alpha_i = \sum_{i=1}^n \operatorname{Var}[X_i] = \operatorname{Var}\left[\sum_{i=1}^n X_i\right]$$

Now consider just the non-standardized sum of random variables

$$\begin{aligned}
\sum_{i=1}^n X_i &= \sum_{i=1}^n X_i \cdot \frac{\sqrt{\sum_{i=1}^n \alpha_i}}{\sqrt{\sum_{i=1}^n \alpha_i}} = \sqrt{\sum_{i=1}^n \alpha_i} \cdot \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n \alpha_i}} \\
&= \underbrace{\sqrt{\sum_{i=1}^n \alpha_i}}_{\to\sqrt{\alpha}} \cdot \underbrace{\frac{\sum_{i=1}^n X_i}{\sqrt{\operatorname{Var}\left[\sum_{i=1}^n X_i\right]}}}_{Z_n \xrightarrow{d} \mathrm{N}(0,1)} \xrightarrow{d} \sqrt{\alpha}\,\mathrm{N}(0,1) = \mathrm{N}(0,\alpha)
\end{aligned}$$

So under the assumptions above, $\sum_{i=1}^{n} X_i$ go in distribution to a normal. But this is a contraddition: X_i can assume only integer values (0, 1, -1) and so will be the sum $\sum_{i=1}^{n} X_i \in \mathbb{Z}$, while normal has domain on $\mathbb{R}$.
If $\sum_{i=1}^{n} X_i \xrightarrow{d} S$ the limit S must satisfy $\mathbb{P}(S \in \mathbb{Z}) = 1$. Instead $\mathbb{P}(N(0, \alpha) \in \mathbb{Z}) = 0$ and we have a contradiction.
Thus

$$Z_n \xrightarrow{d} \mathrm{N}(0,1) \implies \sum_{i=1}^{n} \alpha_i \to +\infty$$

Example 6.3.4. Find

$$\lim_{n\to+\infty} \mathbb{P}\left(\chi_n^2 > n + 7\sqrt{n}\right)$$

where χ_n^2 denotes a chi-square with n degrees of freedom.
Toe evaluate such a limit, take an iid sequence $(Z_n)_{n\in\mathbb{N}}$ with $Z_i \sim \mathrm{N}(0,1)$. We know that $\chi_n^2 \sim \sum_{i=1}^{n} Z_i^2$; considered that $\mathbb{E}[Z_i] = 0, \mathrm{Var}[Z_i] = 1$, moments of Z_i^2 can be obtained as:

$$\mathrm{Var}[Z_i] = \mathbb{E}\left[Z_i^2\right] - (\mathbb{E}[Z_i])^2 = 1 \iff \mathbb{E}\left[Z_i^2\right] = 1$$
$$\mathrm{Var}\left[\chi_n^2\right] = 2n = \mathrm{Var}\left[\sum_{i=1}^{n} Z_i^2\right] \iff 2n = n \cdot \mathrm{Var}\left[Z_i^2\right] \iff \mathrm{Var}\left[Z_i^2\right] = 2$$

Thus $\mathbb{E}\left[Z_i^2\right] = 1$ and $\mathrm{Var}\left[Z_i^2\right] = 2$. Now, back to the probability one obtains:

$$\begin{aligned}\mathbb{P}\left(\chi_n^2 > n + 7\sqrt{n}\right) &= \mathbb{P}\left(\sum_{i=1}^{n} Z_i^2 > n + 7\sqrt{n}\right) = \mathbb{P}\left(\frac{\sum_{i=1}^{n} Z_i^2 - n}{\sqrt{2}\sqrt{n}} > \frac{7}{\sqrt{2}}\right) \\ &= \mathbb{P}\left(\frac{\sum_{i=1}^{n} Z_i^2 - n\,\mathbb{E}\left[Z_i^2\right]}{\sqrt{n}\sqrt{\mathrm{Var}\left[Z_i^2\right]}} > \frac{7}{\sqrt{2}}\right) = \mathbb{P}\left(\frac{\sum_{i=1}^{n} Z_i^2 - n\,\mathbb{E}\left[Z_i^2\right]}{\sqrt{\mathrm{Var}\left[\sum_{i=1}^{n} Z_i^2\right]}} > \frac{7}{\sqrt{2}}\right)\end{aligned}$$

this converges by CLT1, as $n \to +\infty$ to $1 - \Phi(\frac{7}{\sqrt{2}})$ where Φ denotes the distribution function of $\mathrm{N}(0,1)$.

Example 6.3.5. Given any sequence $(X_n)_{n\in\mathbb{N}}$ of real rvs, the *empirical distribution function* is

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}.(X_i \le t), \quad \forall t \in \mathbb{R}$$

If (X_n) is iid F_n can be regarded as an estimate of the distribution function common to the X_n based on the data $(X_1, \dots, X_n)$.
Suppose now that (X_n) is actually iid and denote by F the common distribution function of the X_n. Then

$$\sqrt{n}(F_n(t) - F(t)) \xrightarrow{d} \mathrm{N}(0, F(t)(1 - F(t))), \forall t \in \mathbb{R} : 0 < F(t) < 1$$

Fix infact one such t and define $Y_n = \mathbb{1}.(X_n \leq t)$. Since $(X_n)$is iid (Y_n) is still iid. Now some development we need after:

$$\begin{aligned}
\mathbb{E}[Y_i] &= \mathbb{E}[\mathbb{1}.(X_i \leq t)] = \mathbb{P}(X_i \leq t) = F(t) \\
\operatorname{Var}[Y_i] &= \mathbb{E}\left[Y_i^2\right] - (\mathbb{E}[Y_i])^2 = \mathbb{E}\left[\mathbb{1}.(X_i \leq t)^2\right] - F^2(t) \\
&\overset{(1)}{=} \mathbb{E}[\mathbb{1}.(X_i \leq t)] - F^2(t) = F(t) - F^2(t) = F(t)(1 - F(t)) \\
\mathbb{E}\left[\sum_{i=1}^n Y_i\right] &= n\,\mathbb{E}[Y_i] = nF(t) \\
\operatorname{Var}\left[\sum_{i=1}^n Y_i\right] &= n \operatorname{Var}[Y_i] = nF(t)(1 - F(t))
\end{aligned}$$

where in (1), $\mathbb{E}\left[\mathbb{1}.(X_i \leq t)^2\right] = \mathbb{E}[\mathbb{1}.(X_i \leq t)]$ since $\mathbb{1}.(X_i \leq t)$ is an indicator.
Finally, considering $\sqrt{n}\,[F_n(t) - F(t)]$, we can manipulate it a bit to see

$$\begin{aligned}
\sqrt{n}\,[F_n(t) - F(t)] &= \frac{n\,[F_n(t) - F(t)]}{\sqrt{n}} = \frac{nF_n(t) - nF(t)}{\sqrt{n}} = \frac{n \cdot \frac{1}{n}\sum_{i=1}^n \mathbb{1}.(X_i \leq t) - n \cdot \mathbb{E}[Y_i]}{\sqrt{n}} \\
&= \frac{\sum_{i=1}^n Y_i - \mathbb{E}\left[\sum_{i=1}^n Y_i\right]}{\sqrt{n}} \cdot \frac{\sqrt{\operatorname{Var}\left[\sum_{i=1}^n Y_i\right]}}{\sqrt{\operatorname{Var}\left[\sum_{i=1}^n Y_i\right]}} \\
&= \sqrt{n}\sqrt{F(t)(1 - F(t))} \cdot \frac{\sum_{i=1}^n Y_i - \mathbb{E}\left[\sum_{i=1}^n Y_i\right]}{\sqrt{n}\sqrt{n}\sqrt{F(t)(1 - F(t))}} \\
&= \sqrt{F(t)(1 - F(t))} \cdot \frac{\sum_{i=1}^n Y_i - \mathbb{E}\left[\sum_{i=1}^n Y_i\right]}{\sqrt{\operatorname{Var}\left[\sum_{i=1}^n Y_i\right]}}
\end{aligned}$$

by CLT1 thus this last goes to

$$\sqrt{F(t)(1 - F(t))} \cdot \mathrm{N}(0,1) = \mathrm{N}(0, F(t)(1 - F(t)))$$

Example 6.3.6. Supponiamo che $Y_n \sim \operatorname{Bin}(n,p)$; ora, poiché nel caso della binomiale si fanno estrazioni con reimmissione Y_n può essere scritta come somma di n variabili iid X_i

NB: Esempio dalla triennale

$$Y_n = \sum_{i=1}^n X_i$$

dove X_i è l'indicatrice "bianca alla prova i" e

$$\begin{cases} \mathbb{P}(X_i = 1) = p \\ \mathbb{P}(X_i = 0) = 1 - p \end{cases} \quad , \quad \forall i = 1, \dots, n$$

Quindi, ricordando che la media della binomiale è np e la varianza $np(1-p)$, la seguente quantità standardizzata

$$\frac{Y_n - np}{\sqrt{np(1-p)}} = \frac{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\sqrt{\operatorname{Var}\left[\sum_{i=1}^n X_i\right]}} \xrightarrow{d} \mathrm{N}(0,1)$$

In sintesi se $Y_n \sim \mathrm{Bin}(n,p)$ allora la quantità

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} \mathrm{N}(0,1)$$

e quindi per n grande, la distribuzione di $\frac{Y_n - np}{\sqrt{np(1-p)}}$ può approssimarsi con una $\mathrm{N}(0,1)$

6.3.3 Berry-Esseen theorem

Remark 255. The most important reason for the CLT is so popular is: we are interested in the distrivution of $\sum_{i=1}^n X_i$ but we are not able to evaluate it.
Hence, we replace such unknown distribution with $\mathrm{N}\left(a_n, b_n^2\right)$. If the CLT holds, namely if

$$\frac{\sum_{i=1}^n X_i - a_n}{b_n} \xrightarrow{d} \mathrm{N}(0,1)$$

the error we're making is small, for n large enough. Hence it is crucial to have a quantitative evaluation of the error.
Actually one of the reason of importance of CLT is the following thm which allows to evaluate the error we make in adopting the normal distribution for the sum of random variables.

Theorem 6.3.3 (Berry-Esseen Theorem). *Let's suppose condition of CLT1 plus existence of third moment holds, that is:*

- $(X_n)_{n\in\mathbb{N}}$ *is iid*
- X_i *is non degenerate*
- $\mathbb{E}\left[|X_i|^3\right] < +\infty$

Now consider the difference/error at point x*:*

$$\mathbb{P}\left(\frac{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\sqrt{\mathrm{Var}\left[\sum_{i=1}^n X_i\right]}} \leq x\right) - \Phi(x)$$

where Φ is distribution function of $\mathrm{N}(0,1)$. *By CLT1 the first term goes to standard normal $\Phi(x)$ so the above difference above goes to 0 as $n \to +\infty$. At finite n if we use standard normal we make an error, but this error is supped/bounded:*

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}\left(\frac{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\sqrt{\mathrm{Var}\left[\sum_{i=1}^n X_i\right]}} \leq x\right) - \Phi(x)\right| \leq \frac{c}{\sqrt{n}} \cdot \mathbb{E}\left[\left|\frac{X_i - \mathbb{E}[X]}{\sqrt{\mathrm{Var}[X_i]}}\right|^3\right]$$

where $c \in \left(0, \frac{1}{2}\right)$ (typically set it to 1/2), and note that the error we make does not depend on considered x.

Example 6.3.7. For instance if the assumption by Berry holds and $n = 100$, we can say that the error made at any point x is

$$\leq \frac{1}{2}\frac{1}{10}\,\mathbb{E}\left[\left|\frac{X_i - \mathbb{E}[X_i]}{\sqrt{\mathrm{Var}[X_i]}}\right|^3\right], \quad \forall x \in \mathbb{R}$$

Thus in practice to have a good estimate it's enough to know σ (or making assumption/educated guess).

Remark 256. One last remark on CLT: CLT1 allows to obtain some infos about speed of converge (also said convergence rate) in the Kolmogorov strong law of large numbers (the most important one). We see it below

NB: non fatto quest'anno?

Proposition 6.3.4. *Let's assume the condition of CLT1 and fix a sequence a_n of constants such that*

$$\frac{a_n}{\sqrt{n}} \to 0$$

Now by kolmogorov's strong law we can say that

$$\overline{X}_n - \mu \xrightarrow{a.s.} 0$$

where as before $\mu = \mathbb{E}[X_1]$. Moreover by CLT1 we have that

$$a_n(\overline{X}_n - \mu) = \frac{a_n}{\sqrt{n}}\sqrt{n}(\overline{X}_n - \mu)$$

and by assumption $\frac{a_n}{\sqrt{n}} \to 0$, while for CLT1 $\sqrt{n}(\overline{X}_n - \mu) \to \mathrm{N}\left(0, \sigma^2\right)$ where $\sigma^2 = \mathrm{Var}[X_i]$. Thus the product goes to 0

$$a_n(\overline{X}_n - \mu) \xrightarrow{p} 0$$

further it can be shown that one also obtains

$$a_n(\overline{X}_n - \mu) \xrightarrow{a.s.} 0$$

Remark 257. If we have only LLN we can say only $X_i - \mu \xrightarrow{a.s.} 0$; using clt we can say much more $a_n(\overline{X}_n - \mu) \xrightarrow{a.s.} 0$.

Example 6.3.8. If I take $a_n = \sqrt{n}/\log n$ we have that

$$\frac{a_n}{\sqrt{n}} = \frac{1}{\log n} \to 0$$

and i get that

$$\frac{\sqrt{n}}{\log n}(\overline{X}_n - \mu) \xrightarrow{a.s.} 0$$

but $\sqrt{n}/\log n \to +\infty$ and

$$\frac{\sqrt{n}}{\log n}(\overline{X}_n - \mu) \to 0$$

even if $(\overline{X}_n - \mu)$ is multiplied by something that goes to $+\infty$.

6.4 Additional topics

6.4.1 Borel-Cantelli lemma

Remark 258. Let $\{A_i\}_{i\in\mathbb{N}}$ be a sequence of events, (A_i is any subset of sample space $A_i \subset \Omega$, $A_i \in \mathscr{A}$). Then we can define two new events.

NB: remembering tha tersection means $\forall$ union means $\exists$)

Definition 6.4.1 (Limsup of the sequence)**.** Defined as:

$$\overline{\lim_n} A_i = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{+\infty} A_i = \{\omega \in \Omega : \forall n \geq 1, \exists i \geq n \text{ such that } \omega \in A_i\}$$
$$= \{\omega \in \Omega : \omega \in A_i \text{ for infinitely many } i\}$$

Example 6.4.1. For instance if Bologna plays every sunday, A_i is Bologna wins at time i: limsup is event that Bologna wins infinite number of games.

Remark 259. Note that

- $\overline{\lim}_n A_i$ is still an event and $\overline{\lim}_n A_i \in \mathscr{A}$
- $\overline{\lim}_n A_i$ is true if and only if infinitely many of the A_i are true

NB: liminf non fatto quest'anno

Definition 6.4.2 (liminf of the sequence)**.** Defined as

$$\underline{\lim_n} A_i = \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{+\infty} A_i \qquad = \{\omega \in \Omega : \exists n \geq 1 \text{ such that } \omega \in A_i, \forall i \geq n\}$$

Example 6.4.2. Eg liminf is event there is an n such that from n on, Bologna wins every time.

Remark 260. By the Demorgan Law the complement of the limsup is the liminf of the complement, and the two events are connected by this equation

$$\left(\overline{\lim_n} A_i\right)^c = \left(\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{+\infty} A_i\right)^c = \bigcup_{n=1}^{+\infty} \left(\bigcup_{i=n}^{+\infty} A_i\right)^c = \bigcup_{n=1}^{+\infty} \bigcap_{i=n}^{+\infty} A_i^c = \underline{\lim_n} A_i^c$$

Important remark 57. Borel-Cantelli lemma is a tool to evaluate the probability of the limsup $\mathbb{P}\left(\overline{\lim}_n A_n\right)$ under some assumptions.

Theorem 6.4.1 (Borel-Cantelli)**.** *If*

- $\sum_i \mathbb{P}(A_i) < +\infty$ *(that is converges) then the probability of the limsup is null:* $\mathbb{P}\left(\overline{\lim}_n A_i\right) = 0$*;*
- $\sum_i \mathbb{P}(A_i) = +\infty$ *(that is diverges) and the* A_i *are independent, then* $\mathbb{P}\left(\overline{\lim} A_i\right) = 1$.

Important remark 58 (Two remarks). Regarding Borel-Cantelli:

1. Why the series of probability *necessarily* converges or diverges (can't be oscillating)? This is because it's the limit of a partial sum of positive or null numbers (probabilities).

2. if $\sum_{i=1}^{n} \mathbb{P}(A_n) = +\infty$ but the A_n are not independent, the Borel-Cantelli lemma does not apply (it does not cover any possible situation).

Remark 261. Proof is relatively easy but instead of it we make some examples to appreciate the use of the lemma.

Example 6.4.3. Suppose we have a coin and we throw it infinitely many times (or an urn with white and black balls from which we make drawing with replacement); we assume that the probability of tail is constant, $\mathbb{P}(\mathtt{T}) = \alpha \in (0,1)$, indipendently from the past.
Under these assumptions, for the second point of Borel-Cantelli, we observe *any* finite string of heads and tails infinitely many time with probability 1.
For example fix a finite sequence, say `TTHHT`; define the random variable X_i equal to indicator of the event

$$X_i = \mathbb{1}.(\text{tail at throw } i)$$

We define also all non-overlapping sequences `TTHHT` below:

$$\begin{aligned}
A_1 &= \{X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0, X_5 = 1\}\\
A_2 &= \{X_6 = 1, X_7 = 1, X_8 = 0, X_9 = 0, X_{10} = 1\}\\
A_3 &= \{X_{11} = 1, X_{12} = 1, X_{13} = 0, X_{14} = 0, X_{15} = 1\}\\
&\dots
\end{aligned}$$

A_1 is the event where the string occurs at the first five trials; A_2 from trial 6 to 10 etc. Now we have that

- A_i are independent since defined them using different X_i (which are independent);
- for any A_i:

  $$\mathbb{P}(A_i) = \alpha \cdot \alpha \cdot (1-\alpha) \cdot (1-\alpha) \cdot \alpha = a^3(1-\alpha)^2 > 0$$

  Hence $\sum_{i=1}^{n} \mathbb{P}(A_n) = \sum_{i=1}^{n} a^3(1-\alpha)^2 = +\infty$ since is an infinite sum of positive constant.

Thus we met Borel-Cantelli (second point) requirements and one can conclude that

$$\mathbb{P}\left(\overline{\lim} A_i\right) = 1$$

and thus

$$\mathbb{P}(\text{observe } \mathtt{TTHHT} \text{ infinitely many times}) \geq \mathbb{P}\left(\overline{\lim} A_n\right) \overset{(1)}{=} 1$$

Il $\geq$ presumo perché con gli A_i stiamo solo considerando eventi non overlappanti.

Example 6.4.4. Thanks to Borel-Cantelli it's simple to build an example where $X_n \xrightarrow{L_1} X$ but $X_n \not\xrightarrow{a.s.} X$ Take any sequence A_i of *independent* events such that $\mathbb{P}(A_i) = \frac{1}{i}$ and define

$$X_i = \mathbb{1}.(A_i) = \begin{cases} 1 & \text{if } A_i \text{ is true} \\ 0 & \text{if } A_i \text{ is false} \end{cases}$$

Let $X = 0$ be degenerate. Now:

- we have that $X_i \xrightarrow{L_1} 0$ since:

$$\mathbb{E}\left[|X_i - X|\right] = \mathbb{E}\left[|X_i - 0|\right] = \mathbb{E}\left[|X_i|\right] = \mathbb{E}\left[\mathbb{1}.(A_i)\right] = \mathbb{P}\left(A_i\right) = \frac{1}{i} \to 0$$

- let's see that it $X_i \not\xrightarrow{a.s} X$ does not converge almost surely. Note that
  - A_i are independent by assumption
  - furthermore

  $$\sum_i \mathbb{P}\left(A_i\right) = \sum_{i=1}^{\infty} \frac{1}{i} \overset{(1)}{=} +\infty$$

  being (1) the armonic series.

  Hence by Borel-Cantelli we can say that $\mathbb{P}\left(\overline{\lim_n} A_i\right) = 1$.
  On the other hand the complement events A_i^c

  - are independent (if A_i are independent the complements are still independent)
  - still

  $$\sum_i \mathbb{P}\left(A_i^c\right) = \sum_i \frac{i-1}{i} = +\infty$$

  Thus even here $\mathbb{P}\left(\overline{\lim_n} A_i^c\right) = 1$.
  It follows that the intersection of two almost sure events is still almost sure, that is:

  $$\mathbb{P}\left(\overline{\lim_n} A_i \cap \overline{\lim_n} A_i^c\right) = 1$$

  Now

  - fix an ω in this intersection $\omega \in (\overline{\lim_n} A_i \cap \overline{\lim_n} A_i^c)$.
  - since $\omega \in \overline{\lim_n} A_i$, $X_i(\omega) = \mathbb{1}_{A_i}(\omega) = 1$ for infinitely many n
  - similarly Since $\omega \in \overline{\lim_n} A_i^c$, $X_i(\omega) = \mathbb{1}_{A_i}(\omega) = 0$ for infinitely many n

  Hence $X_i(\omega)$ does not converge to any limit, so X_i does not converge almost surely.

Example 6.4.5. Let $(X_i)_{i\in\mathbb{N}}$ be iid rvs and suppose X_i is non degenerate. Under these assumption:

$$\mathbb{P}\left(X_i \text{ converges to a finite limit}\right) = 0$$

It's intuitive: if every student in a classrom choose a random number from the same distribution, the sequence will not converge to something; let's prove it formally.
Since X_i is non degenerate it can be shown (take this as given) that there are two numbers a, b with $a < b$ such that

$$\mathbb{P}\left(X_i \leq a\right) > 0 \vee \mathbb{P}\left(X_i \geq b\right) > 0$$

Now we define two events

$$A_i = \{X_i \leq a\},$$
$$B_i = \{X_i \geq b\}$$

What is the probability of limsup of A_i? We have that

- A_i are independent (being X_i independent)
- being identically distributed we have that:

$$\sum_i \mathbb{P}(A_i) = \sum_i \mathbb{P}(X_i \leq a) \stackrel{(1)}{=} +\infty$$

  with (1) because summing the same positive number infinite times

Hence $\mathbb{P}\left(\overline{\lim}_n A_i\right) = 1$.
By exactly the same arguments (B_i independent and with $\sum_i \mathbb{P}(B_i) = +\infty$) we conclude that $\mathbb{P}\left(\overline{\lim}_n B_i\right) = 1$.
Hence as before

$$\mathbb{P}\left(\overline{\lim_n} A_i \cap \overline{\lim_n} B_i\right) = 1$$

and then we fix ω in that intersection

$$\omega \in (\overline{\lim_n} A_i \cap \overline{\lim_n} B_i)$$

then $X_i(\omega)$ become a numerical sequence. Again this sequence does not converge:

- since $\omega \in \overline{\lim}_n A_i$, then $X_i(\omega) = X_i \leq a$ for infinitely many n
- otoh since $\omega \in \overline{\lim}_n B_i$, then $X_i(\omega) = X_i \geq b$ for infinitely many n

So having that $a < b$ the sequence can't converge to any limit and formally

$$\mathbb{P}(\omega \in \Omega : X_i(\omega) \text{does not converge}) = 1$$

Remark 262. An incidentally (related to Borel-Cantelli) useful fact is the following

Important remark 59. We have

- recall that for any sequence $(A_i)_{i \in \mathbb{N}}$

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

- in particular, if $\mathbb{P}(A_i) = 0, \forall i$ then

$$\mathbb{P}(\cup_i A_i) \leq \sum_i \mathbb{P}(A_i) = 0 \tag{6.1}$$

- if $\mathbb{P}(A_i) = 1, \forall i$ then then the

$$\mathbb{P}(\cap_i A_i) = 1$$

  Infact:

$$\mathbb{P}(\cap_i A_i) = 1 - \mathbb{P}((\cap_i A_i)^c) = 1 - \mathbb{P}(\cup_i A_i^c) \stackrel{(1)}{=} 1 - 0 = 1$$

  where (1) by applying the previous one, eq 6.1, since $\mathbb{P}(A_i^c) = 0, \forall i$

6.4.2 Stable rvs

Remark 263. It's an important type of random variables, together with infinite divisible rvs

Definition 6.4.3 (Stable rv)**.** A real rv X is said to be stable if exist an iid rvs sequence $\{X_i\}_{i\in\mathbb{N}}$ (with $X_i \sim Z$) and real constant a_n and $b_n > 0$ (actually real sequences i guess) such that

$$\frac{\sum_{i=1}^{n} X_i - a_n}{b_n} \xrightarrow{d} X$$

Example 6.4.6. $X \sim \mathrm{N}(0,1)$ is stable rv. In fact by CLT1, if Z is any non-degenerate rv with $\mathbb{E}\left[Z^2\right] < +\infty$ and if $(X_i)_{i\in\mathbb{N}}$ is iid with $X_i \sim Z$ then

$$\frac{\sum_{i=1}^{n} X_i - n\,\mathbb{E}[Z]}{\sqrt{n}\sqrt{\mathrm{Var}[Z]}} \xrightarrow{d} \mathrm{N}(0,1)$$

Hence it suffices to let $a_n = n\,\mathbb{E}[Z]$ and $b_n = \sqrt{n}\sqrt{\mathrm{Var}[Z]}$.

Remark 264. In general, the stable rvs are those rv which, like the normal can be obtained as the limit in distribution of the partial sums $\sum_{i=1}^{n} X_i$ (suitably normalized) of iid rvs.

NB: Qui la particolarità rispetto la definizione è $X_i \sim X$ e inoltre nello statement vi è direttamente $\sim$, non $\xrightarrow{d}$

Theorem 6.4.2 (Characterization)**.** *X is stable $\iff$ considering the sequence $\{X_i\}_{i\in\mathbb{N}}$ iid with $X_i \sim X$, $\forall n \geq 1$ there are real constants $\alpha_n \in \mathbb{R}$ and $\beta_n > 0$ such that:*

$$\frac{\sum_{i=1}^{n} X_i - \alpha_n}{\beta_n} \sim X$$

Remark 265. The idea of this theorem: given any rv X, take $X_1, \ldots, X_n$ iid with the same distribution as X, $X_i \sim X$. Then, in general, $\sum_{i=1}^{n} X_i \nsim X$. However, if X is stable we can find constants α_n, β_n such that the normalized partial sum has the same distribution X of the summed variables:

$$\frac{\sum_{i=1}^{n} X_i - \alpha_n}{\beta_n} \sim X$$

Example 6.4.7. By applying CLT1 we found $\mathrm{N}(0,1)$ is stable. We can show even with characterization theorem: in fact, if $(X_i)_{i\in\mathbb{N}}$ are iid with $X_i \sim \mathrm{N}(0,1)$ then

$$\sum_{i=1}^{n} X_i \sim \mathrm{N}(0,n)$$

so that

$$\frac{\sum_{i=1}^{n} X_i}{\sqrt{n}} \sim \frac{1}{\sqrt{n}}\mathrm{N}(0,n) = \mathrm{N}(0,1)$$

Hence the previous characterization applies with $\alpha_n = 0$ and $\beta_n = \sqrt{n}$

Remark 266. Other example of stable rvs are the Cauchy and degenerate.

Example 6.4.8 (Cauchy)**.** If X is Cauchy then the characteristic function of X is (take it as given)

$$\varphi_X(t) = e^{-|t|}, \quad \forall t \in \mathbb{R}$$

Now we have to verify the definition. Let's have $\{X_i\}_{i\in\mathbb{N}}$, with $X_i \sim \mathrm{Ca}(\cdot)$, and verify that, if we set $a_n = 0$, $b_n = n$ (obtaining the sample mean):

$$\frac{\sum_{i=1}^n X_i - a_n}{b_n} = \frac{\sum_{i=1}^n X_i}{n} = \overline{X}_n$$

the sample mean is distributed as Cauchy as well; we do this by checking its characteristic function:

$$\varphi_{\overline{X}_n}(t) = \varphi_{\frac{\sum_{i=1}^n X_i}{n}}(t) = \varphi_{\sum_{i=1}^n X_i}\left(\frac{t}{n}\right) \overset{(iid)}{=} \left[\varphi_{X_i}\left(\frac{t}{n}\right)\right]^n = \left[e^{-\left|\frac{t}{n}\right|}\right]^n$$
$$= e^{-|t|}$$

which is still Cauchy. Hence by taking $a_n = 0, b_n = n$ the quantity

$$\frac{\sum_{i=1}^n X_i - a_n}{n} = \overline{X}_n \sim Y_1$$

is Cauchy distributed and so by the characterization theorem Cauchy is example of stable rv.

Example 6.4.9 (Degenerate)**.** Another stable rv; if $\mathbb{P}(X = a) = 1$ for some constant a then trivially

$$\frac{a \cdot n}{n} = a$$

thus

$$\frac{\sum_{i=1}^n X_i - \alpha_n}{\beta_n} \sim X$$

with $X_i \sim \delta_a$, $\alpha_n = 0$ and $\beta_n = n$.

NB: bo secondo me qua ha sbagliato

Example 6.4.10. Let $\alpha \in (0, 2]$ and

$$\phi(t) = \exp\left(-\frac{|t|^\alpha}{2}\right), \quad \forall t \in \mathbb{R}$$

It can be shown that ϕ is the characteristic function of some rv X. Such X is actually stable.
In fact if $X_1, \dots, X_n$ are iid with $X_i \sim X$ then

$$\varphi_{\frac{\sum_{i=1}^n X_i}{n^{1/\alpha}}}(t) = \mathbb{E}\left[e^{i\frac{t}{n^{1/\alpha}}\sum_{i=1}^n X_i}\right] = \varphi_{\sum_{i=1}^n X_i}\left(\frac{t}{n^{1/\alpha}}\right) = \left[\varphi_{X_i}\left(\frac{t}{n^{1/\alpha}}\right)\right]^n$$
$$= \left(\exp\left[-\frac{1}{2}\left|\frac{t}{n^{1/\alpha}}\right|^\alpha\right]\right)^n = \left(\exp\left[-\frac{1}{2}\frac{|t|^\alpha}{n}\right]\right)^n$$
$$= \exp\left(-\frac{1}{2}|t|^\alpha\right)$$

This proves that

$$\frac{\sum_{i=1}^n X_i}{n^{1/\alpha}} \sim X$$

so that once again the previous characterization applies with $\alpha_n = 0$ and $\beta_n = n^{1/\alpha}$

6.4.3 Infinite divisible rvs

Remark 267. The infite divisible rvs are a remarkable subclass of the stable rvs. Losely speaking, a distribution is infinitely divisible if it can be expressed as the sum of an arbitrary number of independent and identically distributed (i.i.d.) random variables.

Definition 6.4.4 (Infinite divisible rv)**.** A real rv X is infinite divisible if and only if $\forall n \geq 1$, there are $X_{n_1}, \ldots, X_{n_n}$ iid rvs such that $\sum_{i=1}^{n} X_{n_i} \sim X$.

Remark 268. Important examples of infinitely divisible rvs are: the normal, the Cauchy, the degenerate, the Poisson, and the Gamma.

Proposition 6.4.3. *If X is stable then X is infinite divisible, but the viceversa does not holds. (so stable are a proper subset of infinite divisible)*

Proof. Infact

- if X is stable, we can write

$$X \sim \frac{\sum_{i=1}^{n} X_i - \alpha_n}{\beta_n}$$

  where $X_1, \ldots, X_n$ are iid. Hence letting

$$X_{n_i} = \frac{X_i - \alpha_n/n}{\beta_n}, \quad i = 1, \ldots, n$$

  one obtains $X_{n_1}, \ldots, X_{n_n}$ iid and

$$\sum_{i=1}^{n} X_{n_i} = \frac{\sum_{i=1}^{n} X_i - \alpha_n}{\beta_n} \sim X$$

- viceversa does not hold: by counterexample we need a infinite divisible which is not stable.
  It is sufficient to note that the only stable random variable X with finite variance/second moment ($\mathbb{E}\left[X^2\right] < \infty$) are the normal $\mathrm{N}\left(\mu, \sigma^2\right)$ and the degenerate.
  Hence if X is infinitely divisible, with finite variance, but neither degenerate nor normal, then X is an example of infinite divisible but not stable rv.
  Example of infinite divisible but not stable are the exponential (or the poisson): the exponential has the second moment but it is neither normal nor degenerate thus it's not stable; however as we noted before is infinite divisible.

□

Remark 269. The following result characterizes the infinite divisible rv having finite second moment.

Theorem 6.4.4. *We have that*

1. *X is infinite divisible and*

2. $\mathbb{E}\left[X^2\right] < +\infty$ *(has finite second moment)*

$\iff X \sim X_1 + X_2 + X_3$ *with* X_1, X_2, X_3 *independent,* X_1 *degenerate,* $X_2 \sim \mathrm{N}\left(0, \sigma^2\right)$ *and* X_3 *generalized Poisson.*

Remark 270. So this result describes the structure of infinite divisible random variables (with finite second moment).
In other words if X is infinitely divisible and $\mathbb{E}\left[X^2\right] < +\infty$ then X has the same disrtibution of the sum of 3 independent rvs, such that one is degenerate, the other is $\mathrm{N}\left(0,\ sigma^2\right)$ and the third is generalized Poisson. Let's see what is a generalized Poisson.

Definition 6.4.5 (Generalized poisson)**.** X is generalized poisson if

$$X \sim \mathbb{1}.(N > 0) \cdot \sum_{i=1}^{N} Z_i$$

where:

- $N \sim \mathrm{Pois}\left(\lambda\right)$
- (Z_i) is any iid sequence of rvs
- $N \perp\!\!\!\perp (Z_i)$

Important remark 60. We expect to find the poisson rv as a special case of this. Infact, if $Z_i = 1, \forall i$:

$$X \sim \mathbb{1}.(N > 0) \cdot N = N$$

but $N \sim \mathrm{Pois}\left(\lambda\right)$. So as expected the poisson is just special case of the generalized Poisson.
Note also that if φ denotes the characteristic function common to the Z_i the characteristic function of X can be written as

$$\begin{aligned}
\varphi_X(t) = \mathbb{E}\left[e^{itX}\right] &= \mathbb{E}\left[\mathbb{1}.(N = 0) \cdot 1 + \sum_{n=1}^{+\infty} \mathbb{1}.(N = n) e^{it \sum_{i=1}^{n} Z_i}\right] \\
&= \mathbb{P}\left(N = 0\right) + \sum_{n=1}^{+\infty} \mathbb{P}\left(N = n\right) \varphi_{\sum_{i=1}^{n} Z_i}(t) \\
&= \mathbb{P}\left(N = 0\right) + \sum_{n=1}^{+\infty} \mathbb{P}\left(N = n\right) \left[\varphi_{Z_i}(t)\right]^n \\
&= \sum_{n=0}^{+\infty} \mathbb{P}\left(N = n\right) \left[\varphi_{Z_i}(t)\right]^n \\
&= \frac{e^{-\lambda}\lambda^n}{n!} \left[\varphi_{Z_i}(t)\right]^n \\
&= e^{-\lambda} \underbrace{\sum_{n=0}^{+\infty} \frac{\left(\lambda\varphi_{Z_i}(t)\right)^n}{n!}}_{e^{\lambda\varphi_{Z_i}(t)}} \\
&= e^{-\lambda} e^{\lambda\varphi_{Z_i}(t)} = e^{\lambda\left(\varphi_{Z_i}(t) - 1\right)}
\end{aligned}$$

This latter is the characteristic function of a generalized poisson rv.

Theorem 6.4.5. *If X is infinite divisible and $\mathbb{P}(a \le X \le b) = 1$ for some a and b (X is bounded) then X is degenerate.*

Proof. Since X is infinite divisible $\forall n \ge 1$ we have $X \sim \sum_{i=1}^n X_{n_i}$, with $X_{n_1}, \dots, X_{n_n}$ iid. Thus:

$$\operatorname{Var}[X] = \operatorname{Var}\left[\sum_{i=1}^n X_{n_i}\right] = \sum_{i=1}^n \operatorname{Var}[X_{n_i}] = n \operatorname{Var}[X_{n_i}] \le n\,\mathbb{E}\left[X_{n_i}^2\right]$$

with last passage due to variance calculation formula.
Now since $\mathbb{P}(a \le X \le b) = 1$, we have that

- $\mathbb{P}\left(X_{n_i} > \frac{b}{n}\right) = 0$. Infact

  $$0 = \mathbb{P}(X > b) = \mathbb{P}\left(\sum_{i=1}^n X_{n_i} > b\right) \ge \mathbb{P}\left(X_{n_i} > \frac{b}{n}, \forall i\right) \overset{(iid)}{=} \left[\mathbb{P}\left(X_{n_i} > \frac{b}{n}\right)\right]^n$$

  and thus $\mathbb{P}\left(X_{n_i} > \frac{b}{n}\right) = 0$

- $\mathbb{P}\left(X_{n_i} < \frac{a}{n}\right) = 0$ by the same argument

Thus X_{n_i} stays between $\frac{a}{n}$ and $\frac{b}{n}$ almost surely

$$\mathbb{P}\left(\frac{a}{n} \le X_{n_i} \le \frac{b}{n}\right) = 1$$

and therefore

$$\mathbb{P}\left(|X_{n_i}| \le \frac{\max(|a|, |b|)}{n}\right) = 1$$

Given this last equation, if we square both terms we can write that the expected value (of the squared rv) is less than the squared "domain superior limit":

$$\mathbb{E}\left[X_{n_i}^2\right] \le \frac{\max(|a|, |b|)^2}{n^2}$$

Hence.

$$\operatorname{Var}[X] \le n\,\mathbb{E}\left[X_{n_i}^2\right] \le n\frac{\max(|a|, |b|)^2}{n^2} = \frac{\max(|a|, |b|)^2}{n}$$

and thus (being valid $\forall n \ge 1$ even an high value I guess)

$$\lim_{n \to +\infty} \frac{\max(|a|, |b|)^2}{n} = 0$$

thus X is degenerate. □

6.4.4 Examples

NB: non fatto quest'anno

Example 6.4.11 (Poisson). $X \sim \operatorname{Pois}(\lambda)$ is infinite divisible. Infact, if $Y_1, \dots, Y_n$ are independent and $Y_i \sim \operatorname{Pois}(\lambda_i)$ then $\sum_{i=1}^n Y_i \sim \operatorname{Pois}(\sum_{i=1}^n \lambda_i)$.
Hence if $X \sim \operatorname{Pois}(\lambda)$, if is sufficient to take $X_{n_1}, \dots, X_{n_n}$ iid rvs with $X_{n_i} \sim \operatorname{Pois}\left(\frac{\lambda}{n}\right)$

non fatto quest'anno **Example 6.4.12** (Normal)**.** $\mathrm{N}\left(\mu, \sigma^2\right)$ is infinite divisible. Infact if $X_1, \ldots, X_n$ idependent, $\mathrm{N}\left(\mu_i, \sigma_i^2\right)$, then $\sum_{i=1}^n X_i \sim \mathrm{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$

non fatto quest'anno **Example 6.4.13** (Gamma)**.** Another example is the gamma. In fact $X \sim \mathrm{Gamma}\left(\alpha, \beta\right)$ iff X is absolutely continuous with density

$$f(x) = \begin{cases} \frac{\alpha^\beta}{\Gamma(\beta)} e^{-\alpha x} x^{\beta 1} & x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Note for $\beta = 1$ we get the $\mathrm{Gamma}\left(\alpha, 1\right) = \mathrm{Exp}\left(\alpha\right)$ so exponential is a special case of gamma.
Now if $Y_1, \ldots Y_n$ indep and $Y_i \sim \mathrm{Gamma}\left(\alpha, \beta_i\right)$ (with common α) then the sum of Y_i is still a gamma, that is $\sum_{i=1}^n Y_i \sim \mathrm{Gamma}\left(\alpha, \sum_{i=1}^n \beta_i\right)$.
By the way, if $Y_1, \ldots Y_n$ are iid $Y_i \sim \mathrm{Exp}\left(\alpha\right)$, then the distribution of the sum $\sum_{i=1}^n Y_i = \mathrm{Gamma}\left(\alpha, n\right)$ (n because $\beta = 1$ and the sum is n).
Using the above results it follows that Gamma is infinite divisible.