Probability theory

A bit of history

Games of chance

- According to popular belief, knights returning from the Crusades introduced a dice game named ”Hazard.” Although the rules were complicated, the game attracted many players.
- The mathematician Gerolamo Cardano, in the middle of the 16th century, starts to investigate the odds of winning in a game of chance. He gives the first definition of the probability of an event as the ratio of the number of favorable outcomes to the total number of outcomes.
- One century later, Pierre de Fermat and Blaise Pascal lay the first theory of probability when answering a question from Antoine Gombaud, a famous Parisian gambler.

A bit of history

A modern view

- In 1933, the Russian mathematician Andrey Kolmogorov gives the first axiomatic description of probability theory.
- This is the approach we will take in the sequel.

Challenges

- To fully understand Kolmogorov's contribution, we must consider cases in which counting the number of favorable outcomes is insufficient.
- For example, take a situation where the measure of interest is a real number, such as a current or a pressure.
- The value itself is not discrete; only the number of times it falls within a given interval can be observed.
- The Fermat-Pascal approach must be extended to these cases.

From intuition to axioms

Sample space

- Let's consider a very simple experiment: a coin is tossed, and the outcome observed.
- What are the possible results ?
- Of course, "head" and "tails" !
- The **sample space** of an experiment is the set of all possible results.
- In the coin tossing example, it is just the set $\Omega = \{\text{heads}, \text{tails}\}$.
- If a die is rolled, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- In a physical system, Ω may be a subset of the real numbers, or something even more complicated.

From intuition to axioms

Events

- If two coins are tossed, the sample space is $\Omega = \{HH, HT, TH, TT\}$.
- An **event** is a statement about the outcome of the experiment, like ”at least one tails was observed”.
- The previous event is the subset $\{HT, TH, TT\}$ of Ω.
- In probability theory, an **event** is a subset of the **sample space**.
- The empty set $\emptyset$ and the sample space Ω are always considered as events.

From intuition to axioms

A bit of logic

- Assertions about events are also events:
  - If A, B are events, so is $A \cup B$.
  - If A is an event, the complementary set $\bar{A}$ is an event.
- For the two coins tossing experiment, if $A = \{HH\}, B = \{HT\}$, then the event $A \cup B = \{HH, HT\}$ is ”the first toss is a head”, while the event $\bar{A} = \{HT, TH, TT\}$ is ”at least one tails was observed”.

Exercise

- Find the sample space Ω associated with the roll of a die.
- What are the respective subsets A, B of Ω corresponding to the observations ”the result is greater than or equal to 3”,”the result is an even number”?
- Describe the events $A \cup B$, $\bar{A}$.

PROBABILITIES

σ-algebras

- Given a sample space Ω, a σ-algebra of events is a set $\mathcal{T}$ of subsets from Ω such that:
  - $\Omega \in \mathcal{T}$.
  - If $A \in \mathcal{T}$, then $\bar{A} \in \mathcal{T}$.
  - For any **countable** sequence A_n in $\mathcal{T}$, the union $\bigcup_n A_n$ is in $\mathcal{T}$.

Probability measure

- Given a σ-algebra $\mathcal{T}$ on Ω, a probability P is a mapping that assigns to $A \in \mathcal{T}$ a real number $P(A) \in [0, 1]$ such that:
  - $P(\Omega) = 1$.
  - For any countable sequence A_n of **pairwise disjoint** events in $\mathcal{T}$,

$$P\left(\bigcup_n A_n\right) = \sum_n P(A_n). \qquad (1)$$

Some properties

Probability space
A sample space Ω together with a σ-algebra of events $\mathcal{T}$ and a probability P is called a **probability space**, denoted by $(\Omega, \mathcal{T}, P)$.

Set operations
Let $(\Omega, \mathcal{T}, P)$ be a probability space. Let A, B be events. Then:

- $P\left(\bar{A}\right) = 1 - P(A)$.
- $P\left(A \cup B\right) = P(A) + P(B) - P\left(A \cap B\right)$.
- If $B \subset A$, then $P\left(A - B\right) = P(A) - P(B)$, where $A - B = A \cap \bar{B}$.

THE DIE ROLL

Defining a Probability

- $\Omega = \{1,2,3,4,5,6\}$, $\mathcal{T} = \mathcal{P}(\Omega)$, the σ-algebra of all subsets.
- How to find a probability on $\mathcal{T}$?
- It is required that $P(\Omega) = 1$.
- If the die is fair, then the probability to observe any number from Ω must be the same.
- Since:
$$1 = P(\Omega) = \sum_{i=1}^{6} P(\{i\}), \tag{2}$$
one deduces $P(\{i\}) = 1/6, i = 1 \ldots 6$.
- The probability of any event can then be obtained by summing the probabilities of its elements.

Exercise

A card game

- A gambler randomly draws a hand of 5 cards from a deck of 52. Can you find a sample space describing this experiment ?
- If the probabilities of all hands are equal, what is the probability of having a four of a kind ?
- Hint: the number of ways to select k elements in a set of n is :

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}. \tag{3}$$

CONDITIONAL PROBABILITY

Motivation

- In many cases, there is some information about the outcome of an experiment.
- As an example, for a die roll, it may be ”the result is odd”.
- The sample space is thus reduced, and the probabilities must be rescaled accordingly.
- For the die example, it is $\{1, 3, 5\}$, and the probability to draw a 3 is $1/3$.

Bayes’ formula

The conditional probability of A knowing B, denoted $P(A|B)$, is given by Bayes’ formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{4}$$

Around Bayes' formula

Useful formulas

- Let (B_n) be a countable sequence of pairwise disjoint events such that $\bigcup_n B_n = \Omega$. For any event A, one has the formula of total probabilities:

$$P(A) = \sum_n P(A|B_n) P(B_n). \tag{5}$$

- Given events $B_1 \ldots B_n, A$, one has:

$$P(A) = P(A|B_1, \ldots, B_n) P(B_n|B_{n-1}, \ldots B_1) \ldots P(B_1). \tag{6}$$

- Conditioning can be reversed using the next formula, provided $P(B) \neq 0$:

$$P(B|A) = \frac{P(A|B) P(B)}{P(A)} \tag{7}$$

EXERCISE

Two factories, denoted F_0, F_1 are producing components. The end customer receives a component that comes from F_0 with probability 0.3. Furthermore, the probability of the component being defective is 0.1 when coming from factory F_0 and 0.06 when coming from factory F_1. If the recived component is defective, what is the probability that it has been produce by F_1 ?

Random variables

Definition
Given two sample spaces E, F equipped with respective σ-algebras of events $\mathcal{T}, \mathcal{F}$, a mapping $X\colon E \to F$ is said to be a random variable if, for any event $A \in \mathcal{F}$, $X^{-1}(A)$ is an event of $\mathcal{T}$.

Example
Consider a die roll with $E = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{T} = \{E, \emptyset, \{2, 4, 6\}, \{1, 3, 5\}\}$. Let $F = \{0, 1\}, \mathcal{F} = \mathcal{P}(F)$. The mapping X that associates to an even number the value 0 and 1 to an odd number is a random variable. However, Y that maps any value less than 3 to 0 and the others to 1 is not.

The law of a random variable

Définition

- In the previous example, the probability of the event $X = 0$ is $1/2$, the probability of $X^{-1}(0) = \{2, 4, 6\}$.
- If X is a random variable on a probability space $(E, \mathcal{T}, P)$ with values in F, the law of X is the probability on $\mathcal{F}$ defined by:

$$P_X(A) = P\left(X^{-1}(A)\right). \tag{8}$$

- In many cases, the outcome of an experiment is impossible to observe, but an aggregated value may be.
- Formally, this is a random variable.
- It is mainly defined by its law.

Discrete random variables

The law of a discrete random variable

- A random variable is said to be discrete if it takes its values in a finite or infinite countable set.
- The law of such a random variable X with values in a set E is entirely determined by its **distribution** $p(x) = P_X(\{x\}), x \in E$.

Expected value

- Let X be a discrete random variable with values in $E \subset \mathbb{R}$. Its expected value is defined as:

$$E[X] = \sum_{x \in E} x p(x), \tag{9}$$

provided:

$$\sum_{x \in E} |x| p(x) < +\infty. \tag{10}$$

EXERCISE

A game of chance

- Two dice are rolled and their values are added.
- Show that the sum is a random variable X with values in $\{2, \ldots, 12\}$.
- Find the distribution of X.
- If you bet at the start, you'll win five times your original bet if $X \geq 10$. Is this rule fair?

Exercise

Bernoulli random variable

- A bernoulli random variable X can take only two values: 0 and 1.
- If $P(X = 1) = p$, can you compute $P(X = 0)$?
- Compute $E[X]$.

Binomial random variable

- A binomial random variable X with parameters $n \in \mathbb{N}, p \in [0, 1]$ has distribution:

$$p(k) = \binom{n}{k} p^k (1 - p)^{n-k}, k = 0 \ldots n. \quad (11)$$

- Find $E[X]$.

Expected value

Properties

- If X, Y are discrete random variables with values in $E \subset \mathbb{R}$ and λ is a real number, then:

$$E[\lambda X + Y] = \lambda E[X] + E[Y]. \tag{12}$$

- The expected value is monotone. If $X \leq Y$, then:

$$E[X] \leq E[Y] \tag{13}$$

Moments

The expected value in a general sense

- If $g\colon \mathbb{R} \to \mathbb{R}$ is piecewise continuous and

$$\sum_{x \in E} |g(x)| p(x) < +\infty, \tag{14}$$

then one defines:

$$E\left[g(X)\right] = \sum_{x \in E} g(x) p(x) \tag{15}$$

- The variance of the discrete random variable X is:

$$V(X) = E\left[\left(X - E\left[X\right]\right)^2\right] \tag{16}$$

- It measures the dispersion of X around its expected value.

Moments

Higher order moments

- Let n be an integer. The n-th moment of X is, if it exists:

$$E\left[X^n\right]. \tag{17}$$

- The moment generating function is the function:

$$t \mapsto m_X(t) = E\left[e^{tX}\right] \tag{18}$$

- If its domain contains 0, then:

$$E\left[X^n\right] = m^{(n)}(0), \tag{19}$$

where $m^{(n)}$ denotes the n-th derivative of m_X.

Conditioning

Independence

- By the Bayes' formula:

$$P\left(X = x | Y = y\right) = \frac{P\left(X = x, Y = y\right)}{P(Y = y)}. \tag{20}$$

- If $P\left(X = x, Y = y\right) = P\left(X = x\right) P\left(Y = y\right)$, then $P\left(X = x | Y = y\right) = P\left(X = x\right)$.
- In such a case, X, Y are said to be **independent.**
- If X, Y are independent, then $E\left[XY\right] = E[X]E[Y]$.
- If $X_1, \ldots, X_n$ are pairwise independent, then:

$$V\left(X_1 + \cdots + X_n\right) = V\left(X_1\right) + \cdots + V\left(X_n\right). \tag{21}$$

Exercise

A pair of dice... again !

- Let X, Y be two independent random variables corresponding to the respective values of two die throws. Compute :

$$E[X], E[Y], E[X+Y].$$

- Same question with the variances.

A taste of estimation theory

- Let $X_1, \dots X_n$ be a sequence of independent random variables and let:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Compute $E\left[\bar{X}\right], V\left(\bar{X}\right)$. What happens if $n \to +\infty$?

Poisson distribution

A model for random events

- Given a time interval of length $[t_0, t_1]$, one counts the number of occurrences X of an event.
- The random variable X is said to have a Poisson distribution with rate λ if:
$$P(X = k) = \frac{(\lambda T)^k}{k!} e^{-\lambda T}, \tag{22}$$
where $T = t_1 - t_0$.
- $E[X] = V(X) = \lambda T$.
- The Poisson distribution models a situation where the occurrences of an event are independent and occur at constant rate.

EXERCISE

Distribution of a sum

- The moment generating function (MGF) of a random variable is the function $m_X(t) = E\left[e^{tX}\right]$. It is characteristic of a distribution. Prove that if X, Y are independent, $m_{X+Y}(t) = m_X(t)m_Y(t)$.
- Compute the MGF of X with Poisson distribution of rate λ.
- Let X, Y be independent random variables with Poisson distributions of respective rates λ, μ. Show that $X + Y$ has Poisson distribution of rate $\lambda + \mu$.

Selecting events

Let X be a random variable with Poisson distribution of rate λ. The underlying events are selected at random with probability p and counted, yielding a random variable Y. Prove that Y has Poisson distribution with rate $p\lambda$.

Real random variables

Definition

- Let $(\Omega, \mathcal{T}, P)$ be a measure space. A mapping $X \colon \Omega \to \mathbb{R}$ is said to be a real random variable if, for any $t \in \mathbb{R}$, $\{\omega | X(\omega) \leq t\}$ is an event in $\mathcal{T}$.
- If X is a real random variable, its cumulative distribution function (CDF) is the mapping:

$$F_x \colon t \in R \mapsto P\left(X \leq t\right). \tag{23}$$

- $\lim_{t \to -\infty} F_x(t) = 0$, $\lim_{t \to +\infty} F_x(t) = 1$.
- F_x is right continuous and admits a limit to the left.
- If F_x is differentiable, its derivative is called the density of X.

Real random variables

Expected value

- Let X be a real random variable with density p_X.
- If $g\colon \mathbb{R} \to \mathbb{R}$ is piecewise continuous, then one defines:

$$E\left[g(X)\right] = \int_{\mathbb{R}} g(t) p_X(t) dt \tag{24}$$

provided:

$$\int_{\mathbb{R}} |g(t)| p_X(t) dt < +\infty. \tag{25}$$

- As special cases:

$$\begin{aligned} P\left(X \in [a,b]\right) &= E\left[1_{[a,b]}(X)\right] = \int_a^b p_X(t) dt \\ F_X(t) &= \int_{-\infty}^{t} p_X(t) dt. \end{aligned} \tag{26}$$

Real random variables

Joint CDF and density

- If X, Y are real random variables, the CDF of the couple (X, Y) is the mapping:

$$F_{X,Y}\colon (s,t) \in \mathbb{R}^2 \mapsto P\left(X \leq s, Y \leq t\right). \tag{27}$$

- Taking the derivative with respect to s, t allows defining the joint density:

$$p_{X,Y}(x,y) = \frac{\partial^2}{\partial_s \partial_t} F_{X,Y}(t,s)|_{x,y}. \tag{28}$$

- X, Y are said to be independent if:

$$F_{X,Y}(s,t) = F_X(s) F_Y(t). \tag{29}$$

Marginal densities

- Let X, Y be a couple of real random variables with joint density $p_{X,Y}$.
- The density of X (resp. Y) can be obtained as:

$$p_X(x) = \int_{\mathbb{R}} p_{X,Y}(x,t)dt \text{ (resp.) } p_Y(y) = \int_{\mathbb{R}} p_{X,Y}(s,y)ds \tag{30}$$

- The conditional density of X knowing Y is defined as:

$$p\left(X|Y=y\right)(s) = \frac{p_{X,Y}(s,y)}{p_Y(y)}. \tag{31}$$

- If X, Y are independent, $p(X|Y=y)(s) = p_X(s)$ or, equivalently, $p_{X,Y}(x,y) = p_X(x)p_Y(y)$.

Characteristic function

Definition

- Let X be a real random variable. Its characteristic function is:

$$\phi_X \colon t \in \mathbb{R} \mapsto E\left[e^{itX}\right]. \tag{32}$$

- Using the density p_X, it can be computed as:

$$\phi_X(t) \int_{\mathbb{R}} e^{itx} p_X(x) dx. \tag{33}$$

Properties

- If two random variables have the same characteristic function, they have the same distribution.
- If X, Y are independent, $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.
- $E\left[X^k\right] = i^{-k}\phi_X^{(k)}(0), k \in \mathbb{N}$.

Uniform distribution

- The uniform distribution on an interval $[a, b]$ has density:

$$p(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b], \\ 0 & \text{otherwise.} \end{cases} \tag{34}$$

- If X has uniform distribution on $[a, b]$, then:

$$\begin{aligned} E[X] &= \frac{a+b}{2}, \; V(X) = \frac{(b-a)^2}{12}, \\ \phi_X(t) &= \frac{e^{itb} - e^{ita}}{t(b-a)}. \end{aligned} \tag{35}$$

Normal distribution

Density

- Let $\mu \in \mathbb{R}, \sigma > 0$. A real random variable X is said to have a normal distribution $\mathcal{N}(\mu, \sigma^2)$ if its density is:

$$p_X(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}. \tag{36}$$

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then:

$$Y = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1). \tag{37}$$

NORMAL DISTRIBUTION

Moments
If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$:

$$\begin{aligned} &E\left[X\right] = \mu,\ V\left(X\right) = \sigma^2, \\ &\phi_X(t) = \exp\left(i\mu t - \sigma^2 t^2/2\right). \end{aligned} \tag{38}$$

Cumulative distribution function

- There is no closed-form expression for the CDF of a $\mathcal{N}(\mu, \sigma^2)$ distribution, but it can be evaluated numerically.
- If Φ is the CDF of a $\mathcal{N}(0, 1)$ distribution, then the CDF of a $\mathcal{N}(\mu, \sigma^2)$ distribution is:

$$\Phi\left(\frac{x-\mu}{\sigma}\right). \tag{39}$$

Central limit theorem

Convergence in distribution
A sequence $(X_n)_{n\mathbb{N}}$ of real random variables is said to converge in distribution to a random variable X if:

$$\forall t \in \mathbb{R}, \lim_{n \to +\infty} P(X_n \leq t) = P(X \leq t). \tag{40}$$

Theorem (Central limit)
If $(X_n)_{n\in\mathbb{N}}$ is a sequence of independent, identically distributed real random variables, then the sequence of random variables $\sqrt{n}\left(\bar{X}_n - \mu\right)$ converges in distribution to $X \sim \mathcal{N}(0, \sigma^2)$, with:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \mu = E\left[X_i\right], \sigma^2 = V\left(X_i\right). \tag{41}$$

Exponential distribution

- The exponential distribution with rate $\lambda > 0$ has density:

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0, \\ 0, x < 0. \end{cases} \tag{42}$$

- If X is exponentially distributed with rate λ :

$$E\left[X\right] = \frac{1}{\lambda}, V(X) = \frac{1}{\lambda^2}, \phi_X(t) = \frac{\lambda}{\lambda - it}. \tag{43}$$

- The exponential distribution is memoryless:

$$P\left(X > t + s | X > s\right) = P\left(X > t\right), \ s, t \geq 0. \tag{44}$$

VECTOR RANDOM VARIABLES

CDF and density

- Let $(\Omega, \mathcal{T}, P)$ be a measure space. Let $X\colon \Omega \to \mathbb{R}^n$ be a mapping. X is said to be a vector (or multivariate) random variable iff:

$$\{\omega | X_1(\omega) \leq t_1, \ldots, X_n(\omega) \leq t_n\} \in \mathcal{T},\ (t_1, \ldots t_n) \in \mathbb{R}^n. \tag{45}$$

- The CDF of X is:

$$F_x(t_1, \ldots, t_n) = P\left(X_1 \leq t_1, \ldots, X_n \leq t_n\right). \tag{46}$$

- Taking the partial derivatives yields the density:

$$p_X(t_1, \ldots, t_n) = \frac{\partial^n F_X(t_1, \ldots, t_n)}{\partial t_1 \ldots \partial t_n}. \tag{47}$$

Vector random variables

Moments

- The expected value of a vector random variable X is just the vector of the expected values of the coordinates:

$$E[X] = (E[X_1], \ldots, E[X_n]) . \quad (48)$$

- The density of $X_i, i = 1 \ldots n$ can be computed by integration:

$$p_{X_i}(x) = \int_{\mathbb{R}^{n-1}} p_X(t_1, \ldots, x, \ldots t_n)\, dt_1, \ldots dt_n, \quad (49)$$

where x occurs in the i-th position.

Vector random variables

Moments

- Higher moments are tensors, generally difficult to compute.
- The covariance matrix of two random variables is:

$$Cov(X,Y) = E\left[XY^t\right] - E\left[X\right]E\left[Y\right]^t. \tag{50}$$

- The density of a vector normal distribution $\mathcal{N}\left(\mu, \Sigma\right)$ is:

$$P_X(x) = \frac{1}{(2\pi)^{n-2}\det(\Sigma)^{-1/2}}\exp\left(\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right), \tag{51}$$

with $\mu = E\left[X\right], \Sigma = Cov(X,X)$.

Vector random variables

Characteristic function

- The characteristic function of a vector random variable X is defined as:

$$\phi_X(t_1, \ldots, t_n) = E\left[e^{i(t_1 X_1 + \cdots + t_n X_n)}\right]. \tag{52}$$

- For a $\mathcal{N}(\mu, \Sigma)$ distribution, it is:

$$\phi_X(t) = \exp\left(i\mu^t t - \frac{1}{2} t^t \Sigma t\right). \tag{53}$$

VECTOR RANDOM VARIABLES

Central limit theorem

- If $X_1, \dots X_n$ are independent, identically distributed vector random variables, then $\sqrt{n}\left(\bar{X}_n - \mu\right)$ converges in distribution to $\mathcal{N}\left(0, \Sigma\right)$ with:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \Sigma = Cov(X_i, X_i). \tag{54}$$

- In practice, almost all problems in data analysis can be solved using the vector CLT.