PT = SOLID FOUNDATIONS

↘ The Ancient Greeks have no mathematical description of randomness. Their approach is mostly philosophical, and differs from one philosopher to another. Democritus (-400) believes in determinism, and argues that chance comes from a lack of knowledge of humans to understand events. Aristotle (-350) views randomness as part of the world. He classifies events in three categories: certain, probable (events that will occur unless an exceptional circumpstance happens), and unknowable (events that cannot be predicted in advance). Epicurus (-300) argues that randomness exists by itself, as a result of the intrinsic randomness present in atoms. [⚠ the concept of "atoms" in Ancient Greece differs from the modern definition; the Greek atom is simply a fundamental, indivisible constituent of the world]

Are you an Epicurian in your approach of understanding randomness in the world you live in?

↘ No more development of randomness in Europe until the XVI^{e}, with the Italian mathematician Cardano who was interested in gambling. In his book Liber de ludo alae (Book on Games of Chance), he is one of the first to associate a fraction with an uncertain event, representing its probability of occurence. Dealing exclusively with random events containing only finitely many outcomes, such as in the case of the cast of a die, the fraction is simply the ratio of the number of favorable outcomes, to the total number of possible outcomes.

↘ These ideas were pushed futher with Pascal, Fermat, Bernoulli (XVII^{e}), and then De Moivre (XVIII^{e}).

↘ At this stage, PT is approached from the point of view of relative frequencies: Bernoulli first introduced the Law of Large Numbers, stipulating that if an experiment is repeated several times in the same conditions then the relative frequency of occurence of some event A converges to a number between 0 and 1 representing the probability of occurence of A.

↗ This is a law of nature; and what makes PT possible.

↘ We refer to a RANDOM EXPERIMENT (RE) as a real-life phenomena that

(i) has a "mass character"; that is, can be repeated many times.

(ii) does not display any "deterministic regularity".

(iii) however, possesses "statistical regularity": relative frequences of events

$$\frac{n_A}{n} = \frac{\#\text{ times that A occured}}{\text{total }\#\text{ of trials}}$$

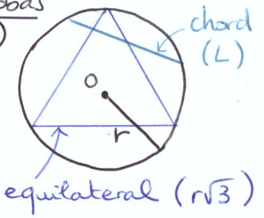
stabilize around some value in $[0,1]$ as the number of "independent" repetitions of the RE increases.

↖ How can we formalize mathematically the notion of independence?

↘ We had to wait Kolmogorov (XXe s) with his axiomatic approach to PT to be able to deal with more complex (and interesting) random events. Indeed, mathematicians so far would consider only events with a finite number of outcomes; but how would you decide for example if a bunch of points spread on a sheet of paper are "randomly" / "uniformly" distributed? (or not...)

When talking about probabilities associated with infinitely (uncountably) many outcomes, traditional counting methods fail. Bertrand's paradox is a famous example illustrating the fact that care must be taken when talking about "at random" or "equally likely" in such cases.

2 Bertrand's paradox (Calcul des probas 1889)



Given a circle of radius r and centre O, choose a chord randomly and denote by L its length. What is the probability $p = \mathbb{P}(L > r\sqrt{3})$ that its length is longer than the side of an inscribed equilateral triangle?

The answer depends on what you understand by choosing the chord "at random".

(i) Given a radius, choose a point J uniformly on it. The chord chosen passes through J and is perpendicular to the radius.

Answer: $p = 1/2$ [Give heuristics]

(ii) Choose uniformly and independently two points A and B, and select the chord AB.

Answer: $p = 1/3$. [Heuristics, please]

(iii) Choose the middle I of the chord uniformly inside the circle.

Answer: $p = 1/4$. [Heuristics, please]

↳ Your conclusion? Do we really have a paradox?

In Bertrand's paradox, if you specify the method used for selecting the chord, relative frequencies will converge to the associated probability p. However, by simply using information contained in the problem statement ("at random"), all three procedures are equally valid and the relative frequency approach will fail.

↘ Kolmogorov proposes a new paradigm: instead of writing directly what probabilities are, Kolmogorov suggests starting with which properties probabilities should have.

impossible/not tractable, as you have uncountably many events

Kolmogorov approach is axiomatic: a general approach in mathematics. Even the Ancient Greeks were doing geometry using axioms.

↳ This is the approach we consider here. Without going too far into theoretical details, we now introduce the main ingredients needed to appreciate Kolmogorov's axioms.

I_ AXIOMS OF PROBABILITY THEORY

To build Kolmogorov's probability model, the first ingredient answers the question: "What are all the things that could possibly happen in our random experiment?". We formalize this by specifying a set Ω, referred to as the SAMPLE SPACE = space of all possible outcomes.

Elements $\omega \in \Omega$ symbolize possible OUTCOMES of the RE.

Examples (i) Coin flip H/T (head / tail)

$\Omega = \{H, T\}$

↳ What if where the coin landed is also of interest to us?

⇒ Several Ωs are possible.

(ii) Cast of a die $\Omega = \{1, 2, 3, 4, 5, 6\}$

(iii) Position of a particle in a fluid $\Omega = \mathbb{R}^3$

(iv) Random motion of a particle in a fluid

$\Omega = \mathcal{C}([0, \infty), \mathbb{R}^3)$

= space of continuous functions $[0, \infty) \to \mathbb{R}^3$

In example (ii), the question "did we throw a 2" is represented by the subset $\{2\}$; the question "did we throw an odd number" by $\{1, 3, 5\}$. More generally, to answer questions about our RE, we use subsets $A \subset \Omega$.

Let $\mathcal{F}$ = collection of subsets of Ω.

↳ We want to answer yes/no questions to elements of $\mathcal{F}$. It turns out that no every $\mathcal{F}$ qualifies.

If asking A or B is sensible, then asking A or B $(A \cup B)$, or A and B $(A \cap B)$ should also be sensible. The complement of A (denoted A^c or $\bar{A}$) should also be sensible. In fact, we need a bit more: if $A_1, A_2, \ldots$ are sensible, then $\bigcup_{n=1}^{\infty} A_n$ should also be sensible.

↳ Convince yourself: with $\Omega = \mathbb{N} = \{1, 2, 3, \ldots\}$, and $A_n = \{n\} \in \mathcal{F}$ $\forall n$, then it would be weird not to be able to answer questions such as "Is it an even number" $= \{2n \mid n \in \mathbb{N}\}$.

A collection of objects satisfying these requirements is called a σ-ALGEBRA.

↑ In mathematics, the symbol σ- indicates stability under countable unions (as opposed to finite unions).

Definition. A family $\mathcal{F}$ of subsets of Ω is said to be a σ-ALGEBRA on Ω if

(i) $\Omega \in \mathcal{F}$

(ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$

(iii) If $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

Immediate consequence: $\mathcal{F}$ is closed under countable intersections:

$$\bigcap_{n=1}^{\infty} A_n = \left[\left(\bigcup_{n=1}^{\infty} A_n^c \right) \right]^c .$$

Also, • $\emptyset \in \mathcal{F}$ since $\emptyset = \Omega^c$.

• Taking $A_3 = A_4 = \ldots = \emptyset$, (iii) yields $A_1 \cup A_2 \in \mathcal{F} \Rightarrow \mathcal{F}$ is closed under finite unions.

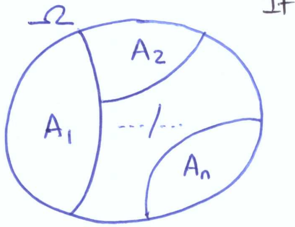
Examples (i) Let Ω be any set. The power set

$$\mathcal{P}(\Omega) = \{ A \mid A \subset \Omega \}$$
$$= \text{set of all subsets of } \Omega$$

is a σ-algebra.

(ii) Trivial σ-algebra $\mathcal{F} = \{\emptyset, \Omega\}$ (no fun)

(iii) A (slightly) more interesting σ-algebra = collection of sets containing a specific A. The smallest such σ-algebra, denoted $\sigma(A)$, is called the "σ-algebra generated by A", and is given by $\sigma(A) = \{\emptyset, A, A^c, \Omega\}$.



It can easily be extended to

$\mathcal{G} = \{A_1, \dots, A_n\}$ = finite PARTITION of Ω

Then

$$\sigma(\mathcal{G}) = \left\{ \bigcup_{i \in I} A_i \;\middle|\; I \subset \{1, \dots, n\} \right\}$$

The smallest one. Its elements can be enumerated easily since all intersections are $\emptyset$.

What if $\mathcal{G}$ is an arbitrary family of subsets of Ω?

Theorem. Let $\mathcal{G}$ = family of subsets of Ω. There exists a unique σ-algebra, denoted $\sigma(\mathcal{G})$, called the σ-algebra generated by $\mathcal{G}$, such that

(i) $\mathcal{G} \subset \sigma(\mathcal{G})$

(ii) If $\mathcal{H}$ is a σ-algebra on Ω and $\mathcal{G} \subset \mathcal{H}$, then $\sigma(\mathcal{G}) \subset \mathcal{H}$

$\sigma(\mathcal{G})$ is thus the smallest one.

3 Let $\{\mathcal{F}_\lambda\}_{\lambda \in \Lambda}$ be a collection of σ-algebras on Ω.

(i) Show that $\mathcal{F} = \bigcap_{\lambda \in \Lambda} \mathcal{F}_\lambda$ is also a σ-algebra on Ω.

(ii) Show that $\bigcup_{\lambda \in \Lambda} \mathcal{F}_\lambda$ is not a σ-algebra.

(iii) Prove the theorem on page 7.

(iv) Let $\Omega = \mathbb{R}$ be the set of real numbers. The BOREL σ-ALGEBRA is the σ-algebra generated by all open intervals of $\mathbb{R}$. It is donoted $\mathcal{B}(\mathbb{R})$.

Immediate consequences: Since all $A_n = (x-n, x) \in \mathcal{B}(\mathbb{R})$ then

$$A = \bigcup_{n \geq 1} A_n = (-\infty, x) \in \mathcal{B}(\mathbb{R})$$
$$A^c = [x, \infty) \in \mathcal{B}(\mathbb{R})$$

Moreover, $B_n = (y, y+n) \in \mathcal{B}(\mathbb{R})$, hence

$$B = \bigcup_{n \geq 1} B_n = (y, \infty) \in \mathcal{B}(\mathbb{R})$$
$$B^c = (-\infty, y] \in \mathcal{B}(\mathbb{R}).$$

Thus, $A^c \cap B^c = [x, y] \in \mathcal{B}(\mathbb{R})$

Closed intervals are also in $\mathcal{B}(\mathbb{R})$. Note that $\mathcal{B}(\mathbb{R})$ is also the σ-algebra generated by all closed intervals of $\mathbb{R}$.

Finally, $\{x\} \in \mathcal{B}(\mathbb{R})$. Take for instance $C_n = (x - \frac{1}{n}, x + \frac{1}{n})$; $\{x\} = \bigcap_{n \geq 1} C_n$.

When working with $\mathbb{R}$, it is customary to take $\mathcal{B}(\mathbb{R})$, and not $\mathcal{P}(\mathbb{R})$. This is a technical point, but an important one. The set $\mathcal{P}(\mathbb{R})$ is too big to construct a consistent theory of probability on it. Recall that elements of the σ-algebra are considered to answer "yes-no" questions. Soon, we will use these sets to assign them probabilities. It turns out that if we were working with $\mathcal{P}(\mathbb{R})$, it would be impossible to assign probabilities to all $A \in \mathcal{P}(\mathbb{R})$ in a way that the result is a uniform distubution on $[0,1]$. This is rather problematic. The set $\mathcal{B}(\mathbb{R})$ is rich enough for practical purposes, and does not lead to such undesirable properties.

The pair $(\Omega, \mathcal{F})$ is called a <u>MEASURABLE SPACE</u>. These are the two ingredients needed to define a probability to elements of $\mathcal{F}$. This must be done in a consistent way.

Foundations of modern PT (Kolmogorov '33)

A probability on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ s.t

(i) $\forall A \in \mathcal{F}, \quad \mathbb{P}(A) \geq 0$

(ii) $\mathbb{P}(\Omega) = 1$

(iii) For any pairwise disjoint $A_1, A_2, \ldots \in \mathcal{F}$,

$$\mathbb{P}\Big(\bigcup_{n\geq 1} A_n\Big) = \sum_{n\geq 1} \mathbb{P}(A_n)$$

AXIOMS

"countable additivity"

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a <u>PROBABILITY SPACE</u>.

The first and second axiom ensure that events have non-negative probabilities, and that a certain event has probability one. The third axiom is motivated from the frequency interpretation

of probability: if A and B are two disjoint events,

n_A = # times A occurs out of n trials.

n_B = # — B ——— n ——

$n_{A\cup B}$ = # — A or B ——— n ——,

then
$$\frac{n_{A\cup B}}{n} = \frac{n_A}{n} + \frac{n_B}{n} \approx \mathbb{P}(A\cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

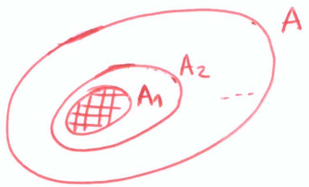
Finite additivity.

Kolmogorov requires a bit more than finite additivity: the probability $\mathbb{P}$ should also be <u>countably additive</u>. This is fundamental not only for the interpretation of the theory, but it will also allow us to take limits.

Indeed, if we only assume

(iii) For any finite collection $A_1, \ldots, A_n \in \mathcal{F}$ of pairwise disjoint events, $\mathbb{P}\big(\bigcup_{i=1}^{n} A_i\big) = \sum_{i=1}^{n} \mathbb{P}(A_i)$.

then we would loose nice continuity properties, such as:



$A_n \uparrow A \Rightarrow \mathbb{P}(A_n) \uparrow \mathbb{P}(A)$

or

$A_n \downarrow A \Rightarrow \mathbb{P}(A_n) \downarrow \mathbb{P}(A)$

In particular,

$A_n \downarrow \emptyset \Rightarrow \mathbb{P}(A_n) \downarrow 0$

<u>Example</u>: Let
- Ω = finite set
- $\mathcal{F} = \mathcal{P}(\Omega)$ = power set of Ω.

Then any probability measure on Ω can be constructed as follows

↘ $\forall \omega \in \Omega$, assign a number $\mathbb{P}(\{\omega\})$ such that $\sum_{\omega\in\Omega} \mathbb{P}(\{\omega\}) = 1$

↘ $\mathbb{P}$ can be extended to all $\mathcal{F}$ using the additive property:

$$\forall A \in \mathcal{F} \qquad \mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}).$$

This example shows the basic idea stated on page 4: to define $\mathbb{P}$, it is not necessary to specify $\mathbb{P}(A)$ for all $A \in \mathcal{F}$. The axioms of probability dictate how $\mathbb{P}$ should behave, and it is enough. Note that this is clean if Ω is finite. For $\Omega = \mathbb{R}$, it is trickier. We return to this point later.

For Ω finite, and equally likely ω, we have

$$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} \quad ; \quad \forall A \in \mathcal{F} \quad \mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

"classical scheme".

- Some elementary properties.

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

Taking $A_1 = A_2 = \dots = \emptyset$ in (iii), we have

$$\mathbb{P}(\emptyset) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n\geq}^{\infty} \mathbb{P}(\emptyset) \quad \Rightarrow \text{We must have } \mathbb{P}(\emptyset)=0$$

(b) Finite additivity $\mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} \mathbb{P}(A_j)$

Taking $A_{n+1} = A_{n+2} = \dots = \emptyset$ in (iii), we have

$$\begin{array}{ccc} \mathbb{P}\left(\bigcup_{j=1}^{\infty} A_j\right) & = & \mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) \\ \overset{(iii)}{\parallel} & & \\ \sum_{j=1}^{\infty} \mathbb{P}(A_j) & \underset{\uparrow}{=} & \sum_{j=1}^{n} \mathbb{P}(A_j) \end{array}$$

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

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

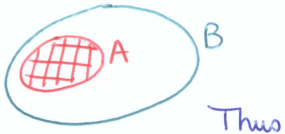
Special case of (b) with $A_1 = A$ and $A_2 = A^c$.

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(\Omega) \overset{(ii)}{=} 1$$

$$\parallel$$

$$\mathbb{P}(A_1) + \mathbb{P}(A_2) = \mathbb{P}(A) + \mathbb{P}(A^c).$$

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



$$\left.\begin{array}{l} A \cup (A^c \cap B) = B \\ A \cap (A^c \cap B) = \emptyset \end{array}\right\} \text{disjoint}$$

Thus

$$\underset{\mathbb{P}(B)}{\underset{\parallel}{\mathbb{P}(A \cup (A^c \cap B))}} = \mathbb{P}(A) + \underbrace{\mathbb{P}(A^c \cap B)}_{\geq 0}$$

$$\Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B).$$

(e) Subadditivity. For any $A_1, A_2, \dots \in \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) \leqslant \sum_{n\geq 1} \mathbb{P}(A_n).$$

Construct disjoint sets B_n such that $\bigcup_{n\geq 1} B_n = \bigcup_{n\geq 1} A_n$:

→ $B_1 = A_1 \qquad \subset A_1$

→ $B_2 = A_2 \setminus A_1 \qquad \subset A_2$

→ $B_3 = A_3 \setminus (A_1 \cup A_2) \subset A_3 \qquad \dots / \dots$

From (d), we have that $\mathbb{P}(B_k) \leq \mathbb{P}(A_k)$

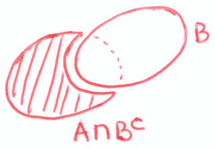
We obtain

$$\mathbb{P}\left(\bigcup_n A_n\right) = \mathbb{P}\left(\bigcup_n B_n\right) = \sum_n \mathbb{P}(B_n) \leq \sum_n \mathbb{P}(A_n)$$

(f) $\forall A, B \in \mathcal{F}, \quad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

$$A \cup B = B \cup (A \cap B^c)$$

disjoint



$$\mathbb{P}(A \cup B) = \mathbb{P}(B) + \boxed{\mathbb{P}(A \cap B^c)}$$

Now, $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$

since $A = (A \cap B) \cup (A \cap B^c)$ [disjoint]

Thus

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

(g) For $A_1 \subset A_2 \subset \dots \in \mathcal{F}$, $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{k\geq1} A_k\right)$

- First note that if $A \subset B$, then $\mathbb{P}(B\setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$. Indeed, A and $B \setminus A = B \cap A^c$ are disjoint, and such that $A \cup (B\setminus A) = B$. Thus $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B\setminus A) \Rightarrow \mathbb{P}(B\setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$

- Since $A_1 \subset A_2 \subset \dots$, we have that $\bigcup_{j=1}^{n-1} A_j = A_{n-1}$

  Putting
  - $A_0 = \emptyset$
  - $B_n = A_n \setminus A_{n-1}$, we see that the B_n are disjoint, and such that $\bigcup_{j=1}^{n} B_j = \bigcup_{j=1}^{n} A_j$.

  Moreover, $\mathbb{P}(B_n) = \mathbb{P}(A_n) - \mathbb{P}(A_{n-1})$

- Thus,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{k\geq1} A_k\right) = \mathbb{P}\left(\bigcup_{k\geq1} B_k\right) &= \sum_{k\geq1} \mathbb{P}(B_k) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(B_k) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} \{\mathbb{P}(A_k) - \mathbb{P}(A_{k-1})\} \\
&= \lim_{n\to\infty} \mathbb{P}(A_n)
\end{aligned}$$

telescoping sum

(h) For $A_1 \supset A_2 \supset \dots \in \mathcal{F}$, $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcap_{k\geq1} A_k\right)$

Use (g) and complementation $\bigcap_k A_k = \left[\bigcup_k A_k^c\right]^c$

Consequence:
- If $\forall n$ $\mathbb{P}(A_n) = 0$, then $\mathbb{P}\left(\bigcup_{n\geq1} A_n\right) = 0$
- If $\forall n$ $\mathbb{P}(A_n) = 1$, then $\mathbb{P}\left(\bigcap_{n\geq1} A_n\right) = 1$

And the converse is also true

An intuitive result, and it is nice that our theory reproduces this.

4

Prove Bonferroni's inequality: for any events $A_1, \dots, A_n$,

$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) \geq \sum_{k=1}^{n} \mathbb{P}(A_k) - \sum_{i<j} \mathbb{P}(A_i \cap A_j)$$

Hint: proceed by recurrence.

It may be useful in some situations to know whether infinitely many of the events $A_1, A_2, \dots$ occured. We need to characterize the set of points $\omega \in \Omega$ which are an element of infinitely many of the A_n. This set is denoted

$[A_n \text{ i.o.}] = \{\omega \in \Omega \mid \omega \in A_n \text{ i.o.}\}$

infinitely often

To characterize this set, note that $\omega \in A_n$ infinitely often if $\forall n$, there is a $k = k(n,\omega) \geq n$ such that $\omega \in A_{k(n,\omega)}$; i.e. $\omega \in \bigcup_{k\geq n} A_k \ \forall n$.

$$\Updownarrow$$

$$\omega \in \underbrace{\bigcap_{n\geq1} \bigcup_{k\geq n} A_k}_{\text{this set is usually written } \limsup A_k}$$

Thus
$$\boxed{[A_n \text{ i.o.}] = \underset{\forall}{\underset{\updownarrow}{\bigcap_{n\geq1}}}\ \underset{\exists}{\underset{\updownarrow}{\bigcup_{k\geq n}}} A_k = \lim\sup A_k.}$$

A_n occurs i.o. if $\forall n$, $\exists k \geq n$ such that A_k occurs

And so $[A_n, \text{i.o.}]$ is indeed an event.

We now prove an important result

Borel - Cantelli lemma

If $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n, \text{i.o.}) = 0$

proof:

$$\mathbb{P}(A_n, \text{i.o.}) = \mathbb{P}\Big(\bigcap_{n\geq 1} \underbrace{\boxed{\bigcup_{k\geq n} A_k}}_{B_n}\Big) = \lim_{n\to\infty} \mathbb{P}\Big(\underbrace{\boxed{\bigcup_{k\geq n} A_k}}_{B_n}\Big)$$

property (h) page 13

prop (e) $\leq \lim_{n\to\infty} \sum_{k\geq n} \mathbb{P}(A_k) = 0$ ∎

II_ RANDOM VARIABLES

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ represents all possible outcomes of your RE, and their probabilities. In practice however, we are often interested in some function of the outcome: Random Variables (RVs) describe observations that one can make from the experiment. A RV is a function $X : \Omega \to \mathbb{R}$, defined in such a way that it makes sense to associate a probability to statements like '$X \in [a,b]$' (proba that our observation is in $[a,b]$), thus requiring that $X^{-1}[a,b] := \{\omega \in \Omega \mid X(\omega) \in [a,b]\}$ is an event.

Definition. Consider a RE with probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A real-valued function $X : \Omega \to \mathbb{R}$ such that $\forall B \in \mathcal{B}(\mathbb{R})$

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

is called a RANDOM VARIABLE (RV)

a.k.a. MEASURABLE.

Note that the terminology 'Random Variable' is rather unfortunate, since X is neither random nor a variable.

- Consequence: For RVs X, the probabilities

$$\mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}) = \mathbb{P}(X \in B)$$

are defined for all $B \in \mathcal{B}(\mathbb{R})$.

a convenient shorthand.

- Remark. In the definition of a RV, it is enough to require that $\{X \in (-\infty, x]\} = \{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}$.

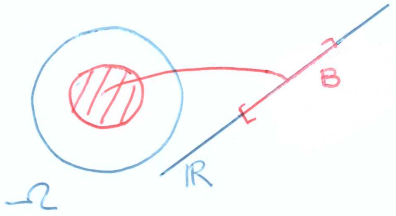
↘ Indeed, let $\mathcal{C}$ = collection of sets $B \in \mathcal{B}(\mathbb{R})$ s.t. $X^{-1}(B) \in \mathcal{F}$ $\subset \mathcal{B}(\mathbb{R})$

Claim: $\mathcal{C}$ is a σ-algebra on $\mathbb{R}$ [def on page 6]

- $\mathbb{R} \in \mathcal{C}$ since $X^{-1}(\mathbb{R}) = \Omega \in \mathcal{F}$
- If $B \in \mathcal{C}$, then $B^c \in \mathcal{C}$.

Indeed, for $B \in \mathcal{C}$, $X^{-1}(B) \in \mathcal{F}$, where $\mathcal{F}$ is a σ-algebra $\Rightarrow [X^{-1}(B)]^c \in \mathcal{F}$.

$$[X^{-1}(B)]^c = \{\omega \in \Omega \mid X(\omega) \in B\}^c = \{\omega \in \Omega \mid X(\omega) \notin B\} = \{\omega \in \Omega \mid X(\omega) \in B^c\} = X^{-1}(B^c)$$



In other words, X^{-1} preserves set operation

$\Rightarrow X^{-1}(B^c) \in \mathcal{F} \Rightarrow B^c \in \mathcal{C}$.

- likewise for stability under countable unions.

↘ All intervals $(-\infty, x] \in \mathcal{C}$ ⇒ smallest σ-algebra will be contained in $\mathcal{C}$ i.e. $\sigma((-\infty, x];\ x \in \mathbb{R}) \subset \mathcal{C}$

But this is exactly $\mathcal{B}(\mathbb{R})$ ⇒ $\mathcal{C} = \mathcal{B}(\mathbb{R})$ ∎

Definition #2 Consider a RE with probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A real-valued function $X : \Omega \to \mathbb{R}$ such that for any $x \in \mathbb{R}$

$$\{X \leq x\} := \{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}$$

is called a RANDOM VARIABLE (RV).

Examples: (i) Constant RVs

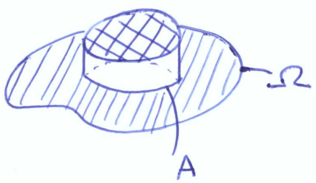
$X \equiv c =$ constant

One has $\{X \leq x\} = \begin{cases} \emptyset & \text{if } x < c \\ \Omega & \text{if } x \geq c \end{cases}$

↖ all in $\mathcal{F}$

(ii) Random indicators

Take $A \in \mathcal{F}$ and define $X = \mathbb{1}_A = \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A \end{cases}$



Then $\{\mathbb{1}_A \leq x\} = \begin{cases} \emptyset & \text{if } x < 0 \\ A^c & \text{if } 0 \leq x < 1 \\ \Omega & \text{if } x \geq 1 \end{cases}$

Thus $\mathbb{1}_A$ is a RV indeed.

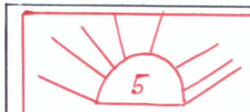
(iii) Simple RVs

Take $\{A_i\}$ = partition of Ω, and put

$X := \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}$, $\alpha_i \in \mathbb{R}$

Then $\{X \leq x\} = \bigcup_{i \mid \alpha_i \leq x} A_i \in \mathcal{F}$ so X is a RV indeed.

↗ In fact, $\{A_i\}$ does not need to be a partition of Ω (why?)



For $x \in \mathbb{R}$, put $x^+ = \max(x, 0)$ its positive part, and $x^- = -\min(0, x)$ its negative part.

Prove that X is a RV, then X^+/X^- are RVs too.

Proposition: Let X be a RV.

Then $\sigma(X) := \{X^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})\}$ is a σ-algebra on Ω, called the "σ-algebra generated by X".

↖ Usually, it is smaller than $\mathcal{F}$: $\sigma(X) \subset \mathcal{F}$.

$\sigma(X)$ represents the information that is obtained by observing X

Examples (i) Indicators $X = \mathbb{1}_A$; $A \in \mathcal{F}$.

Then $\sigma(X) = \{\emptyset, \Omega, A, A^c\}$

[By observing X, you can only answer the following questions: did A occured? did A not occured? X does not contain more information than this]

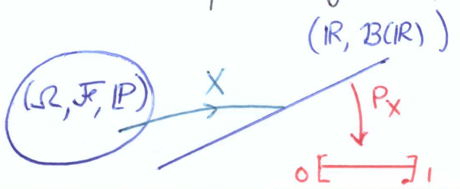
(ii) Simple RVs $X = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}$, $\{A_i\}$ partition

$\sigma(X) = \left\{ \bigcup_{i \in I} A_i , \; I \subset \{1, -, n\} \right\}$.

Definition: The DISTRIBUTION of a RV X is, $\forall B \in \mathcal{B}(\mathbb{R})$,

$$P_X(B) := \mathbb{P}(X \in B) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}).$$

↳ P_X is a probability on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For different $\mathbb{P}$s on $(\Omega, \mathcal{F})$, we will get different distubutions P_X for the same X


$(\mathbb{R}, \mathcal{B}(\mathbb{R}))$
$(\Omega, \mathcal{F}, \mathbb{P})$
X
P_X
0
1

"P_X is induced by $\mathbb{P}$"

We say that X and Y are identically distributed, and we write $X \overset{d}{=} Y$, if $P_X = P_Y$.

Note that we shifted our interest from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ towards the newly defined $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$

"backstage"

Definition. The DISTRIBUTION FUNCTION of a RV X is $F_X(x) := P_X((-\infty, x]) = \mathbb{P}(X \leq x)$

aka the cumulative distribution function (cdf)

We have the following result:

Theorem. (i) F_X is non-decreasing: $x < y \Rightarrow F_X(x) \leq F_X(y)$

(ii) F_X is right-continuous:

$$F_X(x) = F_X(x+) := \lim_{y \downarrow x} F_X(y)$$

(iii) $\lim_{x \to -\infty} F_X(x) = 0$; $\lim_{x \to +\infty} F_X(x) = +1$

proof (i) Follows from the monotonicity of P_X: for $(-\infty, x] \subset (-\infty, y]$, we have that

$$F_X(x) = P_X((-\infty, x]) \leq P_X((-\infty, y]) = F_X(y)$$

(ii) Follows from the continuity of P_X: let $x_n \downarrow x$ and put $A_n := (-\infty, x_n]$. Then $A_n \downarrow A := (-\infty, x]$ as $n \to \infty$, so that $F_X(x_n) = P_X(A_n) \downarrow P_X(A) = F_X(x)$.

(see page 13)

(iii) Follows again from the continuity of P_X since

$B_n := (-\infty, -n] \downarrow \emptyset$

$C_n := (-\infty, n] \uparrow \mathbb{R}$, as $n \to \infty$.

Then

$$\lim_{n \to \infty} F_X(-n) = \lim_{n \to \infty} P_X(B_n) = P_X(\emptyset) = 0$$

$$\lim_{n \to \infty} F_X(n) = \lim_{n \to \infty} P_X(C_n) = P_X(\mathbb{R}) = 1.$$ ■

F_X is not left-continuous. One could be tempted to proceed as in (ii) for right-continuity, arguing that for $x_n < x$, $x_n \uparrow x$, $(-\infty, x_n] \uparrow (-\infty, x]$. However, this is not true, as we have $(-\infty, x_n] \uparrow (-\infty, x)$. Prove it.

We have the following important result:

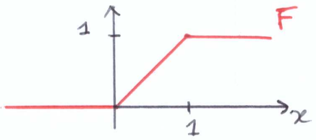
Theorem: For any $F_X : \mathbb{R} \to \mathbb{R}$ which satisfies (i)-(iii), there exists a unique P_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $F_X(x) = P_X((-\infty, x])$, $\forall x \in \mathbb{R}$.

This theorem has huge consequences: to specify the distribution of a RV X, we do not need to compute $P_X(B)$ for all $B \in \mathcal{B}(\mathbb{R})$ [no good]; but instead it is enough to know the expression of F_X; a univariate function with nice properties.

Sketch of proof: For any $x < y$, put $P_X((x,y]) = F_X(y) - F_X(x)$, which is indeed positive from (i).

Then go on and assign a probability to $\bigcup_{i=1}^{n} (x_i, y_i]$, where $-\infty \leq x_1 < y_1 < x_2 < y_2 < \dots < y_n \leq +\infty$. But this is not enough, you then need to extend this measure to the σ-algebra generated by such sets; which existence and uniqueness is ensured by Carathéodory extension theorem. ■

Example: Take $F(x) = \begin{cases} 0 & \text{if } x<0 \\ x & \text{if } 0\le x\le 1 \\ 1 & \text{if } x\ge 1 \end{cases}$



Then there exists a probability on IR with this distribution function. look: it satisfies $P_X((x,y]) = |y-x|$, for $x<y$, $(x,y]\subset[0,1]$.

⇒ This probability is called the uniform distribution on $[0,1]$, and denoted $\mathcal{U}(0,1)$.

[Q: What is the distribution function of $\mathcal{U}(x,y)$, $x<y$?]

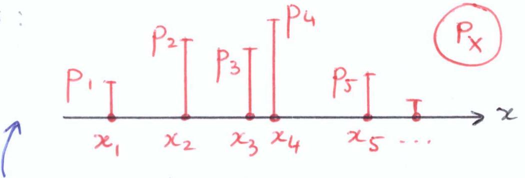
Important classes of distributions on IR.

- Discrete probabilities on IR.

For some countable set $C\subset\mathbb{R}$, $P_X(C)=1$.

Example: $C=\{0,1\}$, $C=\mathbb{N}$ .../...

Then for some $\{x_i\}_{i\geq 1}\subset\mathbb{R}$ and $\{p_i>0\}_{i\geq 1}$ with $\sum p_i=1$, one has:

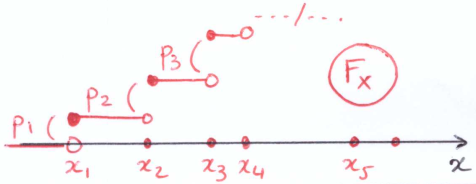


We only need to know the probability of particular outcomes: $p_X(x) = \mathbb{P}(X=x)$ for some $x\in C$.

↑ called the PROBABILITY MASS FUNCTION (pmf).

Equivalently,

$$F_X(x) = \sum_{i=1}^{\infty} p_i \mathbb{1}(x_i\le x)$$



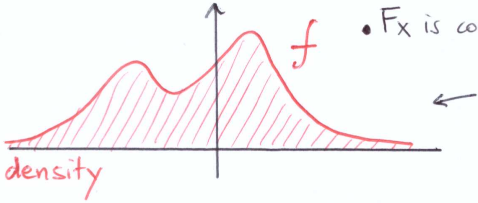
- Absolutely continuous probabilities on IR.

A probability P_X on IR is Absolutely Continuous (AC) if there exists a function $f:\mathbb{R}\to\mathbb{R}_+$ called the DENSITY (aka the probability density function (pdf)) such that

$$\underset{\overset{\shortparallel}{F_X(x)}}{P_X((-\infty,x])} = \int_{-\infty}^{x} f(u)\,du \quad , \quad x\in\mathbb{R}$$

↑ Often the Riemann integral, but sometimes we need to consider more general integrals → Lebesgue integral.

This implies that
- $P_X((a,b]) = \int_a^b f(u)\,du$
- $f(x) = F_X'(x)$ [if f is continuous at x]
- F_X is continuous on IR



← by the way, $\int_{\mathbb{R}} f(u)\,du = 1$

Remark: Any integrable function $f\geq 0$ such that $\int f(u)\,du = 1$ specifies a probability on IR since $F(x) := \int_{-\infty}^{x} f(u)\,du$ satisfies conditions (i)-(iii) on page 19, and theorem p.20 applies.

Notation: If X is AC and f continuous at x,
$P_X((x,x+\Delta)) = f(x)\Delta\,(1+o(1))$.

- Mixed distributions on $\mathbb{R}$.

Mixed distutution are a mixture of discrete and AC distributions. Specifically,

$$P_X = p\, P_d + (1-p)\, P_{ac}$$

$p \in (0,1)$ discrete AC

- Singular distributions on $\mathbb{R}$.

Such distributions have no density (so not AC with respect to the lebesgue measure) nor a probability mass function (since each discrete point $x \in \mathbb{R}$ has zero probability).

Example: Cantor's distribution. (aka Cantor's staircase)

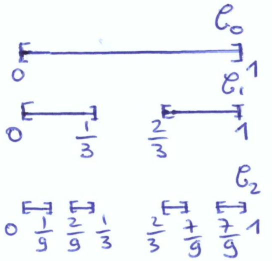
The support of this distribution is Cantor's set $\mathcal{C}$, constructed recursively as follows.

$$\mathcal{C}_0 = [0,1]$$

$$\mathcal{C}_1 = [0,\tfrac{1}{3}] \cup [\tfrac{1}{3}, 1]$$

"remove the middle part"

$$\mathcal{C}_2 = [0,\tfrac{1}{9}] \cup [\tfrac{2}{9}, \tfrac{1}{3}] \cup [\tfrac{2}{3}, \tfrac{7}{9}] \cup [\tfrac{7}{9}, 1]$$



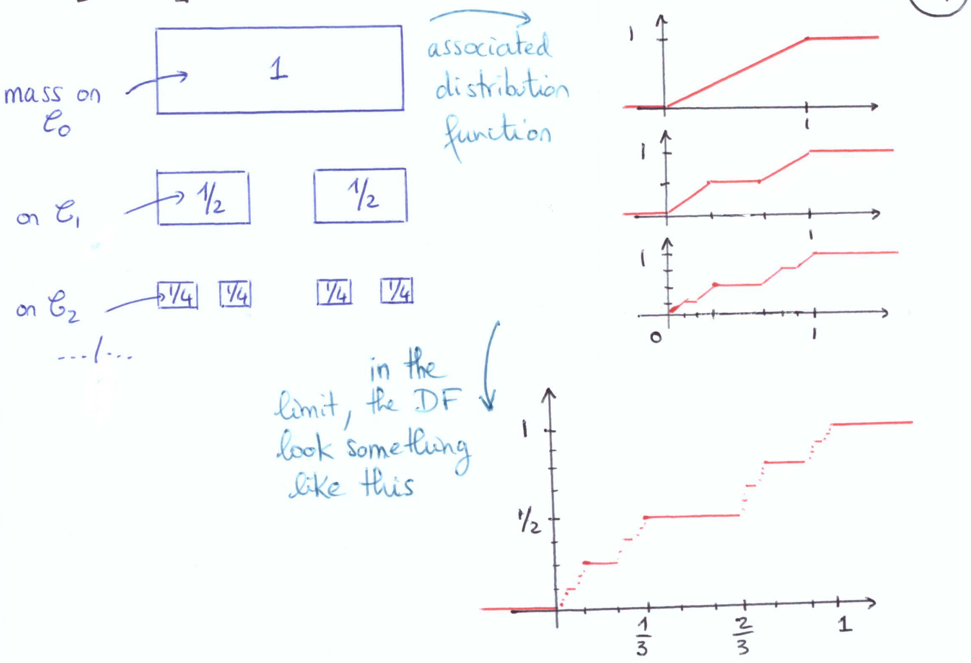
⋮

$$\mathcal{C} = \bigcap_i \mathcal{C}_i$$

Cantor's distribution is defined such that for any $\mathcal{C}_i$ $(i \geq 0)$, the probability of a particular interval $\mathcal{C}_i^j$ $(j = 1, \dots, 2^i)$ is uniform, equal to 2^{-i}.

$$\mathcal{C}_i = \bigcup_{j=1}^{2^i} \mathcal{C}_i^j$$

[Picture]



in the limit, the DF look something like this

It turns out that any distribution on $\mathbb{R}$ is a mixture of discrete, AC, and singular distributions:

Theorem (lebesgue's decomposition)

Any probability on $\mathbb{R}$ has the unique representation of the form

$$P_X = \alpha_d\, P_d + \alpha_{ac}\, P_{ac} + \alpha_s\, P_s,$$

where $\alpha_d, \alpha_{ac}, \alpha_s \geq 0$

$\alpha_d + \alpha_{ac} + \alpha_s = 1$

P_d is discrete, P_{ac} is AC, P_s is singular

- Random Vectors

All what was said about RVs generalizes to random vectors $X = (X_1, \dots, X_d) : \Omega \to \mathbb{R}^d$, where each X_i $(i \leqslant d)$ is a RV.

↘ DF : $F_X(x_1, \dots, x_d) = \mathbb{P}(X_1 \leqslant x_1, \dots, X_d \leqslant x_d)$ where $(x_1, \dots, x_d) \in \mathbb{R}^d$

As in the univariate case, the resulting distribution P_X of X is uniquely specified by F_X.

↘ discrete : same as for discrete RVs

↘ AC : has density $f \geqslant 0$ satisfying

$$F_X(x_1, \dots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(u_1, \dots, u_d)\, du_1 \dots du_d$$

Theorem : (i) $X = (X_1, \dots, X_d)$ is discrete iff all X_i are discrete

(ii) If $X = (X_1, \dots, X_d)$ is AC, then so is X_i for any $i \leqslant d$, and

the "marginal" density of X_i →

$$f_{X_i}(x) = \underbrace{\int \cdots \int}_{d-1} f(u_1, \dots, u_{i-1}, x, u_{i+1}, \dots, u_d)\, du_1 \dots du_{i-1}\, du_{i+1} \dots du_d$$

(some appropriate domain)

7

Suppose that (X, Y) has the joint density function $f(x,y) = \frac{1}{\pi}$ if $x^2 + y^2 \leqslant 1$.

Derive the marginal density functions $f_X(x)$ and $f_Y(y)$ of X and Y.

III _ INDEPENDENCE

The notion of independence goes as follows :

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

(i) A countable set of events $A_1, A_2, \dots \in \mathcal{F}$ are independent if

$$\mathbb{P}(A_{k_1} \cap A_{k_2} \cap \dots \cap A_{k_n}) = \prod_{i=1}^{n} \mathbb{P}(A_{k_i}),$$

for all $n < \infty$ and $k_1, \dots, k_n \in \{1, 2, \dots\}$

(ii) A countable set $\mathcal{F}_1, \mathcal{F}_2, \dots \subset \mathcal{F}$ of σ-algebras are independent if any finite set of events $A_1, \dots, A_n$ from distinct $\mathcal{F}_i$ are independent

(iii) A countable set $X_1, X_2, \dots$ of RVs are independent if the σ-algebras $\sigma(X_i)$ generated by these RVs are independent.

Equivalently, $\forall B_1 \in \mathcal{B}(\mathbb{R})$ ⋮ $\forall B_n \in \mathcal{B}(\mathbb{R})$

$$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i),$$

since $\sigma(X_i) = \{ X_i^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R}) \}$

And this makes perfect intuitive sense with the relative frequency interpretation of probability

How do we know (in practice) if RVs are independent ? We have the following result :

Theorem: (i) RVs $X_1, .., X_n$ are independent iff $\forall x_1, .., x_n \in \mathbb{R}$, $F(x_1, .., x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$

(ii) Discrete RVs $X_1, .., X_n$ are independent iff $\forall x_1, .., x_n \in \mathbb{R}$ $\mathbb{P}(X_1 = x_1, .., X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i)$

(iii) AC RVs $X_1, .., X_n$ are independent iff $\forall x_1, .., x_n \in \mathbb{R}$ $f(x_1, .., x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$

The "$\Rightarrow$" part is relatively straightforward and follows directly from the definition of independent RVs. The reverse direction can be more tedious; except in the discrete case (ii), which is left as an exercise.

Remarks (i) In the definition p.26 and the theorem above, we consider n RVs $X_1, .., X_n$ with DF $F_{X_1}, .., F_{X_n}$ that are independent. How do we even know that such a thing exists at all? That is, how do we ensure the existence of a probability space on which $X_1, ..., X_n$ exist and are independent?

$\Rightarrow$ Take
- $\Omega = \mathbb{R}^n$
- $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$
- $\mathbb{P}$ = proba on $(\Omega, \mathcal{F})$ whose DF is given by $\prod_{i \leq n} F_{X_i}(x_i)$

For $\omega \in \Omega$, put $X_i(\omega) = \omega_i$

By the way, we can always take $(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ and define $X(\omega) = \omega$ "canonical representation"

(ii) In fact, we need a bit more than this: we need the existence of infinitely (countably) many independent RVs $X_1, X_2, \dots$, which brings some additional difficulties. However, it can be done.

Idea: Let $U \sim \mathcal{U}(0,1)$, and denote by $U = \sum_{n \geq 1} U_n 2^{-n}$ its binary expansion.

Then it is possible to show that all the U_n are independent and take values 0 and 1 with equal probability (and the converse is also true).

↳ Then, re-order the U_ns in a two dimensional array:

$$\begin{matrix} U_1 & U_3 & U_6 & U_{10} & \cdots \rightarrow Y_1 := \frac{1}{2} U_1 + \frac{1}{2^2} U_3 + \frac{1}{2^3} U_6 + \cdots \\ U_2 & U_5 & U_9 & \cdots & \rightarrow Y_2 := \frac{1}{2} U_2 + \frac{1}{2^2} U_5 + \frac{1}{2^3} U_9 + \cdots \\ U_4 & U_8 & \cdots & & \rightarrow Y_3 := \frac{1}{2} U_4 + \cdots \\ U_7 & \cdots & & & \\ \vdots & & & & \end{matrix}$$

↑

The $Y_1, Y_2, \dots$ are then independent RVs on the proba space $([0,1], \mathcal{B}[0,1]$, uniform measure$)$, where each Y_i is uniformly distributed on $[0,1]$.

↳ Finally, denoting F_{X_i} the DF of X_i, and $Q_i(x) := \inf\{u \mid F_{X_i}(u) > x\}$, we know that $Q_i(U) \sim F_{X_i}$ for $U \sim \mathcal{U}(0,1)$.

(prove it)

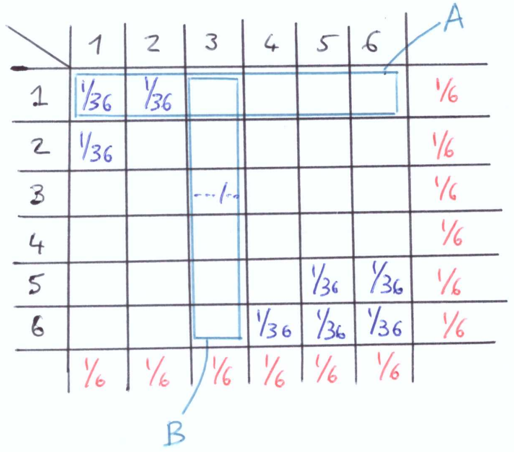
Putting $X_i := Q_i(Y_i)$ gives the desired result ■

Example: We throw two dice

- $\Omega = \{(1,1),(1,2),\dots,(6,6)\}$
- $\mathcal{F}$ = power set = $\mathcal{P}(\Omega)$
- $\mathbb{P}$ such that $\mathbb{P}((i,j)) = \frac{1}{36} \quad \forall i,j$.

Consider A = obtain a 1 on the first dice $\in \mathcal{F}$
B = obtain a 3 on the second dice $\in \mathcal{F}$

Then $A = \{(1,1),(1,2),\boxed{(1,3)},(1,4),(1,5),(1,6)\}$
$B = \{\boxed{(1,3)},(2,3),(3,3),(4,3),(5,3),(6,3)\}$



	1	2	3	4	5	6	
1	1/36	1/36					1/6
2	1/36						1/6
3			--/--				1/6
4							1/6
5					1/36	1/36	1/6
6				1/36	1/36	1/36	1/6
	1/6	1/6	1/6	1/6	1/6	1/6	

We see that

$$\mathbb{P}(A\cap B) = \frac{1}{36} = \mathbb{P}(A)\,\mathbb{P}(B)$$

$\frac{1}{6}$

↑
And this holds for any $A, B \in \mathcal{F}$!

Note that you can also define two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that

$X_1((i,j)) = i$ = number on the first dice
$X_2((i,j)) = j$ = number on the second dice.

X_1 and X_2 are independent discrete RVs.

Remark: $A \cap B \neq \emptyset$, but they are independent.

In fact, two disjoint sets A and B are rarely independent since

$$\mathbb{P}(A\cap B) \underset{A\cap B=\emptyset}{\overset{}{=}} \mathbb{P}(\emptyset) = 0 \underset{\text{in general}}{\neq} \mathbb{P}(A)\,\mathbb{P}(B)$$

"disjoint" → set theory

"independent" → measure theoretic concept