Probability Theory

LECTURE 1

Course Information

- Lecturer: Oana Lang, e-mail: oana.lang@ubbcluj.ro
- Lectures: Thursday 2pm - 4pm between 27 February - 5 June 2025, Room 6/II, 1 M. Kogalniceanu, Level 2.
  Seminars/Tutorials:

Ziua	Orele	Frecventa	Sala	Anul	Formatia	Tipul	Cadrul didactic
Luni	16-18		A313	1 Inteligenta Artificiala in limba engleza	1011	Seminar	Conf. LISEI Hannelore
Joi	12-14		A312	1 Inteligenta Artificiala in limba engleza	1012	Seminar	Lect. LANG Oana
Joi	14-16		6/II	1 Inteligenta Artificiala in limba engleza	IA1	Curs	Lect. LANG Oana

Ziua	Orele	Frecventa	Sala	Anul	Formatia	Tipul	Cadrul didactic
Luni	8-10		A321	2 Matematica informatica - linia de studiu engleza	821	Seminar	Asist. MICU Tudor
Miercuri	8-10		A321	2 Matematica informatica - linia de studiu engleza	822	Seminar	Asist. MICU Tudor

- Examination:
  - 70% written examination in June
  - 30% Coursework - Deadline: 30 May 2025
  - extra 10% possible based on seminar activity - minimal requirement: 5 'stars' per term where 1 $\star$ is given for a demonstration at the board.
- Pre-requisites: basic concepts in set theory

Outline for today

- Why should we learn Probability Theory?
  - What is *uncertainty*?
- What is a *probability*? (classical definition)
- What is a *probability space?* (axiomatic definition)
  - Outcomes, events, sample space
- Rules of Probability
- Real-world applications
- Time for questions and discussion

Why should we learn Probability Theory?

Suppose you are planning to wait for a friend at the train station. The train is coming from Bucharest and it is **supposed to** arrive on time. All of a sudden, the website is updated and it is announced that the train left Sighișoara one hour behind the schedule.

*Given the new information, are you **sure** that the train will arrive on time in Cluj-Napoca?*

Why should we learn Probability Theory?
What is *Uncertainty*?

- Uncertainty surrounds us every day:
  - Before entering the lecture room, you couldn't be certain about the number of colleagues you are going to find in the room.
  - When you toss a coin, roll a die, or buy a lottery ticket.
  - *I wonder if I will risk and lose the bet?*
  - *Is that really* ***possible****?*
  - *According to meteorological forecasts it* ***might*** *rain tomorrow.*
  - Viruses can attack a computer system at *unpredictable/uncertain* times and can affect an *unpredictable/uncertain* number of files/directories.

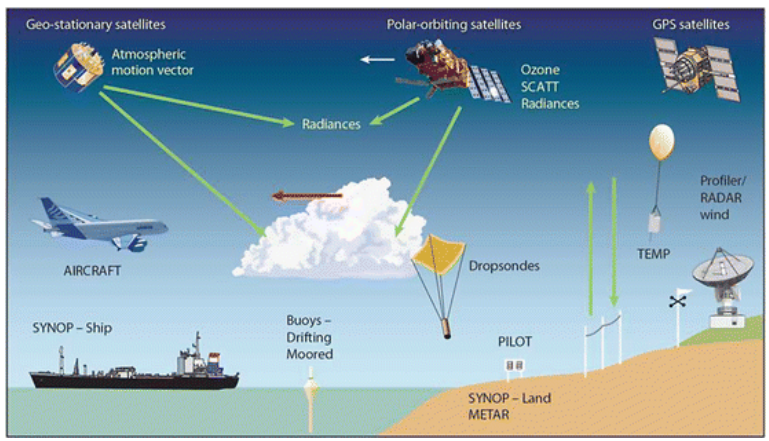
$\Rightarrow$ we are forced to make decision *under uncertainty.*

- *Uncertainty* refers to a condition when a situation/event cannot be predicted *for sure*, with no error. ([2])

We will use Probability Theory to measure and quantify uncertainty. This will allow us to make decisions under uncertainty.

Further Examples

- Medical diagnosis: the chance of having a particular medical condition given a positive test result is calculated using probabilities (tests for diseases).
- Weather forecasting: meteorologists use historical data and current conditions to estimate the probability of rain, storms, or other weather events occurring in a given location. They can predict, say 70% chance of rain, but they don't know the future situation *for sure*.



What is Probability Theory?

Probability Theory is a branch of mathematics that deals with the study of *random* phenomena.

random = happening by chance with no cause or reason; unpredictable, unintended
- *aleatorius* (lat.) = random
- *alea* (lat.) = dice (for a game); dice game

↪ One measures the chances of success or the risk of failure of events.

Random experiments, trials, events

The terms *trial* or *experiment* are used in probability theory to describe virtually any process or action whose *outcome* is not known in advance with certainty $\rightarrow$ it has a **random** behaviour.

The random event is the result of an experiment.

- roll two game dice
  $\rightarrow$ both dice show 1
- draw a playing card
  $\rightarrow$ a 3 was drawn
- lottery draw (6 out of 49)
  $\rightarrow$ the first number drawn is 23

Outcomes, events, and probabilities
Sample space

Intuitively: the probability of an event represents the *chance* that the event will happen.

- When you toss a (fair) coin: 50-50 chance of turning up heads or tails $\Rightarrow$ the probability of each side is equal to $\frac{1}{2}$.
- When making predictions (e.g. in forecasting): it is common to speak about probability as a *likelihood* e.g a company's profit is *likely* to rise this year.
- Two software companies are competing for an important contract. We know that company A is *twice as likely* to win the contract as company B $\Rightarrow$ the probability to win the contract for company A is equal to $\frac{2}{3}$, while this probability is just $\frac{1}{3}$ for company B.

Outcomes, events, and probabilities
Sample space

- Probabilities appear when we consider and *weight* possible results of some *experiments*. Some of these results are more *likely* than others.

- A collection of all elementary results, or **outcomes** of an experiment is called **sample space**.
- A set of outcomes is an **event**. $\Rightarrow$ events are subsets of the sample space.

Outcomes, events, and their probabilities
Sample space

- Example 1: A tossed die can produce 6 possible outcomes: $1, 2, 3, \ldots,$ or 6 dots. Each outcome is an event, in this case. We can observe also other events: an even/odd number of dots, a number of dots less than 4, etc.
- Example 2: Consider a football game between U Cluj and CFR Cluj. The sample space consists of 3 outcomes:

$$\Omega = \{\text{U Cluj wins, CFR Cluj wins, they tie}\}.$$

If we combine these outcomes in all possible ways, we obtain $2^3 = 8$ events: U Cluj wins, loses, ties, gets at least a tie, gets at most a tie, no tie, gets *some* result, gets *no result*.

- The event "some result" is the entire sample space $\Omega \Rightarrow$ it should have probability 1.
- The event "no result" is empty (it does not contain any outcome) $\Rightarrow$ it has probability 0.

What is the *probability* of an event? (classical definition)

Definition 1

We consider an experiment which has finitely many equally probable outcomes. The probability that the event A occurs is

$$\mathbb{P}(A) = \frac{\text{the number of favorable outcomes for the occurrence of } A}{\text{number of all possible outcomes within the experiment}}.$$

Definition 2

Let A be a random event appearing in an experiment; the experiment is repeated n times (under the same given conditions) and denote by k_n how many times the event A appears; the *relative frequency of the event* A is the number

$$h_n(A) = \frac{k_n}{n}$$

and k_n is the *absolute frequency of the event* A.

- After repeating an experiment n times (n sufficiently large), under the same conditions, the relative frequency $h_n(A)$ of the event A is approximately equal to the probability $\mathbb{P}(A)$

$$h_n(A) \approx \mathbb{P}(A), \text{ if } n \to \infty.$$

$\Rightarrow$ In the long run, the probability of an event can be viewed as a *proportion* of times this event happens i.e. its relative frequency.

Dice Rolling - History

The correspondence between B. Pascal and P. Fermat, in which they investigated the dice rolling problem of the French nobleman and gambler Chevalier de Méré is famous:
It is said that de Méré had been betting that in four rolls of a die at least one six would turn up. He was consistently winning and to get more people to play, he changed the game bet: in 24 rolls of two dice, a pair of sixes would turn up. But with this second bet, de Méré lost and felt that 25 rolls were necessary to make the game favorable.
We will calculate and compare the probabilities of the following events:
A : we obtain at least one six in 4 rolls of a die;
B : we obtain at least one pair of sixes in 24 rolls of two dice;
C : we obtain at least one pair of sixes in 25 rolls of two dice.

Dice Rolling

For this problem it is easier to determine the probabilities of the contrary events $\bar{A}, \bar{B}$ and $\bar{C}$. The event $\bar{A}$ means no six is obtained in 4 rolls of a die.

$$\Longrightarrow \mathbb{P}(\bar{A}) = \frac{5^4}{6^4} \Longrightarrow \mathbb{P}(A) = 1 - \frac{5^4}{6^4} \approx 0.5177.$$

The event $\bar{B}$ means no pair of sixes is obtained in 24 rolls of two dice

$$\mathbb{P}(\bar{B}) = \frac{35^{24}}{36^{24}} \Longrightarrow \mathbb{P}(B) = 1 - \frac{35^{24}}{36^{24}} \approx 0.4914.$$

The event $\bar{C}$ means no pair of sixes is obtained in 25 rolls of two dice

$$\mathbb{P}(\bar{C}) = \frac{35^{25}}{36^{25}} \Longrightarrow \mathbb{P}(C) = 1 - \frac{35^{25}}{36^{25}} \approx 0.5055.$$

Comparing now the calculated probabilities we notice

$$\mathbb{P}(B) < \frac{1}{2} < \mathbb{P}(C) < \mathbb{P}(A).$$

Set operations in Probability

- Events are *sets* of outcomes $\Rightarrow$ in order to learn how to compute probabilities of events, we use set operations.

- The **complement** of an event A is an event that occurs every time when A does not occur. It consists of outcomes excluded from A. *Notation:* A^c or $\bar{A}$.
- A **union of events** $A, B, C, \ldots$ is an event which consists of all the outcomes in all these events. It occurs if any of $A, B, C, \ldots$ occurs.
- An **intersection** of events $A, B, C, \ldots$ is an event which consists of outcomes which are common in all these events. It occurs if each $A, B, C, \ldots$ occurs.
- A **difference** of events A and B consists of all outcomes included in A but excluded from B. It occurs when A occurs and B does not occur.

Set operations in Probability

- The events A and B are disjoint if their intersection is empty, that is

$$A \cap B = \emptyset.$$

- The events $A_1, A_2, A_3, \ldots$ are **mutually exclusive** or **pairwise disjoint** if any two of these events are disjoint, that is

$$A_i \cap A_j = \emptyset \quad \text{for any} \quad i \neq j.$$

- The events $A_1, A_2, A_3, \ldots$ are **exhaustive** if their union equals the whole sample space, that is

$$A_1 \cup A_2 \cup \ldots = \Omega.$$

What is a Probability Space? (axiomatic definition)

In probability theory, we use set notation to define key concepts:

- Sample Space (Ω): The sample space is the set of all possible outcomes of a random experiment.
- Events (in $\mathcal{F}$): Events are subsets of the sample space Ω. Formally, let $\mathcal{F}$ be a sigma-algebra over Ω. Any subset $A \in \mathcal{F}$ is considered an *event*.
- Probability Measure ($\mathbb{P}$): The probability measure $\mathbb{P}$ assigns a real number between 0 and 1 to each event in $\mathcal{F}$.
- Outcomes (ω): An outcome ω is a specific element in the sample space. For example, if the experiment is rolling a six-sided die, a possible outcome might be a specific number like 3 or 6.

What is a probability space?

Consider a non empty set Ω. A σ-field or σ-algebra $\mathcal{F}$ on Ω is a collection of subsets of Ω with the following properties:

1. The empty set $\emptyset$ belongs to $\mathcal{F}$.
2. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ where $A^c = \Omega \backslash A$.
3. Given a countable collection of sets $A_i, i = 1, 2, \ldots$ from $\mathcal{F}$, their union

$$\bigcup_{i=1}^{\infty} A_i$$

   also belongs to $\mathcal{F}$.

- Note that if $A_i, i = 1, 2, \ldots$ is a countable collection of sets from $\mathcal{F}$, their intersection $\bigcap_{i=1}^{\infty} A_i = \left(\bigcup_{i=1}^{\infty} A_i^c\right)^c$ also belongs to $\mathcal{F}$.

What is a probability space?

Let $\mathcal{F}$ be a σ-field on a non empty set Ω. A probability measure $\mathbb{P}$ on Ω is a function

$$\mathbb{P} : \mathcal{F} \to [0, 1]$$

such that:

1. $\mathbb{P}(\Omega) = 1$.
2. For any (infinite) sequence of disjoint sets $\{A_i\}_{i \geq 1}$, that is $A_i \cap A_j = \emptyset$, if $i \neq j$, then

$$\mathbb{P}\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} \mathbb{P}\left(A_i\right).$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*.

Rules of Probability

- Complement rule:

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

- Probability of a union

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

  - For mutually exclusive events:

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

Rules of Probability

Theorem 3

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $A, B \in \mathcal{F}$. Then the following properties are true:

(1) $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$ *and* $0 \leq \mathbb{P}(A) \leq 1$.

(2) $\mathbb{P}(\emptyset) = 0$.

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

(4) If $A \subseteq B$, *then* $\mathbb{P}(A) \leq \mathbb{P}(B)$, *i.e.* $\mathbb{P}$ *is monotone.*

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

Rules of Probability

Proof:

(1) Since A and $\bar{A}$ are disjoint and $A \cup \bar{A} = \Omega$, we have $\mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(\bar{A})$. Taking into account that $\mathbb{P}(\Omega) = 1$, it follows that $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$. But $\mathbb{P}(\bar{A}) \geq 0$ implies $\mathbb{P}(A) \leq 1$.

(2) We apply (1) for $A = \Omega$ and use $\mathbb{P}(\Omega) = 1$.

(3) In virtue of the equality $A = (A \cap B) \cup (A \backslash B)$ we have $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \backslash B)$.

(4) We have $A \cap B = A$, since $A \subseteq B$. Then by (1) and (3) it follows that $0 \leq \mathbb{P}(B \backslash A) = \mathbb{P}(B) - \mathbb{P}(A)$.

(5) The following property holds $A \cup (\bar{A} \cap B) = A \cup B$, where the union on the left side of the equality is composed of disjoint sets. Thus,

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(\bar{A} \cap B) = \mathbb{P}(A) + \mathbb{P}(B \backslash A) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B), \end{aligned}$$

where we also used property (3).

Rules of Probability

Theorem 4

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $(A_n)_{n\geq 1}$ is a sequence of events from $\mathcal{F}$. The inclusion-exclusion principle holds

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{\substack{i,j=1 \\ i<j}}^{n} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n-1}\mathbb{P}(A_1 \cap \cdots \cap A_n)$$

for all $n \in \mathbb{N}^*$.

Rules of Probability

Proof:

We use the induction method. For $n = 2$ the property was proved in Theorem 3 - (5). Assuming the property is true for $n \in \mathbb{N}^*$, we prove that it is also true for $n + 1$. By Theorem 3 - (5) we write

$$\mathbb{P}\left(\bigcup_{i=1}^{n+1} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) + \mathbb{P}\left(A_{n+1}\right) - \mathbb{P}\left(\bigcup_{i=1}^{n} \left(A_i \cap A_{n+1}\right)\right)$$

and applying the property (inclusion-exclusion principle) for n sets we have

$$\mathbb{P}\left(\bigcup_{i=1}^{n} \left(A_i \cap A_{n+1}\right)\right)$$

$$= \sum_{i=1}^{n} \mathbb{P}\left(A_i \cap A_{n+1}\right) - \sum_{\substack{i,j=1 \\ i<j}}^{n} \mathbb{P}\left(A_i \cap A_j \cap A_{n+1}\right) + \cdots + (-1)^{n-1} \mathbb{P}\left(A_1 \cap \cdots \cap A_n \cap A_{n+1}\right).$$

Using these relations in (1) it follows that the inclusion-exclusion principle holds also for $n + 1$ sets.

Rules of Probability

Definition 5

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

$(A_n)_{n\geq 1}$ is an *increasing sequence* of events from $\mathcal{F}$, if $A_n \subseteq A_{n+1}$ for each $n \in \mathbb{N}^*$.

$(A_n)_{n\geq 1}$ is a *decreasing sequence* of events from $\mathcal{F}$, if $A_{n+1} \subseteq A_n$ for each $n \in \mathbb{N}^*$.

Rules of Probability

Theorem 6

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. Then the following properties are true:

(1) If $(A_n)_{n\geq 1}$ is an increasing sequence of events from $\mathcal{F}$, then

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right).$$

(2) If $(A_n)_{n\geq 1}$ is a decreasing sequence of events from $\mathcal{F}$, then

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right).$$

Rules of Probability

Proof:

(1) We define in $\mathcal{F}$ the sequence $(B_n)_{n\geq 1}$ of events by

$$B_1 = A_1, \quad B_n = A_n \backslash A_{n-1} \text{ for } n \geq 2.$$

Since $(A_n)_{n\geq 1}$ is an increasing sequence of events, we have

$$\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n \text{ and } B_i \cap B_j = \emptyset \text{ for } i \neq j \in \mathbb{N}^*.$$

We can write

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) = \mathbb{P}(A_1) + \sum_{n=1}^{\infty} \mathbb{P}(A_{n+1} \backslash A_n)$$
$$= \lim_{n\to\infty} \left(\mathbb{P}(A_1) + \mathbb{P}(A_2 \backslash A_1) + \ldots \mathbb{P}(A_n \backslash A_{n-1})\right).$$

By Theorem 3 we have

$$\mathbb{P}(A_{n+1} \backslash A_n) = \mathbb{P}(A_{n+1}) - \mathbb{P}(A_n) \qquad \text{for all } n \in \mathbb{N}^*,$$

because $A_n \subseteq A_{n+1}$. Thus,

$$\mathbb{P}(A_1) + \mathbb{P}(A_2 \backslash A_1) + \ldots \mathbb{P}(A_n \backslash A_{n-1}) = \mathbb{P}(A_n).$$

Rules of Probability

Proof:

Then by (3)

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}\left(A_n\right).$$

(2) We apply (1) for $B_n = \bar{A}_n$. So $(B_n)_{n\geq 1}$ becomes an increasing sequence of events from $\mathcal{F}$ and it holds by the previous result that

$$\lim_{n\to\infty} \mathbb{P}\left(B_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right).$$

Due to De Morgan's laws this equality implies

$$\lim_{n\to\infty} \left(1 - \mathbb{P}\left(A_n\right)\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} \bar{A}_n\right) = \mathbb{P}\left(\overline{\bigcap_{n=1}^{\infty} A_n}\right) = 1 - \mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right).$$

Therefore, $\lim_{n\to\infty} \mathbb{P}\left(A_n\right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right)$.

Rules of Probability

Remark:

The operations of union and intersection are commutative:

$$A \cup B = B \cup A, \quad A \cap B = B \cap A;$$

associative:

$$(A \cup B) \cup C = A \cup (B \cup C), \quad (A \cap B) \cap C = A \cup (B \cap C);$$

and distributive:

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C), \quad (A \cap B) \cup C = (A \cup C) \cap (B \cup C),$$

It holds: $A \cup \bar{A} = S, A \cap \bar{A} = \emptyset$ and the laws of De Morgan:

$$\overline{A \cup B} = \bar{A} \cap \bar{B}, \quad \overline{A \cap B} = \bar{A} \cup \bar{B}.$$

Real-world Applications

- Networks: *Conditional probabilities* are used in network security for detecting anomalies or potential security threats. It helps assess the probability of a network event to be a security threat based on historical data.
- Machine Learning and AI: Probabilities are used in Bayesian machine learning, where models are updated based on new data. Bayesian networks use probability rules to model relationships between variables and make predictions based on observed data.

Real-world Applications

- Finance and Risk Management: Probability models are extensively used in finance for risk assessment, portfolio optimization, and pricing of financial derivatives. For example, the Black-Scholes model for option pricing involves probability distributions.
- Epidemiology: Probability models are essential in epidemiology to understand the spread of diseases, predict outbreaks, and assess the effectiveness of intervention strategies.

Real-world Applications

In all these applications, the use of probability theory provides a structured and quantitative framework to model uncertainty, make predictions, and guide decision-making processes. This mathematical foundation allows practitioners to assess risks, optimize processes, and make informed choices in the face of uncertainty.

LECTURE 2

Outline

- Basics of Probability - Quick Review
- Conditional Probability
- Independence

Why do we need conditional probabilities?
Quick example

Suppose you are planning to wait for a friend at the train station. The train is coming from Bucharest and it is supposed to arrive on time, with probability 70%. All of a sudden, the website is updated and it is announced that the train left Sighișoara one hour behind the schedule $\Rightarrow$ the probability for the train to arrive on time **becomes** 5%.

$\Rightarrow$ new information influences the probability of meeting your friend on time.

The updated probability is called **conditional probability**, where the new information (the train left Sighișoara late) is a condition.

Conditional Probability

Definition 7

On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, consider two events $A, B \in \mathcal{F}$ such that $\mathbb{P}(B) \neq 0$. The conditional probability of A given B is defined by

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{1}$$

- This represents *the probability that event A occurs, when event B is known to have occurred*.
- It can be rewritten to obtain a general formula for the probability of intersection of two events:

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

Conditional Probability

- Remark that, given the new information = occurence of the event B, we replace the (standard) formula for the *unconditional probability* of A

$$\mathbb{P}(A) = \frac{\text{number of outcomes in } \quad A}{\text{number of outcomes in } \quad \Omega}$$

by the *conditional probability* of A given B

$$\mathbb{P}(A \mid B) = \frac{\text{number of outcomes in } \quad A \cap B}{\text{number of outcomes in } \quad B}.$$

Conditional Probability

Example 1

Consider a deck of 52 cards. Let

A: the event of drawing a red card
B: the event of drawing a queen.

We want to find the conditional probability for

$\mathbb{P}(A|B)$: the probability of drawing a red card given that a queen is drawn. By definition

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

In our particular case:

$$\mathbb{P}(A \cap B) = \mathbb{P}(\text{drawing a red queen})$$

$$\mathbb{P}(B) = \mathbb{P}(\text{drawing a queen})$$

Note: if event A occurred i.e. the first card was red, then the probability of drawing a second red card becomes slightly smaller. We say that the two events are **dependent**.

Conditional Probability

Example 1

We have

$$\mathbb{P}(A \cap B) = \mathbb{P}(\text{drawing a red queen}) = \frac{2}{52}$$

$$\mathbb{P}(B) = \mathbb{P}(\text{drawing a queen}) = \frac{4}{52}$$

Substituting these values into the formula:

$$\mathbb{P}(A|B) = \frac{\frac{2}{52}}{\frac{4}{52}}$$

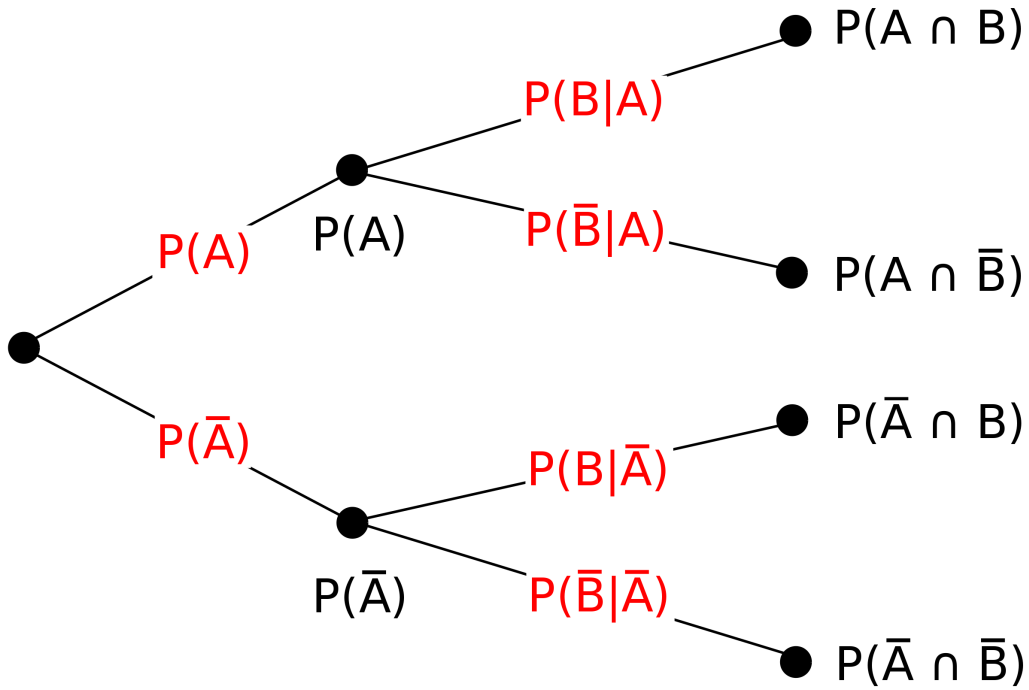
Simplifying:

$$\mathbb{P}(A|B) = \frac{2}{4} = \frac{1}{2}$$

Therefore $\mathbb{P}(A|B) = \frac{1}{2}$.

Conditional Probability

Visual interpretation



Tree diagram - conditional probabilities

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

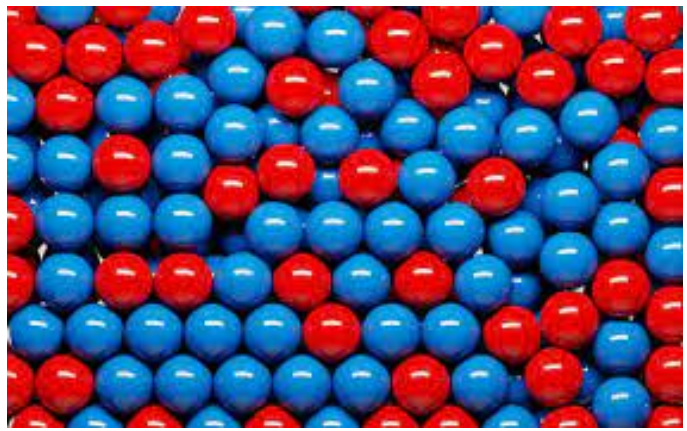
Conditional Probability

Example 2

An urn contains 4 blue marbles and 5 red marbles. Two marbles are successively drawn without replacement.

a) Knowing that the first marble is red, what is the probability that the second marble is blue?

b) What is the probability that both balls are red?



Conditional Probability

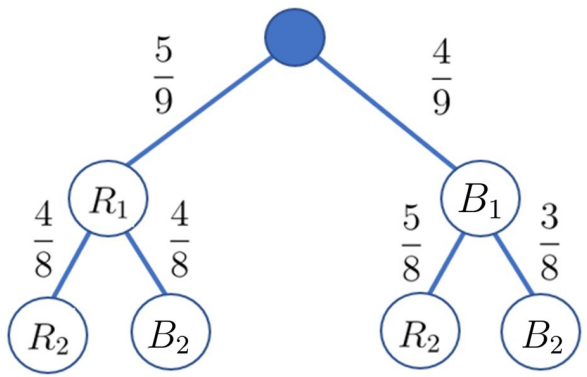
Example 2

For $i \in \{1, 2\}$ consider the events:

R_i : the i^{th} drawn marble is red
$B_i = \bar{R}_i$: the i^{th} drawn marble is blue

a) conditional probability: $\mathbb{P}(B_2 \mid R_1) = \frac{4}{8}$.

b)
$\mathbb{P}(R_1 \cap R_2) = \mathbb{P}(R_2 \mid R_1)\,\mathbb{P}(R_1) = \frac{4}{8} \cdot \frac{5}{9}$.



Independent events

Definition 8

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and two events $A, B \in \mathcal{F}$. The events A and B are called independent if

$$\boxed{\mathbb{P}(A \mid B) = \mathbb{P}(A)} \tag{2}$$

that is *the occurrence of B does not affect the probability of A.*

$\Rightarrow$ Substituting this into

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A \mid B) \tag{3}$$

gives

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

$\Rightarrow$ *conditional probability* equals *unconditional probability* when the events are **independent**.

Independent events

Alternative definition

Definition 9

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The events $A, B \in \mathcal{F}$ are said to be independent events if

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)} \tag{4}$$

Remark:

- *Independent events* A and $B \Rightarrow$ the occurrence of the event A not affecting the occurrence of the event B and vice versa.
- *Dependent events* A and $B \Rightarrow$ the occurrence of the event A affecting the occurrence of the event B or vice versa.

Independent events

Examples:
(1) When you toss two coins, the outcome of one does not affect the other, therefore the events are independent.

(2) When you roll a die and toss a coin, the outcome of one does not affect the outcome of the other, therefore the events are independent.

(3) You draw a first card from a deck with 52 cards; denote the event B_1: the first drawn card is black; you draw a second card; denote the event B_2 : the second drawn card is black. By knowing that the first card was black, you made the probability of drawing a second black card slightly smaller. The two events B_1 and B_2 are dependent.

Independent events

Proposition 1

In a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ let $A, B \in \mathcal{F}$. Then the following assertions are equivalent:

(1) A and B are independent.

(2) $\bar{A}$ and B are independent.

(3) A and $\bar{B}$ are independent.

(4) $\bar{A}$ and $\bar{B}$ are independent.

Independent events

Proof.

In order to prove the equivalences we use the properties from Theorem 3 and definition 9.

(1)⇔(2):

$$\begin{aligned}\mathbb{P}(A\cap B)=\mathbb{P}(A)\mathbb{P}(B) &\Leftrightarrow \mathbb{P}(B)-\mathbb{P}(A\cap B)=\mathbb{P}(B)-\mathbb{P}(A)\mathbb{P}(B)\\ &\Leftrightarrow \mathbb{P}(B)-\mathbb{P}(A\cap B)=\mathbb{P}(B)\mathbb{P}(\bar{A})\\ &\Leftrightarrow \mathbb{P}(B\setminus A)=\mathbb{P}(B)\mathbb{P}(\bar{A}) \Leftrightarrow \mathbb{P}(\bar{A}\cap B)=\mathbb{P}(\bar{A})\mathbb{P}(B).\end{aligned}$$

(1)⇔(3):

$$\begin{aligned}\mathbb{P}(A\cap B)=\mathbb{P}(A)\mathbb{P}(B) &\Leftrightarrow \mathbb{P}(A)-\mathbb{P}(A\cap B)=\mathbb{P}(A)-\mathbb{P}(A)\mathbb{P}(B)\\ &\Leftrightarrow \mathbb{P}(A)-\mathbb{P}(A\cap B)=\mathbb{P}(A)\mathbb{P}(\bar{B})\\ &\Leftrightarrow \mathbb{P}(A\setminus B)=\mathbb{P}(A)\mathbb{P}(\bar{B}) \Leftrightarrow \mathbb{P}(A\cap \bar{B})=\mathbb{P}(A)\mathbb{P}(\bar{B}).\end{aligned}$$

(1)⇔(4):

$$\begin{aligned}&\mathbb{P}(A\cap B)=\mathbb{P}(A)\mathbb{P}(B) \Leftrightarrow \mathbb{P}(A\cup B)=\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A)\mathbb{P}(B)\\ &\Leftrightarrow 1-\mathbb{P}(A\cup B)=(1-\mathbb{P}(A))(1-\mathbb{P}(B)) \Longleftrightarrow 1-\mathbb{P}(A\cup B)=\mathbb{P}(\bar{A})\mathbb{P}(\bar{B})\\ &\Leftrightarrow \mathbb{P}(\bar{A}\cap \bar{B})=\mathbb{P}(\bar{A})\mathbb{P}(\bar{B}).\end{aligned}$$

Independent events

Definition 2

- *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then $A_1, ..., A_n \in \mathcal{F}$ are* ***independent events*** *if*

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_m}) = \mathbb{P}(A_{i_1}) \cdot \ldots \cdot \mathbb{P}(A_{i_m})$$

for each subset $\{i_1, \ldots, i_m\} \subset \{1, ..., n\}$ with $2 \leq m \leq n$.

- $A_1, ..., A_n \in \mathcal{F}$ *are* ***pairwise independent events*** *if*

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j) \quad \textit{for all} \quad i, j \in \{1, \ldots, n\}, i \neq j.$$

Remark:
The independence of two (or more) events means that the occurrence of one has no effect on the occurrence or not of the other(s). It can be extended to sets, (collections of) σ-algebras, or even *random variables*.

Independent events

Exercises:
1) What does it mean that three events A, B, C are independent?
2) How many conditions must be verified for n events $B_1, \ldots, B_n$ to be independent?
3) You roll a die twice. Consider the events:
A: the first number is 6; B: the second number is 5; C: the first number is 1;

- Are the following events independent or dependent?

a) A and B;
b) A and C;
c) B and C.

- Are the following events disjoint?

a) A and B;
b) A and C;
c) B and C.

Independent events

Answers:
1) $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C), \quad \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C)$ and $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.
2) $C_n^2 + C_n^3 + ... + C_n^n = 2^n - C_n^1 - C_n^0 = 2^n - n - 1$ conditions.
3) • a) A and B are independent;
b) A and C are dependent;
c) B and C are independent.
• a) A and B are not disjoint (they both can appear at the same time within the experiment);
b) A and C are disjoint (they cannot appear at the same time within the experiment);
c) B and C are not disjoint.

LECTURE 3

Outline

- Independent events - continuation
- Total Probability Rule
- Bayes' Rule
- Applications

Independent events

Example
Let the four faces of a regular tetrahedron be painted as follows: one red, one blue, one green. The fourth face is painted with all three colours. The tetrahedron is thrown, and the following events are considered:
R: the tetrahedron falls on a surface which contains red colour.
B: the tetrahedron falls on a surface which contains blue colour
G: the tetrahedron falls on a surface which contains green colour.
Are the 3 events R, B, G independent?

Independent events

Answer: R, B, G are not independent, because $\mathbb{P}(R \cap B \cap G) = \frac{1}{4} \neq \mathbb{P}(R)\mathbb{P}(B)\mathbb{P}(G) = \frac{1}{2^3}$. But R, B, G are pairwise independent, since it holds

$$\mathbb{P}(R \cap B) = \mathbb{P}(R)\mathbb{P}(B); \mathbb{P}(B \cap G) = \mathbb{P}(B)\mathbb{P}(G); \mathbb{P}(R \cap G) = \mathbb{P}(R)\mathbb{P}(G).$$

Independent events

Example 3

We have the following scenario:

- 90% of planes depart on time
- 80% of planes arrive on time
- 75% of planes depart on time **and** arrive on time

(a) Alex is meeting Sabrina. Sabrina's plane departed on time. What is the probability that Sabrina will arrive on time?

(b) Alex has met Sabrina, and she arrived on time. What is the probability that her plane departed on time?

(c) Are the events *departing on time* and *arriving on time* independent?

Independent events

Example 3

Let

$$A = \{ \text{ arriving on time } \}$$
$$D = \{ \text{ departing on time } \}.$$

We have:

$$\mathbb{P}(A) = 0.8, \quad \mathbb{P}(D) = 0.9, \quad \mathbb{P}(A \cap D) = 0.75.$$

Then

(a) $\mathbb{P}(A \mid D) = \frac{\mathbb{P}(A \cap D)}{\mathbb{P}(D)} = \frac{0.75}{0.9} = 0.8333.$

(b) $\mathbb{P}(D \mid A) = \frac{\mathbb{P}(A \cap D)}{\mathbb{P}(A)} = \frac{0.75}{0.8} = 0.9375.$

(c) Events are not independent, since

$$\mathbb{P}(A \mid D) \neq \mathbb{P}(A), \quad \mathbb{P}(D \mid A) \neq \mathbb{P}(D), \quad \mathbb{P}(A \cap D) \neq \mathbb{P}(A)\mathbb{P}(D).$$

Note that $\mathbb{P}(A \mid D) > \mathbb{P}(A)$ and $\mathbb{P}(D \mid A) > \mathbb{P}(D) \Rightarrow$ departing on time increases the probability of arriving on time, and vice versa.

In general: the two conditional probabilities $\mathbb{P}(A \mid B)$ and $\mathbb{P}(B \mid A)$ are not the same.

Total Probability Rule

- Consider a partition of the sample space Ω with mutually exclusive and exhaustive events $B_1, B_2, \ldots, B_k$. That is

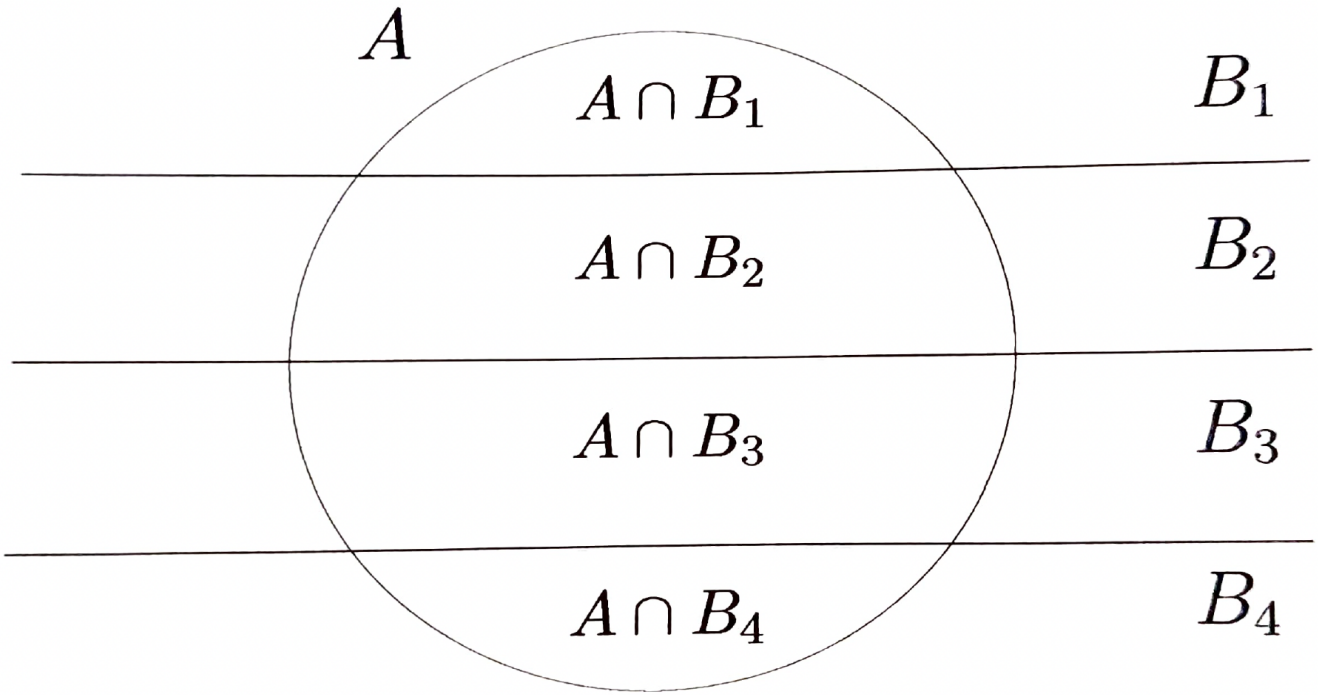
$$B_i \cap B_j = \emptyset \text{ for any } i \neq j \text{ and } B_1 \cup \ldots \cup B_k = \Omega.$$

- Assume that these events also partition the event A, that is

$$A = (A \cap B_1) \cup \ldots \cup (A \cap B_k)$$

and this is also a union of mutually exclusive events.

Total Probability Rule



Partition of the sample space Ω and the event A[1]

[1] M. Baron, *Probability and Statistics for Computer Scientists*

Total Probability Rule

Then

$$\mathbb{P}(A) = \sum_{j=1}^{k} \mathbb{P}\left(A \cap B_j\right)$$

$\Rightarrow$ The Total Probability Rule or The Law of Total Probability:

$$\boxed{\mathbb{P}(A) = \sum_{j=1}^{k} \mathbb{P}\left(A \mid B_j\right) \mathbb{P}(B_j)} \tag{5}$$

- Relates the unconditional probability of an event A with its conditional probabilities.
- It is used when it is easier to compute conditional probabilities of A given additional information.

Total Probability Rule

- Overall, probability $\mathbb{P}(A)$ is just a "weighted average" of the conditional probabilities $\mathbb{P}(A \mid B_j)$, where the "weights" are $\mathbb{P}(B_j)$.
- These "weights" add up to 1, since $B_1, \ldots, B_k$ are a partition of all the possible outcomes, whose total probability is 1.
- This means that the overall probability $\mathbb{P}(A)$ will always lie somewhere between the conditional probabilities $\mathbb{P}(A \mid B_j)$, with more weight given to the more probable scenarios.

Bayes' Rule

Consider the following scenario:

- There exists a test for a certain infection (say Covid19) which is:
  - 99% reliable for healthy patients
  - 95% reliable for infected patients
- Let us define the following events:
  V: the patient has the virus
  S: the test is positive

$\Rightarrow$ if the patient has the virus, the test will show this with probability $\mathbb{P}(S \mid V) = 0.95$.
$\Rightarrow$ if the patient does not have the virus, the test will show this with probability $\mathbb{P}(S^c \mid V^c) = 0.99$.

- Consider a patient whose test result is positive. Knowing that sometimes the test is wrong, the patient wants to now if she/he has indeed the virus, that is we need to find out $\mathbb{P}(V \mid S)$.
- Since $S \cap V = V \cap S$ we have $\mathbb{P}(V)\mathbb{P}(S \mid V) = \mathbb{P}(S)\mathbb{P}(V \mid S)$ and then

$$\mathbb{P}(V \mid S) = \frac{\mathbb{P}(S \mid V)\mathbb{P}(V)}{\mathbb{P}(S)}.$$

Bayes' Rule

- Describes the probability of an event B, based on prior knowledge of conditions that may be related to the event:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)} \tag{6}$$

- Bayes' Rule is to probability "what Pythagora's theorem is to geometry".
- It allows to update the probability of an event using new information $\rightarrow$ a method for adjusting or refining current predictions given new or additional evidence.
- Widely used today in machine learning.

Thomas Bayes
(1701-1761)

Bayes' Rule

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

- Heuristics:

$$\text{Posterior} = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Evidence}}$$

- Probability $\Rightarrow$ measures a "degree of belief". Bayes' theorem links the degree of belief in a proposition/system before and after accounting for *evidence* (information, direct observation).
- $\mathbb{P}(B) =$ *the prior* is the initial degree of belief in B before any updates (before any new information/observation).
- $\mathbb{P}(B \mid A) =$ *the posterior* is the degree of belief after incorporating information regarding the fact that A is true.
- $\frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$ incorporates the *evidence* A provides for B.

Bayes' and Total Probability Rule

- For $k = 2$ the total probability rule becomes

$$\mathbb{P}(A) = \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c)$$

which together with Bayes' rule gives

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c)}.$$

- In general

$$\mathbb{P}(A) = \sum_{j=1}^{k} \mathbb{P}\left(A \mid B_j\right) \mathbb{P}(B_j)$$

and then

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{k} \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}.$$

Applications

In Filtering

$\Rightarrow$ all these are probabilistic tools which can be used to estimate the state of a dynamical system

- Conditional probability is closely linked to *conditional expectation*. These can be defined for *random variables* = some "functions" defined on our sample space.
- Random variables are sometimes solutions corresponding to sets of equations = a *mathematical model*, which can describe the current/past/future state of a dynamical system $\Rightarrow$ making inferences about such a solution (about its conditional expectation/probability) $\leftrightarrow$ making inferences about the state of the dynamical system itself.
- Explicit calculations can be performed only for simple cases. For more complicated (= realistic) models we can only *approximate* the conditional probability using specific *approximation methods*.

Applications

In Filtering

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

$$\text{Posterior} = \mathbb{P}(B \mid A) = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Evidence}}$$

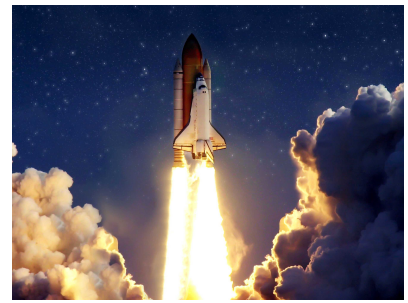
$$\pi_{t-1}^{z_{0:t-1}} \xrightarrow[\substack{\text{model}\\ \text{forecast}\\ \text{prediction}}]{\mathcal{M}_t} \mathcal{M}_t\left(\pi_{t-1}^{z_{0:t-1}}\right) =: p_t \xrightarrow[\substack{\text{assimilation}\\ \text{analysis}\\ \text{update}}]{g_t^{z_t}\star} g_t^{z_t} \star p_t = \pi_t^{z_{0:t}}.$$

π_t is the *posterior distribution* which gives information about the state of the dynamical system at time t, taken into account the information provided by observations made up to time t (modelled via $z_{0:t}$).

LECTURE 4

Random Variables and Random Vectors

Are these useful?



Assume a spaceship is launched.

- The costs involved are estimated in millions of dollars/pounds/euros $\Rightarrow$ trying several times would be extremely expensive $\Rightarrow$ spaceship's performance is first *simulated* $\rightarrow$ this allows experts to *evaluate* the associated risks (reliability, safety, etc) in advance.

- *Computer simulations* are used to estimate quantities/costs/outputs for which the direct computation is too risky/difficult/expensive or even impossible.
- *Monte Carlo Methods* are used for computing probabilistic characteristics associated with such simulations: instead of generating the *real* process (spaceship launch, weather forecast etc) we only simulate some associated *random variables*. This suffices in order to infer the probabilistic behaviour of the system.

Random Variables and Random Vectors

- The implementation of Monte Carlo methods reduces to generating random variables from given *distributions*.

  $\Rightarrow$ random variables and vectors provide a rigorous framework for calculating probabilities, defining distributions, and formulating statistical hypotheses.

- In statistical applications, these concepts become the foundation for parameter estimation, hypothesis testing, and building predictive models.

- Random variables and random vectors: *discrete* or *continuous*.

Random Variables

Definition 3

*A **random variable** is a measurable function*

$$X : \Omega \to \mathbb{R}. \tag{8}$$

- *Measurable:*

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$
$$\forall A \in \mathcal{B}(\mathbb{R}), \quad X^{-1}(A) \in \mathcal{F}$$

  where

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

- Its range can be the set of all real numbers, $(0, \infty)$, the integers $\mathbb{Z}$, an interval (a, b) etc.
- Once the experiment is completed and the outcome $\omega \in \Omega$ is known, the value $X(\omega)$ of the random variable becomes *determined*.
- A random variable is a quantity that depends on *chance*.

- The **Borel σ-algebra** $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra containing all open sets in $\mathbb{R}$. Explicitly, the Borel σ-algebra satisfies:

$$\mathcal{B}(\mathbb{R}) = \bigcap \{\mathcal{A} \mid \mathcal{A} \text{ is a } \sigma\text{-algebra containing all open sets in } \mathbb{R}\}$$

meaning it is the intersection of all σ-algebras that contain the open sets.

- The **Borel σ-algebra** $\mathcal{B}(\mathbb{R})$ is generated by the collection of all open intervals of $\mathbb{R}$. Specifically, the collection of sets:

$$(-\infty, x] = \{y \in \mathbb{R} : y \leq x\}, \quad x \in \mathbb{R}$$

is sufficient to generate $\mathcal{B}(\mathbb{R})$ through countable set operations (unions, intersections, complements).

- Any other set (e.g., an open or closed set) in $\mathbb{R}$ can be constructed using countable unions or intersections of sets of the form $(-\infty, x]$.

Even though open intervals (a, b) are part of $\mathcal{B}(\mathbb{R})$, the collection of sets

$$(-\infty, x] = \{y \in \mathbb{R} \mid y \leq x\}, \quad x \in \mathbb{R}$$

is also sufficient to generate $\mathcal{B}(\mathbb{R})$. Open sets can be approximated using $(-\infty, x]$. Specifically,

$$(a, b) = \bigcup_{n=1}^{\infty} \left((-\infty, b - \tfrac{1}{n}] \cap (-\infty, a + \tfrac{1}{n}]^c \right).$$

Since a σ-algebra is closed under countable unions, this shows that open sets can be built from $(-\infty, x]$. A closed interval $[a, b]$ can be expressed as an intersection of countably many sets of the form $(-\infty, x]$:

$$[a, b] = \bigcap_{n=1}^{\infty} \left((-\infty, b + \tfrac{1}{n}] \cap (-\infty, a - \tfrac{1}{n}]^c \right).$$

Since σ-algebras are closed under countable intersections, this means that closed intervals also belong to the Borel σ-algebra.

Random Variables

- Remarks:
  1. We can also say that $X : \Omega \to \mathbb{R}$ *is a random variable if*

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F} \quad \text{for each} \quad x \in \mathbb{R}. \tag{9}$$

  2. Notation:

$$X^{-1}(A) = \{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\} \quad \text{for each} \quad A \in \mathcal{B}(\mathbb{R}).$$

  3. *Indicator function* of the event $A \in \mathcal{F}$

$$\mathbb{1}_A(\omega) \quad = \quad \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \in \bar{A}. \end{cases}$$

  This is a discrete random variable (check the definition above).

Random Variables

Examples

E1 Throw two dice, the sum of the obtained numbers is a random variable
$X : \Omega \to \{2, 3, ..., 12\}$, where Ω contains all elementary events, i.e. the sample space is

$$\Omega = \left\{(\omega_i^1, \omega_j^2) : i, j \in \{1, 2, ..., 6\}\right\},$$

where (ω_i^1, ω_j^2) is the elementary event. The first die indicates the number i and the second die indicates the number j, for
$i, j \in \{1, 2, ..., 6\} \Longrightarrow X(\omega_i^1, \omega_j^2) = i + j$ for each $i, j \in \{1, 2, ..., 6\}$.
For example $\mathbb{P}(X = 5) = \frac{4}{36}$ (since
$5 = 1 + 4 = 4 + 1 = 2 + 3 = 3 + 2$); $\mathbb{P}(X = 6) = \frac{5}{36}$, etc.

E2 A player throws two coins $\Rightarrow \Omega = \{(H, T), (H, H), (T, H), (T, T)\}$

The random variable X indicates how many times tail T has appeared: $\Rightarrow X : \Omega \to \{0, 1, 2\}$

$\Rightarrow \mathbb{P}(X = 0) = \mathbb{P}(X = 2) = \frac{1}{4}, \mathbb{P}(X = 1) = \frac{1}{2}$.

Random Variables

Examples

E3 ([2]) Consider an experiment in which we toss 3 fair coins and we count the number of heads. The same *model* suits the numbers of girls/boys in a family with 3 children, the number of 1's or 0's in a random binary string of 3 characters, and so on.

- Let X be the numbers of *heads* (girls, 1's, etc). Prior to the experiment, we don't know the value of X, all we know about it is that X is an integer between 0 and 3. We assume that each value is an event, and therefore we can compute the corresponding probabilities:

$$\mathbb{P}(X = 0) = \mathbb{P}(\text{3 tails}) = \mathbb{P}(TTT) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

$$\mathbb{P}(X = 1) = \mathbb{P}(HTT) + \mathbb{P}(THT) + \mathbb{P}(TTH) = \frac{3}{8}$$

Random Variables

Examples

$$\mathbb{P}(X = 2) = \mathbb{P}(HHT) + \mathbb{P}(HTH) + \mathbb{P}(THH) = \frac{3}{8}$$

$$\mathbb{P}(X = 3) = \mathbb{P}(HHH) = \frac{1}{8}$$

Overall

x	$\boldsymbol{P}\{X = x\}$
0	1/8
1	3/8
2	3/8
3	1/8
Total	1

The table contains everything we can know about the *random variable* X **before** the experiment. Before finding out the outcome $\omega \in \Omega$ we cannot say what is the actual value of X - all we can do is to list all *possible* values of X together with their corresponding probabilities.

Discrete and Continuous Random Variables

- ***Discrete***: X is said to be a discrete random variable if its range $X(\Omega)$ is a countable set i.e. $X(\Omega) = \{x_i : i \in I\}$ where $I \subseteq \mathbb{N}$ (set of indices) and $\mathbb{P}(X = x_i) > 0$ for each $i \in I$.
  - it has finitely many values $(x_1, \ldots, x_n)$ or countably infinitely many values $(x_1, \ldots, x_n, \ldots)$; the values can be listed
  - *numerical* random variables: the number of cars in a parking lot, the number of sixes in 100 dice rolls, the number of defective parts during a production, the number of items sold at a store on a certain day, the number of customers that enter a certain shop on a given day, the number of voters who showed up to the polls.
  - *categorical* random variables: weather forecast e.g. rainy, cloudy, foggy, clear $\rightarrow$ classification into categories.

Discrete and Continuous Random Variables

- ***Continuous***: the set of its possible values is uncountable. Its possible values comprise either a single interval from $\mathbb{R}$, or a union of disjoint intervals of $\mathbb{R}$.
  - The values cannot be listed.
  - E.g. the running time of a machine until first defection, software installation time, code execution time, the temperature in a certain city within a year, amount of rainfall in a certain city over a year, the speed of a car passing a speed camera, the time it takes to complete an exam for a 60 minute test.

Random Variables

Theoretical examples

- Let X and Y be random variables. Then $aX + b$ (where $a, b \in \mathbb{R}$), $|X|$, $\min\{X, Y\}$, $\max\{X, Y\}$, $X + Y$, $X - Y$ and $X \cdot Y$ are random variables. Moreover, if $Y(\omega) \neq 0$ for all $\omega \in \Omega$, then also $\dfrac{X}{Y}$ is a random variable.

- Let $(X_n)_{n\geq 1}$ be a sequence of random variables such that $\sup\limits_{n\geq 1} X_n$, $\inf\limits_{n\geq 1} X_n \in \mathbb{R}$ for each $\omega \in \Omega$. Then $\sup\limits_{n\geq 1} X_n$, $\inf\limits_{n\geq 1} X_n$, $\limsup\limits_{n\to\infty} X_n := \inf\limits_{n\geq 1}(\sup\limits_{k\geq n} X_k)$, $\liminf\limits_{n\to\infty} X_n := \sup\limits_{n\geq 1}(\inf\limits_{k\geq n} X_k)$are random variables.

Random Variables and Their Distributions

Definition 4

The collection of all probabilities related to X is the distribution of X.

Definition 5

The function $P : \mathbb{R} \to [0, 1]$ with

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

is the probability mass function (pmf).

Definition 6

The probability function

$$P_X : \mathcal{B}(\mathbb{R}) \to [0, 1]$$
$$P_X(B) = \mathbb{P}(X \in B) \text{ for each } B \in \mathcal{B}(\mathbb{R})$$

is called the probability distribution function or the law of X.

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

Proposition 7

Let X be a random variable. Then the mapping $P_X : \mathcal{B}(\mathbb{R}) \to [0,1]$ defined by

$$P_X(B) = \mathbb{P}(X \in B) \text{ for each } B \in \mathcal{B}(\mathbb{R}) \tag{10}$$

is a probability over $\mathcal{B}(\mathbb{R})$ i.e. $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ is a probability space.

Proof.

(i) $P_X(\mathbb{R}) = \mathbb{P}(\omega \in \Omega : X(\omega) \in \mathbb{R}) = \mathbb{P}(\Omega) = 1$.

(ii) $P_X(B) \geq 0$ for all $B \in \mathcal{B}(\mathbb{R})$.

(iii) If $(B_n)_{n\geq 1}$ are pairwise disjoint events, then

$$
\begin{aligned}
P_X\left(\bigcup_{n=1}^{\infty} B_n\right) &= \mathbb{P}\left(\left\{\omega \in \Omega : X(\omega) \in \bigcup_{n=1}^{\infty} B_n\right\}\right) \\
&= \mathbb{P}\left(X^{-1}\left(\bigcup_{n=1}^{\infty} B_n\right)\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} X^{-1}\left(B_n\right)\right) \\
&= \sum_{n=1}^{\infty} \mathbb{P}\left(X^{-1}\left(B_n\right)\right) = \sum_{n=1}^{\infty} P_X\left(B_n\right).
\end{aligned}
$$

Note that $\left(X^{-1}\left(B_n\right)\right)_{n\geq 1}$ are pairwise disjoint events, since $(B_n)_{n\geq 1}$ are pairwise disjoint events and X is measurable. □

Random Variables - Probability mass function (pmf)

If X is a discrete random variable with range $X(\Omega) = \{x_i : i \in I\}$, then P_X is completely determined by the values

$$\mathbb{P}(X = x_i) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x_i\}) \quad i \in I$$

i.e. by the probability mass function of the discrete random variable X. We say that X has a *discrete distribution*. We write

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_i & \dots \\ p_1 & p_2 & \dots & p_i & \dots \end{pmatrix} = \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}$$

where

$$I \subseteq \mathbb{N}, p_i := \mathbb{P}(X = x_i) > 0, i \in I, \sum_{i \in I} p_i = 1.$$

Random Variables - Cumulative Distribution Function (cdf)

Definition 8

The cumulative distribution function (cdf) of a random variable X is defined as $F : \mathbb{R} \to \mathbb{R}$

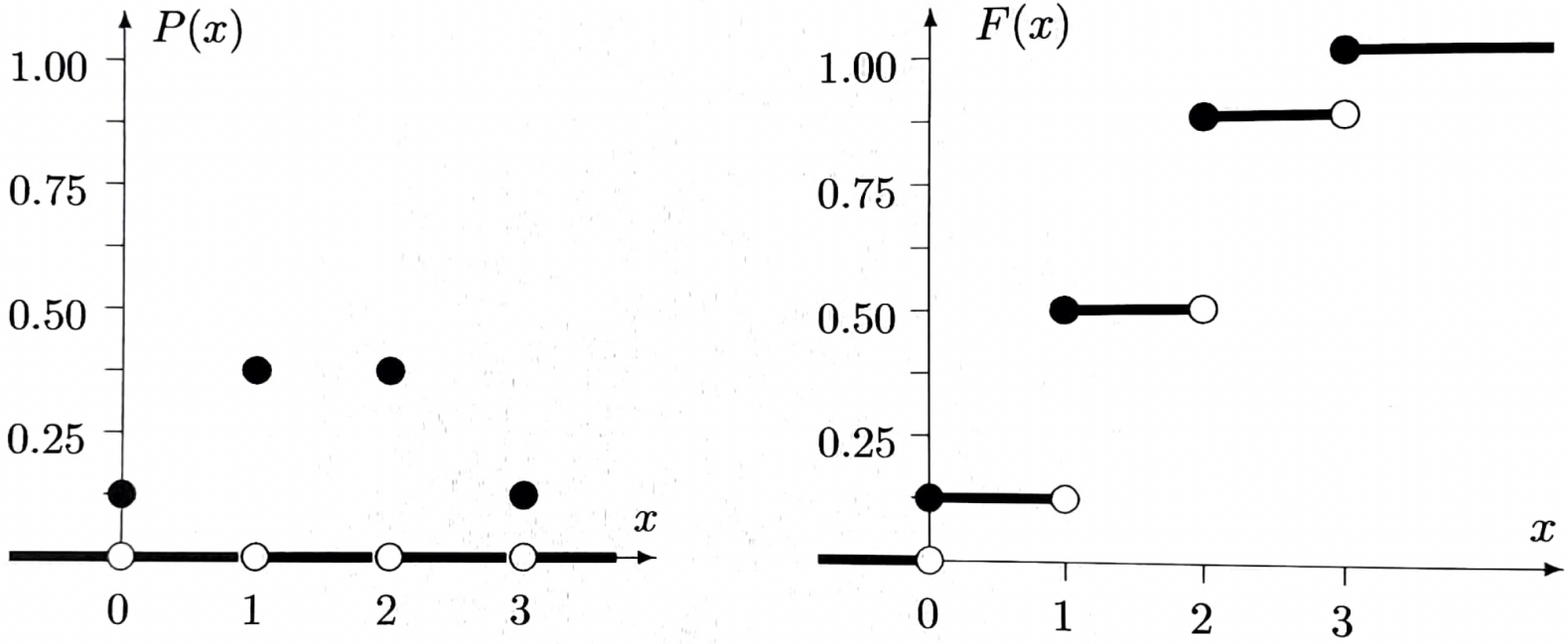
$$F(x) = \mathbb{P}(X \leq x).$$

- The set of possible values of X is called the support of the distribution F.
- In the discrete case:

$$F(x) = \mathbb{P}(X \leq x) = \sum_{y \leq x} P(y), \quad x \in \mathbb{R}$$

  and $F(x)$ has jumps of magnitude $P(x)$.
- In the continuous case $P(x) = 0$ so there are no jumps.

Random Variables - Discrete Case - Example



The probability mass function (pmf) $P(x)$ and the cumulative distribution function (cdf) $F(x)$ for Example E3. White circles denote points which are excluded. Picture from [2].

Random Variables - Discrete Case - Examples

Examples:

(1) If X is a discrete random variable and takes the values $\{x_i : i \in I\}$, then its distribution function is $F : \mathbb{R} \to \mathbb{R}$

$$F(x) = \mathbb{P}(X \leq x) = \sum_{\substack{i \in I \\ x_i \leq x}} \mathbb{P}\left(X = x_i\right), x \in \mathbb{R}.$$

(2) Consider the discrete random variable X such that

$X \sim \begin{pmatrix} -1 & 1 & 2 \\ 0.5 & 0.25 & 0.25 \end{pmatrix}$. Then the cumulative distribution function of X is $F : \mathbb{R} \to \mathbb{R}$ given by

$$F(x) = \begin{cases} 0, & \text{if } x < -1 \\ 0.5, & \text{if } -1 \leq x < 1 \\ 0.5 + 0.25 = 0.75, & \text{if } 1 \leq x < 2 \\ 0.5 + 0.25 + 0.25 = 1, & \text{if } 2 \leq x \end{cases}$$

Random Variables - Discrete Case

For every outcome ω, the discrete variable X takes one and only one value x. This makes the events $\{X = x\}$ disjoint and exhaustive, and therefore,

$$\sum_x P(x) = \sum_x \mathbb{P}(X = x) = 1.$$

Note that the cdf $F(x)$ is a non-decreasing function of x, always between 0 and 1, with

$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to +\infty} F(x) = 1.$$

$F(x)$ is constant between any two subsequent values of X. It jumps by $P(x)$ at each possible value x of X.

Random Variables - Discrete Case

Remember that one way to compute the probability of an event is to add probabilities of all the outcomes corresponding to that particular event. Hence, for any set A,

$$\mathbb{P}(X \in A) = \sum_{x \in A} P(x)$$

If A is an interval, its probability can be computed directly from the cdf $F(x)$ that is

$$\mathbb{P}(a < X \leq b) = F(b) - F(a).$$

pmf vs. cdf

- Any pmf $P(x)$ can assign positive probabilities to a finite or countable set only. We need to have $\sum_x P(x) = 1 \Rightarrow$ we can have at most 2 values of x with $P(x) \geq 1/2$, at most 4 with $P(x) \geq 1/4$ etc. Therefore we can list all x for which $P(x) > 0 \rightarrow$ the set of x_i with $P(x_i) > 0$ cannot be an interval (uncountable), it has to be at most countable.
- For all continuous random variables, the probability mass function is always equal to zero i.e. $P(x) = 0$ for all $x \in \mathbb{R}$. We can instead use the *cumulative distribution function* (cdf).

Random Variables - Properties of the cdf

Theorem 9

The cumulative distribution function $F : \mathbb{R} \to \mathbb{R}$ *of a random variable* X *has the following properties:*
(1) $\mathbb{P}(a < X \leq b) = F(b) - F(a)$ *for* $a < b$*;*
(2) $\mathbb{P}(X = b) = F(b) - F(b - 0)$ *for* $b \in \mathbb{R}$*;*
(3) F *is monotonically increasing;*
(4) F *is right-continuous, i.e.* $F(b + 0) = \lim_{x \searrow b} F(x) = F(b)$ *for each* $b \in \mathbb{R}$*;*
(5) $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$.

Remark: The properties (3), (4) and (5) from Theorem 9 characterize a cumulative distribution function, i.e. if a function $F : \mathbb{R} \to [0, 1]$ has these properties, then there exists a probability space and a random variable X which has F as its cumulative distribution function.

LECTURE 5

Random Variables - Continuous Case

- In the continuous case, the cdf $F(x)$ is a continuous function.
- Assume that $F(x)$ has a derivative i.e. $F'(x)$ exists. This is typically true for all commonly used distributions.

Definition 10

Probability density function (pdf, density) is the derivative of the cumulative distribution function (cdf) i.e. $f(x) = F'(x)$.

Definition 11

A distribution is called continuous if it has a density.

Random Variables - Continuous Case

To put it differently:

Definition 12

Let X be a random variable and let $F : \mathbb{R} \to \mathbb{R}$ be its cumulative distribution function. If there exists a function $f : \mathbb{R} \to \mathbb{R}$ such that

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt \text{ for all } x \in \mathbb{R},$$

then f is the probability density function of X. If the random variable X admits a density function, then X is said to be a continuous random variable. We say that it has a continuous distribution.

Random Variables - Continuous Case

- So $F(x)$ is the antiderivative of the density $f(x)$. By the fundamental theorem of calculus we know that

$$\int_a^b f(x)dx = F(b) - F(a) = \mathbb{P}(a \leq X \leq b)$$

- So we have

$$f(x) = F'(x)$$

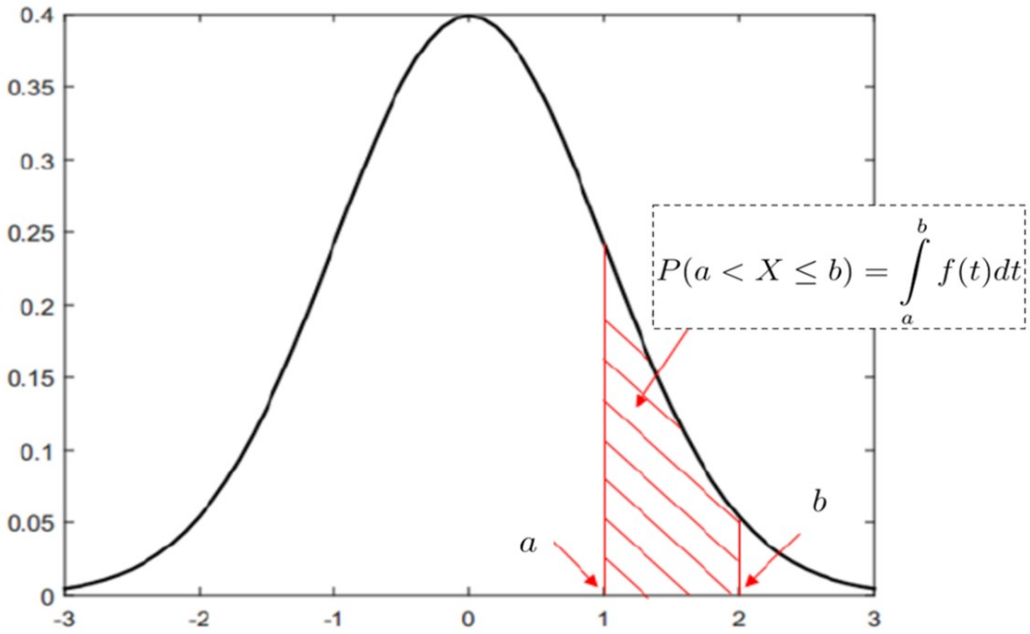
$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx$$

- We can see now that

$$P(x) = \mathbb{P}(x \leq X \leq x) = \int_x^x f = 0$$

- Geometrically, probabilities are represented by areas.

Random Variables - Continuous Case



Probabilities are areas under the density curve

Random Variables - Continuous Case

- The role of the density function for continuous distributions is very similar to the role of the probability mass function (pmf) for discrete distributions. Most concepts can be translated from the discrete case to the continuous case by replacing the pmf $P(x)$ with the pdf $f(x)$ and integrating instead of summing (see table on next slides).

Random Variables - Discrete and Continuous

Distribution	**Discrete**	**Continuous**
Definition	$P(x) = P\{X = x\}$ (pmf)	$f(x) = F'(x)$ (pdf)
Computing probabilities	$\boldsymbol{P}\{X \in A\} = \sum_{x \in A} P(x)$	$\boldsymbol{P}\{X \in A\} = \int_A f(x)dx$
Cumulative distribution function	$F(x) = \boldsymbol{P}\{X \leq x\} = \sum_{y \leq x} P(y)$	$F(x) = \boldsymbol{P}\{X \leq x\} = \int_{-\infty}^{x} f(y)dy$
Total probability	$\sum_x P(x) = 1$	$\int_{-\infty}^{\infty} f(x)dx = 1$

Pmf $P(x)$ in discrete case vs pdf $f(x)$ in continuous case ([2]).

Theorem 13

Let $a, b \in \mathbb{R}, a < b$. If X is a continuous random variable having the cumulative distribution function F and density function f, then the following properties hold:

(1) $f(x) \geq 0$ *for all* $x \in \mathbb{R}$;

(2) $\int_{\mathbb{R}} f(t)dt = 1$;

(3) For $a, b \in \mathbb{R}$ *with* $a < b$ *we have* $\mathbb{P}(X = b) = 0$ *and*

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b)$$
$$= F(b) - F(a) = \int_a^b f(t)dt;$$

(4) For $B \in \mathcal{B}(\mathbb{R})$ *we have* $\mathbb{P}(X \in B) = \int_B f(x)dx$;

(5) F *is continuous and* $F'(x) = f(x)$, *if* F *is derivable in* x.

Random Variables

Remark: Any function $f : \mathbb{R} \to \mathbb{R}$ that satisfies properties (1) and (2) of the above theorem is a density function.

Example:

Let $F : \mathbb{R} \to \mathbb{R}$ be given by $F(x) = \begin{cases} d, & \text{if } x < 0 \\ ax^2 + bx + c, & \text{if } 0 \leq x < 2 \\ e, & \text{if } x \geq 2, \end{cases}$

where $a, b, c, d, e \in \mathbb{R}$ are fixed. Find a, b, c, d, e such that F is the cumulative distribution function of a continuous random variable X with $\mathbb{P}(1 < X < 2) = \frac{3}{4}$. Write the density function of X.

Random Variables

From Theorem 9 we have that $0 = \lim_{x \to -\infty} F(x) = d$ and $1 = \lim_{x \to \infty} F(x) = e$. By Theorem 13 F is continuous, since X is continuous, and thus $c = F(0) = F(0-0) = d = 0$ and $1 = e = F(2) = F(2-0) = 4a + 2b + c = 4a + 2b$. By Theorems 9 and 13, we have

$$\mathbb{P}(1 < X < 2) = F(2) - F(1) = e - a - b - c = 1 - a - b = \frac{3}{4}.$$

Hence, we deduce that $a = \frac{1}{4}, b = 0, c = 0, d = 0, e = 1$.

We point out that $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = \begin{cases} \frac{x}{2}, & \text{if } x \in (0,2) \\ 0, & \text{otherwise} \end{cases}$ is the density function of X.

Independent Random Variables

Definition 14

Consider the discrete random variables

$$X \sim \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I} \quad \text{and} \quad Y \sim \begin{pmatrix} y_j \\ q_j \end{pmatrix}_{j \in J}.$$

X *and* Y *are independent random variables if*

$$\mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i)\,\mathbb{P}(Y = y_j) \quad \forall i \in I, j \in J.$$

Notation:

$$\mathbb{P}(X = x_i, Y = y_j) := \mathbb{P}(\{X = x_i\} \cap \{Y = y_j\}), \forall i \in I, j \in J.$$

Independent Random Variables

Remark: Consider the events

$$A_i = \{X = x_i\} = \{\omega \in \Omega : X(\omega) = x_i\}$$

and

$$B_j = \{Y = y_j\} = \{\omega \in \Omega : Y(\omega) = y_j\}, i \in I, j \in J.$$

The discrete random variables X and Y are independent $\iff \forall (i,j) \in I \times J$ the events A_i and B_j are independent.

Random Vectors

Definition 15

If X and Y are random variables, then the pair (X, Y) is called a random vector. Its distribution is called the joint distribution of X and Y. Individual distributions of X and Y are then called marginal distributions.

Similarly to a single variable, the joint distribution of a vector is a collection of probabilities for a vector (X, Y) to take a value (x, y). Recall that two vectors are equal

$$(X, Y) = (x, y)$$

if $X = x$ **and** $Y = y$. Here "**and**" means *intersection*, so the joint probability mass function of X and Y is

$$P(x, y) = \mathbb{P}\{(X, Y) = (x, y)\} = \mathbb{P}(X = x \cap Y = y).$$

Random Vectors

Note that $\{(X, Y) = (x, y)\}$ are exhaustive and mutually exclusive events for different pairs (x, y), therefore

$$\sum_x \sum_y P(x, y) = 1.$$

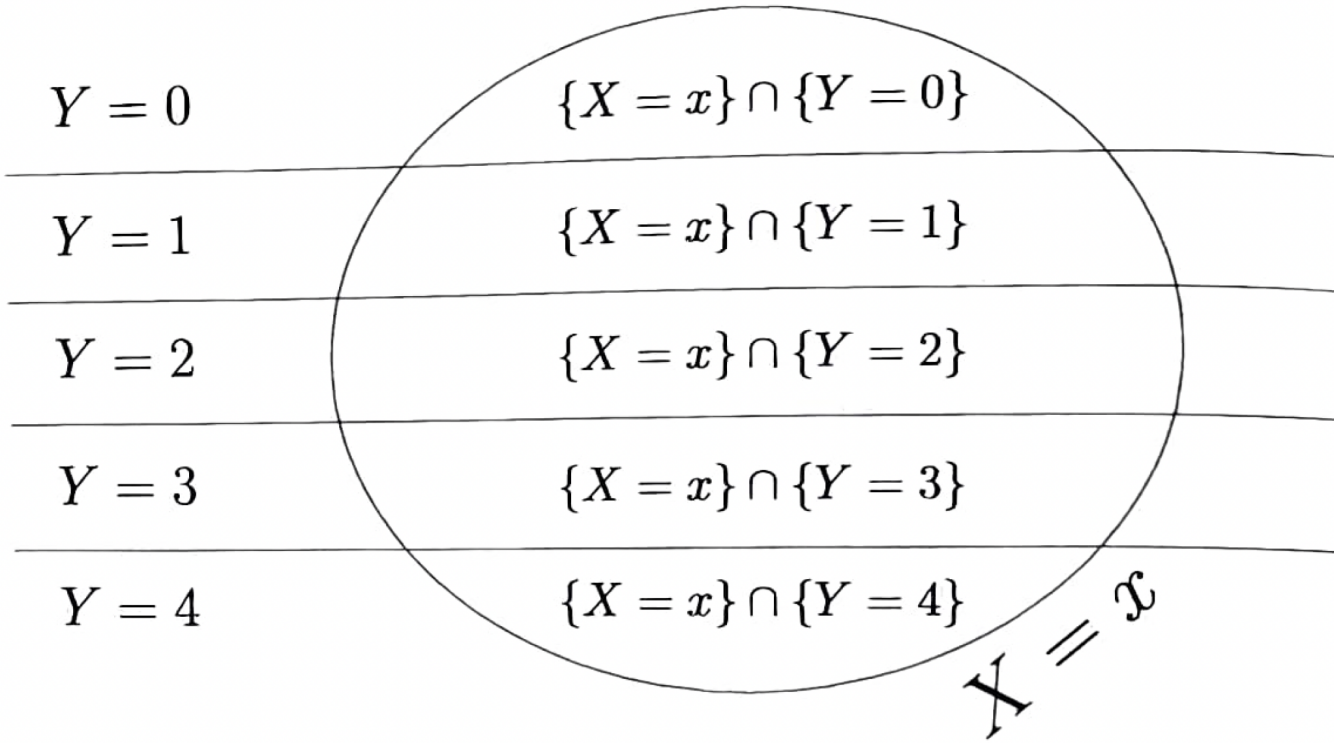
The joint distribution of (X, Y) carries the complete information about the behavior of this random vector. In particular, the marginal probability mass functions of X and Y can be obtained from the joint pmf by the **addition rule**:

$$P_X(x) = \mathbb{P}(X = x) = \sum_y P_{(X,Y)}(x, y)$$

$$P_Y(y) = \mathbb{P}(Y = y) = \sum_x P_{(X,Y)}(x, y)$$

$\Rightarrow$ to get the marginal pmf of one variable, we add the joint probabilities over all values of the other variable.

Random Vectors



Addition rule: computing marginal probabilities from joint distributions ([2]).

LECTURE 6

Discrete Random Vectors

- $\mathbb{X} = (X_1, \ldots, X_n)$ is a discrete random vector, if each component $X_1, \ldots X_n$ is a discrete random variable.
- Discrete random vectors are characterized by their joint probability distribution. For example a discrete random vector with two components has distribution

$$\mathbb{X} = (X, Y) \sim \begin{pmatrix} (x_i, y_j) \\ p_{ij} \end{pmatrix}_{(i,j) \in I \times J}$$

where $I, J \subseteq \mathbb{N}$ are index sets,
$p_{ij} = \mathbb{P}\left((X, Y) = (x_i, y_j)\right) = \mathbb{P}\left(X = x_i, Y = y_j\right)$,

$$p_{ij} > 0 \forall i \in I, j \in J, \text{ where } \sum_{(i,j) \in I \times J} p_{ij} = 1.$$

Discrete Random Vectors

The joint probability mass function of the discrete random vector $\mathbb{X} = (X, Y)$ is often given by

Y / X	$\cdots$	y_j	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	$\cdots$	p_{ij}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

, where

$p_{ij} = \mathbb{P}\left((X, Y) = (x_i, y_j)\right)$.

Discrete Random Vectors

- If X and Y are independent discrete random variables, then

$$p_{ij} = \mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i)\,\mathbb{P}(Y = y_j) \quad \forall i \in I, j \in J.$$

- If X and Y are independent discrete random variables, and their probability distributions are known, then one can compute p_{ij} for (X, Y).

If the joint probability mass function of the discrete random vector $\mathbb{X} = (X, Y)$ is known

$$(X, Y) \sim \begin{pmatrix} (x_i, y_j) \\ p_{ij} \end{pmatrix}_{(i,j) \in I \times J}$$

how can we compute the probability distribution of X and Y, respectively?

Discrete Random Vectors

Answer:

$$\mathbb{P}(X = x_i) = \sum_{j \in J} p_{ij} \quad \forall i \in I$$

$$\mathbb{P}(Y = y_j) = \sum_{i \in I} p_{ij} \quad \forall j \in J.$$

Random Vectors - Joint Distribution

Definition 16

$(X_1, ..., X_n)$ is a random vector if each component X_i, $i \in \{1, 2, ..., n\}$ is a random variable.

Definition 17

The function $F : \mathbb{R}^n \to \mathbb{R}$ defined by

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) \text{ for each } x_1, \ldots, x_n \in \mathbb{R}$$

is called the joint cumulative distribution function (or joint distribution function) of the random vector $(X_1, \ldots, X_n)$.

Remark: Two random vectors have the same distribution if and only if they have the same joint cumulative distribution function.

Joint PDF and joint CDF

Joint Cumulative Distribution Function (CDF)

The joint cumulative distribution function $F_{X,Y}(x,y)$ of two random variables X and Y is defined as:

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y).$$

Remark (continuous RV):

Joint Probability Density Function (PDF)

The joint probability density function $f_{X,Y}(x,y)$ is defined using the second partial derivative of the joint CDF $F_{X,Y}(x,y)$, that is:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x\, \partial y} F_{X,Y}(x,y)$$

This can also be written as:

$$f_{X,Y}(x,y) = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y} F_{X,Y}(x,y)\right) = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x} F_{X,Y}(x,y)\right).$$

Random Vectors - Joint Distribution

Example 1:

The random vector (X, Y) is given by the *contingency table*:

X \ Y	0	3
-2	0.4	0.3
4	0.2	0.1

We compute

$$F_{(X,Y)}(3,5) = \mathbb{P}(X \leq 3, Y \leq 5) = \mathbb{P}(X = -2, Y \in \{0,3\}) = 0.7;$$

$$F_{(X,Y)}(5,2) = \mathbb{P}(X \leq 5, Y \leq 2) = \mathbb{P}(X \in \{-2,4\}, Y = 0) = 0.6\,.$$

$$F_{(X,Y)}(4,3) = \mathbb{P}(X \leq 4, Y \leq 3) = \mathbb{P}(X \in \{-2,4\}, Y \in \{0,3\}) = 1\,.$$

$$F_{(X,Y)}(-4,-3) = \mathbb{P}(X \leq -4, Y \leq -3) = \mathbb{P}(\emptyset) = 0.$$

Example 2:

Suppose $X, Y \in \{0, 1\}$, and their joint pmf is:

$\mathbb{P}(X = x, Y = y)$	$y = 0$	$y = 1$
$x = 0$	0.1	0.2
$x = 1$	0.3	0.4

The joint cdf $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ is computed by summing the pmf over the appropriate region. For example:

$$F_{X,Y}(0, 0) = \mathbb{P}(X = 0, Y = 0) = 0.1$$

$$F_{X,Y}(0, 1) = \mathbb{P}(X = 0, Y = 0) + \mathbb{P}(X = 0, Y = 1) = 0.1 + 0.2 = 0.3$$

$$F_{X,Y}(1, 0) = \mathbb{P}(X = 0, Y = 0) + \mathbb{P}(X = 1, Y = 0) = 0.1 + 0.3 = 0.4$$

$$F_{X,Y}(1, 1) = \text{sum of all entries} = 1.0$$

Continuous case - joint cdf and joint pdf

Example:

Let the joint cdf of X and Y be:

$$F_{X,Y}(x,y) = \begin{cases} xy, & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0, & \text{otherwise} \end{cases}$$

The joint pdf is the mixed partial derivative of the joint cdf:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x\, \partial y} F_{X,Y}(x,y).$$

Computing the derivatives:

$$\frac{\partial}{\partial x} F_{X,Y}(x,y) = y, \quad \frac{\partial^2}{\partial x\, \partial y} F_{X,Y}(x,y) = 1.$$

So the joint pdf is:

$$f_{X,Y}(x,y) = \begin{cases} 1, & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

Random Vectors - Joint Distribution

Theorem 18

If (X, Y) is a random vector having the joint cumulative distribution function F, then the following properties hold:

(1) F *is monotone increasing in each argument.*

(2) F *is right-continuous in each argument.*

(3) $\lim\limits_{\substack{x\to\infty \\ y\to\infty}} F(x, y) = 1.$

(4) We have

$$\lim_{x\to-\infty} F(x, y) = 0 \quad \text{for all } y \in \mathbb{R}$$

and

$$\lim_{y\to-\infty} F(x, y) = 0 \quad \text{for all} \quad x \in \mathbb{R}.$$

(5) It holds that

$$\lim_{y\to\infty} F_{(X,Y)}(x, y) = F_X(x)$$

$$\lim_{x\to\infty} F_{(X,Y)}(x, y) = F_Y(y).$$

Expectation and Variance of a Random Variable

- The *distribution* of a random variable X encodes the entire information about the behaviour of X.
- This information can be summarised in a couple of basic characteristics which provide information about the average value of a random variable, the most likely value, the spread of the random variable, etc $\rightarrow$ *expectation, variance, standard deviation, covariance, correlation*.
- In general, the random variable X can take different values with different probabilities. The expected value is not just the average of its values, but a *weighted average*.

Expectation of a Random Variable

Definition 19

The expectation or the expected value of a random variable X is its mean, or the average value. If X is a discrete random variable taking the values $\{x_i : i \in I\}$, then the expectation of X is the number

$$\mu = \mathbb{E}[X] = \sum_{i \in I} x_i \mathbb{P}(X = x_i) \tag{11}$$

if the series is absolutely convergent i.e. $\sum_{i \in I} |x_i| \mathbb{P}(X = x_i) < \infty$. If a continuous random variable X has f as its density function, then the expectation (or mean value, or expected value) of X is the number

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f(x) dx \tag{12}$$

if the integral is absolutely convergent, i.e. $\int_{-\infty}^{\infty} |t| f(t) dt < \infty$.

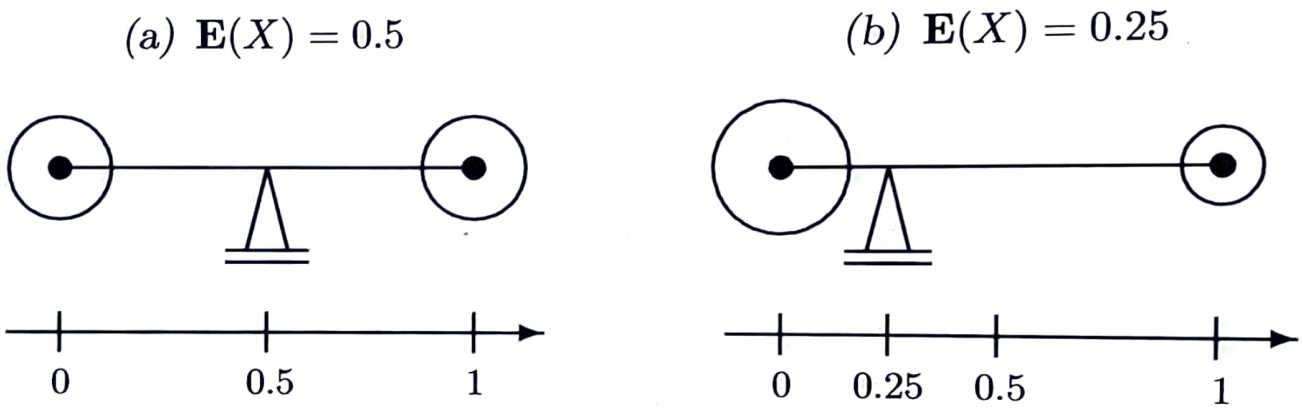
Expectation of a Random Variable

Example ([2]): Consider a random variable which takes values 0 with probability $1/2$ and value 1 with probability $1/2$, $P(0) = P(1) = 1/2$. If we observe this variable many times, we shall see that $X = 1$ about 50% of times and $X = 0$ about 50% of times $\Rightarrow$ it makes sense to have $\mathbb{E}[X] = 1/2$.

How do unequal probabilities for $X = 0$ and $X = 1$ affect the expectation $\mathbb{E}[X]$?

Assume that $P(0) = 0.75$ and $P(1) = 0.25 \Rightarrow$ in the long run, X is 1 only $1/4$ times, and it equals 0 for the rest of times. Suppose we earn 1 RON every time we get $X = 1 \Rightarrow$ on average, we earn 1 every four times. To put it differently: we earn 0.25 per each observation. In this situation $\mathbb{E}[X] = \frac{1}{4}$.

Expectation of a Random Variable



Expectation as center of gravity. ([2])

(a) Equal masses $= \frac{1}{2}$ at points 0 and 1 connected via a firm and weightless rod. The *masses* represent the probabilities $P(0)$ and $P(1)$, respectively. The system is balanced at its *center of gravity* $= \frac{1}{2}$.

(b) The masses at 0 and 1 are unequal, according to $P(0)$ and $P(1)$. The system is balanced at $\frac{1}{4}$.

Expectation of a Random Variable

- Likewise, formula (12) characterises the *center of gravity* for a physical system/model with masses $P(x)$ allocated at points x.
- $\mathbb{E}[X]$ is the *center of mass/gravity center/balance point* for the distribution of X. This means that on average, the values of X tend to cluster around the expected value $\mathbb{E}[X]$.
- The variable X itself is random and it takes different values with different probabilities $P(x)$. However, the expectation $\mathbb{E}[X]$ is a non-random quantity.

Expectation of a Random Variable - A good Forecast for X

- *Forecast* of a random variable: predicting the distribution or behavior of that random variable based on available information or on a mathematical/statistical model.
- $\mathbb{E}[X]$ is a good forecast for X:
  - On average, the values of X tend to cluster around $\mathbb{E}[X]$.
  - Balancing errors: When deviations occur (i.e. X takes values different from the expected value), these deviations tend to balance out over many repetitions. Some values may be higher than the expected value, while others may be lower, but on average, they cancel each other out.
  - Law of Large Numbers: The Law of Large Numbers states that as the number of repetitions of a random experiment increases, the average of the outcomes converges to the expected value. This means that in the long run, the observed average tends to approach the expected value.

Expectation of a Random Variable

Remark: The mean value of a random variable is the most frequently used measure of central tendency of its values, a measure of the center of the distribution.

Example: The expectation of the continuous random variable X with uniform distribution on the interval $[a, b]$ is

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f(x) dx = \frac{1}{b-a} \int_a^b x dx = \frac{a+b}{2},$$

where f is the density function.

Expectation of a Random Variable

Theorem 20

If X and Y are both discrete or both continuous random variables which have the expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ respectively, then the following properties hold:

(1) $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$ *for all* $a, b \in \mathbb{R}$.

(2) $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

(3) If $X(\omega) \geq 0$ *for each* $\omega \in \Omega$, *then* $\mathbb{E}[X] \geq 0$.

(4) If $X[\omega] \geq Y[\omega]$ *for each* $\omega \in \Omega$, *then* $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

(5) If $h : \mathbb{R} \to \mathbb{R}$ *is a function such that* $h(X)$ *is a random variable then: If* X *is a discrete random variable taking the values* $\{x_i : i \in I\}$, *then*

$$\mathbb{E}\left[h(X)\right) = \sum_{i \in I} h(x_i)\mathbb{P}(X = x_i),$$

while for a continuous random variable X *with density function* f, *we have*

$$\mathbb{E}\left[h(X)\right] = \int_{\mathbb{R}} h(x)f(x)dx.$$

Expectation of a Random Variable

- Examples of functions of a RV: downloading time depends on the connection speed, profit of a company depends on the number of items sold etc.
- The expected value $\mathbb{E}[X]$ shows where the average value of a random variable is located, where the variable X is *expected* to be, plus or minus some error. But how much can these error be and how much a variable can *vary* around its expected value?

Expectation of a Random Variable

- Example ([2]): Consider 2 people: one receives either 0 or 100 emails with 50% chance of each, while the other receives either 48 or 52 emails with 50% chance. What is common in these 2 scenarios? How do they differ?

  We can see that
  $$\mathbb{E}[X] = \mathbb{E}[Y] = 50.$$

  Nonetheless, in the second case the number of emails is always close to 50, while in the first case it always differs from it by $50 \Rightarrow$ the second random variable Y is more *stable* (it has *low variability*), while the first random variable X has *high variability*.

$\Rightarrow$ the *variability* of a random variable is measured by its distance from its expected value $\mathbb{E}[X]$, that is $X - \mathbb{E}[X]$. This is random and has expected value 0, so we cannot efficiently use it to provide information about the distribution $\rightarrow$ use $(X - \mathbb{E}[X])^2$ which makes all deviations positive.

List of classical discrete probability distributions

See a Python notebook here.

Discrete uniform distribution

$X \sim Unif(n), n \in \mathbb{N}^*$

$$X \sim \begin{pmatrix} 1 & 2 & \dots & n \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix}$$

Example: Let X be the outcome of the roll of a die.

$$\Rightarrow X \sim \begin{pmatrix} 1 & 2 & \dots & 6 \\ \frac{1}{6} & \frac{1}{6} & \dots & \frac{1}{6} \end{pmatrix}$$

Python: discrete uniform distribution

```
import numpy as np

# Generating random values from a uniform distribution
def unidrnd(n, *args):
    return np.random.uniform(*args, size=n)

# Generating random integers
def randi(n, *args):
    return np.random.randint(*args, size=n)

# Example usage
n = 10  # Number of random values to generate
uniform_values = unidrnd(n, 0, 10)  # Generates n random values from \
                                    # uniform distribution between 0 and 10
random_integers = randi(n, 0, 100)   # Generates n random integers between 0 and 100

print("Random values from uniform distribution:", uniform_values)
print("Random integers:", random_integers)
```

Bernoulli distribution

$X \sim Bernoulli(p)$, $p \in (0,1)$

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

Example: Bernoulli Trials - Repeated independent trials of an experiment, such that there are only two possible outcomes for each trial - which we classify as either *success* or *failure* - and their probabilities remain the same throughout the trials are called **Bernoulli trials**.
We denote the two possible outcomes by A (*success*) having the probability $p = \mathbb{P}(A)$, respectively, $\bar{A}$ (*failure*), with probability $1 - p = \mathbb{P}(\bar{A})$.
Recall example 10: $\mathbb{1}_A = 0 \Leftrightarrow \bar{A}$ emerges and $\mathbb{1}_A = 1 \Leftrightarrow A$ emerges;
$\Longrightarrow \mathbb{1}_A \sim Bernoulli(p)$ with $p := \mathbb{P}(A)$

$$\mathbb{1}_A \sim \begin{pmatrix} 0 & 1 \\ 1-\mathbb{P}(A) & \mathbb{P}(A) \end{pmatrix}.$$

Python: Bernoulli distribution

```
import numpy as np

n = 10000
p = 0.3
r = np.random.rand(1, n)
X = (r <= p).astype(int)  # data vector with Bernoulli(p) distribution

print(X)
```

Binomial distribution

$X \sim Bino(n, p), n \in \mathbb{N}^*, p \in (0, 1)$
The binomial distribution describes the number of *successes* in a series of independent Bernoulli trials.

- *success* appears with probability p, *failure* with probability $1 - p$;
- the experiment is repeated n times;
- the random variable X denotes the number of successes in n repetitions of the experiment

$\Rightarrow$ the possible values of X are $0, 1, \ldots, n$.
A random variable X with range of values $\{0, \ldots, n\}$ is called **binomially distributed** with parameters $n \geq 1$ and $p \in (0, 1)$ if its probability mass function is

$$\mathbb{P}(X = k) = C_n^k p^k (1 - p)^{n-k}, \quad k \in \{0, \ldots, n\}.$$

Binomial distribution

Remember the binomial formula

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}.$$

where $C_n^k p^k (1-p)^{n-k}$ represents the coefficient of x^k in the binomial expansion

$$(px+1-p)^n = \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} x^k \,.$$

Binomial distribution

Example:

(1) A die is rolled 10 times. Let X be the random variable that shows how many times the number 6 appeared $\Longrightarrow$ $X \sim Bino(10, \frac{1}{6})$.

(2) In an urn are n_1 white marbles and n_2 black marbles.
Experiment: A marble is drawn randomly, its color is noted and then it is put back into the urn, i.e. we sample *with replacement* The experiment is repeated n times; consider the random variables:
$X_1 =$ the number of white marbles drawn; $X_2 =$ the number of black marbles drawn;
$\Longrightarrow$ $X_1 \sim Bino(n, p_1)$ with $p_1 = \frac{n_1}{n_1+n_2}$; $X_2 \sim Bino(n, p_2)$ with $p_2 = \frac{n_2}{n_1+n_2}$.

(3) In each scanning cycle a radar registers the presence of an object with probability 0.9. The probability that the radar detects an object at least once in 4 scanning cycles is

$$\sum_{k=1}^{4} C_4^k 0.9^k 0.1^{4-k} = 1 - 0.1^4 = 0.9999\,.$$

Python: Binomial distribution

```
def binornd(n, p, size=None):
    """
    Generate random values from a binomial distribution.

    Parameters:
        n (int): Number of trials.
        p (float): Probability of success.
        size (int or tuple of ints, optional): Output shape.
                                               If None, a single value is returned.

    Returns:
        int or ndarray: Random values from Binomial distribution.
    """
    return np.random.binomial(n, p, size)
```

Hypergeometric distribution

$X \sim Hyge(n, n_1, n_2),\ n, n_1, n_2 \in \mathbb{N}^*, n \leq n_1 + n_2.$

A random variable X has **hypergeometric distribution** with parameters $n, n_1, n_2 \in \mathbb{N}^*$ and $n \leq n_1 + n_2$, if its probability mass function is

$$\mathbb{P}(X = k) = \frac{C_{n_1}^k C_{n_2}^{n-k}}{C_{n_1+n_2}^n} \quad \text{for} \quad k \in \{0, 1, \ldots, \min\{n, n_1\}\}.$$

- It describes the probability of obtaining a certain number of successes (or items of interest) in a fixed-size sample drawn from a finite population, where each draw is made without replacement.

Hypergeometric distribution

Example:

(1) In a box there are n_1 white marbles and n_2 black marbles of the same size. We randomly draw a marble from the box and **do not place** the marble back into the box, i.e. we sample *without replacement*. If X denotes the number of white marbles in $n \leq n_1 + n_2$ repetitions of this experiment, then $X \sim Hyge(n, n_1, n_2)$.

(2) A team of 5 persons is randomly selected from a group of 10 women and 11 men. The probability that more women than men are selected is

$$\frac{C_{10}^5 C_{11}^0 + C_{10}^4 C_{11}^1 + C_{10}^3 C_{11}^2}{C_{21}^5}.$$

Python: Hypergeometric distribution

```
import numpy as np

def hygernd(n1, n2, n, size=None):
    """
    Generate random values from a hypergeometric distribution.

    Parameters:
        n1 (int): Number of successes in the population.
        n2 (int): Number of failures in the population.
        n (int): Number of draws.
        size (int or tuple of ints, optional): Output shape. If None, a single value is ret

    Returns:
        int or ndarray: Random values from Hypergeometric distribution.
    """
    return np.random.hypergeometric(n1, n2, n, size)
```

Geometric distribution

$X \sim Geo(p)$, $p \in (0,1)$
Consider an infinite sequence of Bernoulli trials in which the outcome of any trial is either success with probability p or failure with probability $1-p$, where $p \in (0,1)$, and let X be a random variable denoting the total number of failures that occur *before* the first success. Then X has **geometric distribution** with parameter p. Its probability mass function is

$$\mathbb{P}(X = k) = p(1-p)^k \quad \text{for} \quad k \in \{0, 1, 2...\} = \mathbb{N}.$$

For $k \in \mathbb{N}$ the number $p(1-p)^k$ represents the coefficient of x^k in the *geometric series*

$$\frac{p}{1-(1-p)x} = \sum_{k=0}^{\infty} p(1-p)^k x^k$$

which is convergent if (the ratio) $|(1-p)x| < 1$.

- Models the number of trials needed to achieve the first success in a sequence of independent Bernoulli trials.

Geometric distribution

Example: Suppose we shuffle a standard deck of 52 cards, and we turn over the top card. We put the card back in the deck and reshuffle. We repeat this process until we get a Jack; let X denote the number of trials before we get a Jack. $\implies X \sim Geo\left(\frac{4}{52}\right)$; if Y denotes the number of trials until (inclusive) we get a Jack, then

$$\mathbb{P}(Y = k) = \frac{4}{52}\left(\frac{48}{52}\right)^{k-1} \quad \text{for} \quad k \in \{1, 2...\} \implies Y - 1 \sim Geo\left(\frac{4}{52}\right).$$

Python: Geometric distribution

```
import numpy as np

def geornd(p, size=None):
    """
    Generate random values from a geometric distribution.

    Parameters:
        p (float or array_like of floats): The probability of success for each trial.
        size (int or tuple of ints, optional): Output shape.
        If None, a single value is returned.

    Returns:
        ndarray or scalar: Random values from the geometric distribution.

    """
    return np.random.geometric(p, size=size)

# Example usage
p = 0.3  # Probability of success
num_values = 10  # Number of random values to generate

random_geo_values = geornd(p, size=num_values)
print("Random values from geometric distribution:", random_geo_values)
```

Poisson distribution

$X \sim Poiss(\lambda),\ \lambda > 0.$
A random variable X has **Poisson distribution** with parameter $\lambda > 0$, if its probability mass function is

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{for} \quad k \in \{0, 1, 2, \ldots\}.$$

A random variable X having Poisson distribution usually denotes the total number of occurrences of some phenomenon during a fixed period of time or within a fixed region in space. For example, X represents:

- the number of accidents which took place on a certain place in a certain time interval;
- the number of telephone calls received at a switchboard during a fixed period of time
- the number of particles emitted from a radioactive source which strike a certain target during a fixed period of time.

Poisson distribution

Example: The number of failures X in a certain electricity power station during a period of 120 days is modeled using a Poisson distribution with parameter $\lambda = 0.02$. The probability that there are at least two failures during the 120 days is

$$1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) = 1 - 1.02 \cdot e^{-0.02} \approx 0.0002.$$

Python: Poisson distribution

```
import numpy as np

def poissrnd(lmbda, size=None):
    """
    Generate random values from a Poisson distribution.

    Parameters:
        lmbda (float or array_like of floats): Expectation of interval. Should be >= 0.
        size (int or tuple of ints, optional): Output shape.
        If None, a single value is returned.

    Returns:
        ndarray or scalar: Random values from the Poisson distribution.

    """
    return np.random.poisson(lmbda, size=size)

# Example usage
lmbda = 2.5  # Expectation of interval
num_values = 10  # Number of random values to generate

random_poiss_values = poissrnd(lmbda, size=num_values)
print("Random values from Poisson distribution:", random_poiss_values)
```

LECTURE 7

Variance of a Random Variable

Definition 21

Let X be a random variable with expectation $\mathbb{E}[X]$. The variance or dispersion of X is the number

$$\boxed{\sigma^2 = V(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]} \tag{13}$$

if the expectation of $(X - \mathbb{E}[X])^2$ exists.

The value $\sigma = \sqrt{V(X)}$ is called the standard deviation of X.

- For a discrete random variable this is given by

$$\sum_{i \in I} (x_i - \mu)^2 \mathbb{P}(X = x_i).$$

- The variance and standard deviation are measures of the horizontal spread or dispersion of the random variable X around its expected value $\mathbb{E}(X)$.

Expectation and Variance of a Random Variable

Theorem 22

If X and Y are random variables, then the following properties hold:

(1) $V(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

(2) $V(aX + b) = a^2 V(X)$ *for all* $a, b \in \mathbb{R}$.

Proof.

(1) By the linearity property of the expectation we have

$$V(X) = \mathbb{E}\left[X^2 - 2\mathbb{E}[X]X + (\mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2.$$

(2) Using Theorem 20 we obtain

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

Covariance and Correlation Coefficient

The expected value, variance, and standard deviation characterise the distribution of a single random variable $\rightarrow$ we need a measure also for two random variables.

$\Rightarrow$ *covariance* and *correlation*

Covariance and Correlation Coefficient

Definition 23

Let X and Y be two random variables, and let $\mathbb{E}[X]$ and $\mathbb{E}[Y]$, respectively be their expectations. The covariance of the random variables X and Y is the number (if it exists)

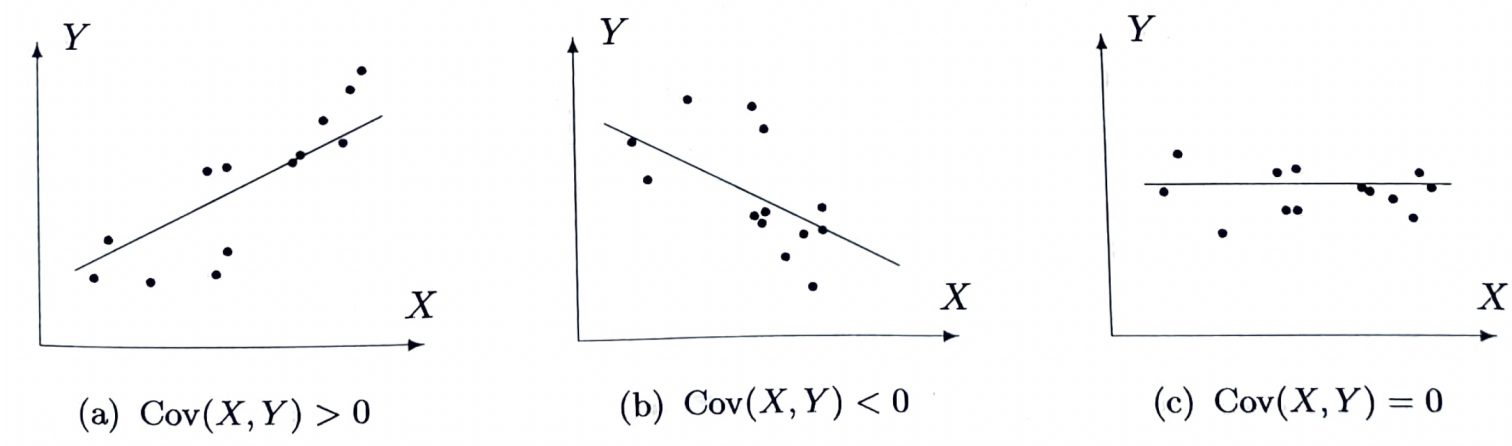
$$cov(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])\right]. \tag{14}$$

The correlation coefficient of X and Y is defined by

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} \tag{15}$$

if $cov(X,Y), V(X), V(Y)$ exist and $V(X) \neq 0, V(Y) \neq 0$.

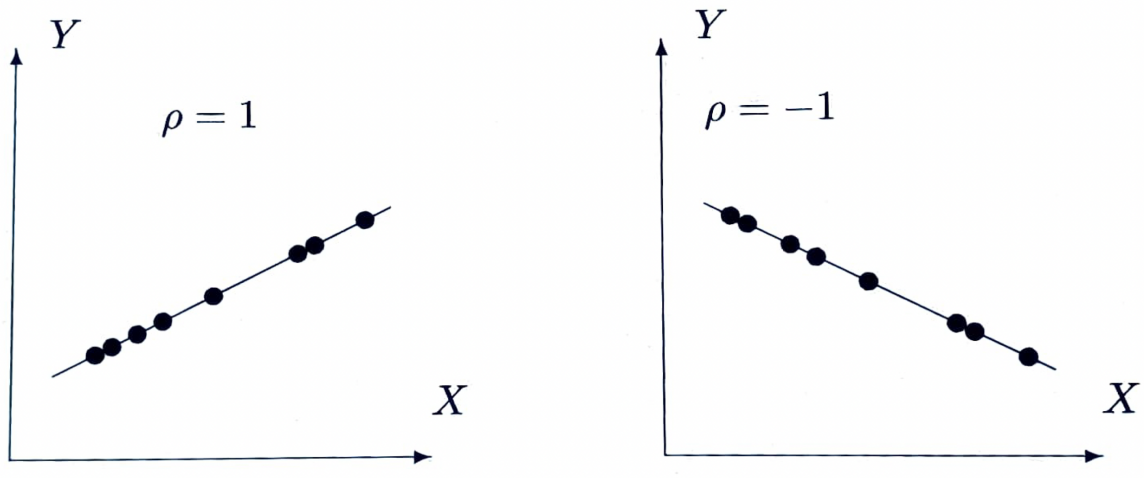
Covariance and Correlation Coefficient



Positive, negative, and zero covariance ([2])

- If $Cov(X, Y) > 0$ then positive deviations $(X - \mathbb{E}[X])$ are more likely to be multiplied by positive $(Y - \mathbb{E}[Y])$ and negatie $(X - \mathbb{E}[X])$ are more likely to be multiplied by negative $(Y - \mathbb{E}[Y])$. That is, large X implies large Y and small X implies small Y.
- If $Cov(X, Y) < 0$ then large X implies small Y and small X implies large Y.

Covariance and Correlation Coefficient



Perfect correlation ([2])

- The correlation coefficient is a statistical measure of the strength of a linear relationship between the two variables X and Y, it is a rescaled, normalised covariance. Its values can range from -1 to 1. A correlation coefficient of 0 means there is no linear relationship between the two variables i.e. the two random variables are uncorrelated.
- Covariance has a measurement unit (units of X multiplied by units of Y) $\rightarrow$ in order to be able to decide whether X and Y are strongly or weakly correlated, we have to compare $Cov(X, Y)$ with the magnitude of X and Y. Correlation coefficient is dimensionless.

Covariance and Correlation Coefficient

Theorem 24

If X, Y and Z are random variables, then the following properties hold:

(1) $cov(X,Y) = \mathbb{E}[X \cdot Y] - \mathbb{E}[X]\mathbb{E}[Y]$.

(2) $V(aX+bY) = a^2V(X) + b^2V(Y) + 2ab\,cov(X,Y)$ *for all* $a, b \in \mathbb{R}$.

(3) $\rho^2(X,Y) \leq 1$.

In particular, for two independent random variables X and Y, we have that $cov(X,Y) = 0$ and therefore

$$V(X+Y) = V(X) + V(Y).$$

Covariance and Correlation Coefficient

Proof.

(1) By the linearity property of the expectation (see Theorem 20) we can write

$$\begin{aligned} cov(X,Y) &= \mathbb{E}\Big[X \cdot Y - \mathbb{E}[X]Y - \mathbb{E}[Y]X + \mathbb{E}[X]\mathbb{E}[Y]\Big] \\ &= \mathbb{E}[X \cdot Y] - \mathbb{E}[X]\mathbb{E}[Y]. \end{aligned}$$

(2) We have

$$\begin{aligned} V(aX+bY) &= \mathbb{E}\left[\big(aX+bY-\mathbb{E}[aX+bY]\big)^2\right] \\ &= a^2\mathbb{E}\Big[\big(X-\mathbb{E}[X]\big)^2\Big] + b^2\mathbb{E}\Big[\big(Y-\mathbb{E}[Y]\big)^2\Big] \\ &+ 2ab\mathbb{E}\Big[\big(X-\mathbb{E}[X]\big)\big(Y-\mathbb{E}[Y]\big)\Big] \\ &= a^2V(X) + b^2V(Y) + 2ab\,cov(X,Y). \end{aligned}$$

(3) It holds $\Big(t(X-\mathbb{E}[X]) + Y - \mathbb{E}[Y]\Big)^2 \geq 0$ for each $\omega \in \Omega$ and each $t \in \mathbb{R}$.

Covariance and Correlation Coefficient

Then by Theorem 20 we have

$$\mathbb{E}\Big[\big(t(X-\mathbb{E}[X])+Y-\mathbb{E}[Y]\big)^2\Big] \geq 0 \text{ for all } t \in \mathbb{R}.$$

We rewrite as

$$V(X)t^2 + 2cov(X,Y)t + V(Y) \geq 0 \quad \forall t \in \mathbb{R}.$$

Since $V(X) > 0$ this inequality holds if and only if $4cov^2(X,Y) - 4V(X)V(Y) \leq 0$. So, $\rho^2(X,Y) \leq 1$.

Moments of a Random Variable

Definition 25

Let $k \in \mathbb{N}^\star$ and let X be a random variable. The number $\mathbb{E}[X^k]$ (if it exists) is called the moment of order k of X, while the number $\mathbb{E}\left[|X|^k\right]$ (if it exists) is called the absolute moment of order k of X.

The expectation (if it exists) $\mathbb{E}\left[(X - \mathbb{E}[X])^k\right]$ is called the central moment of order k of X.

Discrete	Continuous
$\mathbf{E}(X) = \sum_x xP(x)$	$\mathbf{E}(X) = \int xf(x)dx$
$\begin{aligned}\mathrm{Var}(X) &= \mathbf{E}(X-\mu)^2 \\ &= \sum_x (x-\mu)^2 P(x) \\ &= \sum_x x^2 P(x) - \mu^2\end{aligned}$	$\begin{aligned}\mathrm{Var}(X) &= \mathbf{E}(X-\mu)^2 \\ &= \int (x-\mu)^2 f(x)dx \\ &= \int x^2 f(x)dx - \mu^2\end{aligned}$
$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbf{E}(X-\mu_X)(Y-\mu_Y) \\ &= \sum_x \sum_y (x-\mu_X)(y-\mu_Y)P(x,y) \\ &= \sum_x \sum_y (xy)P(x,y) - \mu_x\mu_y\end{aligned}$	$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbf{E}(X-\mu_X)(Y-\mu_Y) \\ &= \iint (x-\mu_X)(y-\mu_Y)f(x,y)\,dx\,dy \\ &= \iint (xy)f(x,y)\,dx\,dy - \mu_x\mu_y\end{aligned}$

Expected value, variance, covariance for discrete and continuous distributions ([2]).

List of classical continuous probability distributions

See a Python notebook here.

Continuous uniform distribution: $X \sim Unif([a, b])$

- Plays a central role: a random variable with any other type of distribution can be generated from a uniform random variable (a RV with uniform distribution).
- Used when we pick a value arbitrarily from a given interval $[a, b]$, without any particular preference for higher, lower, or medium values $\Rightarrow$ it has constant density

$$f(x) = \frac{1}{b-a} \quad x \in (a, b) \tag{16}$$

$\Rightarrow$ the rectangular area below the density graph is equal to 1.

- There is no uniform distribution on the entire $\mathbb{R}$ (it's impossible to have a constant pdf that integrates to 1 over the whole $(-\infty, \infty)$) $\Rightarrow$ when we want to choose a random number from $(-\infty, \infty)$ we cannot do it uniformly.
- E.g: height of adults in a certain range (any height within that range is equally likely), temperature in a specific region, errors in a code, etc.

Continuous uniform distribution - The Uniform Property

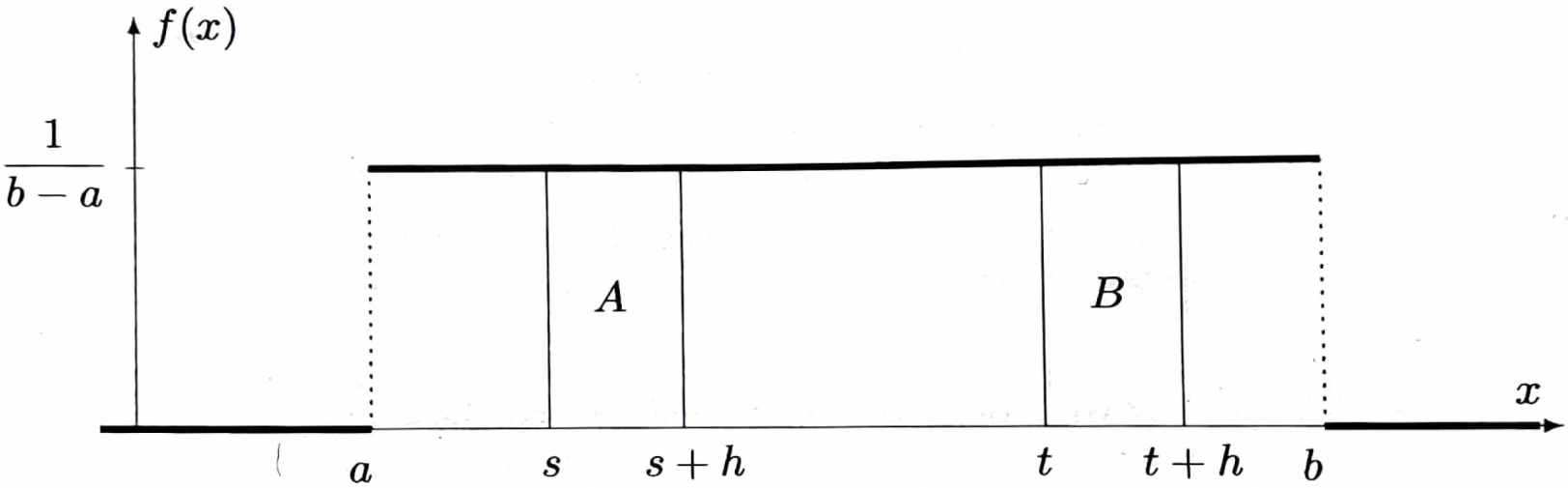
For any $h > 0$ and $t \in [a, b-h]$, the probability

$$\mathbb{P}(t < X < t+h) = \int_t^{t+h} \frac{1}{b-a} dx = \frac{h}{b-a}$$

is independent of t.

- i.e. the probability is only determined by the length of the interval (a, b), not by its *position* on the real line.

Continuous uniform distribution



Uniform (continuous) density and the *uniform property*. ([2])

$$\mathbb{P}(t < X < t+h) = \mathbb{P}(s < X < s+h)$$

Continuous uniform distribution

Uniform distribution

$$
\begin{aligned}
(a, b) &= \text{range of values} \\
f(x) &= \frac{1}{b-a}, \quad a < x < b \\
\mathbf{E}(X) &= \frac{a+b}{2} \\
\mathrm{Var}(X) &= \frac{(b-a)^2}{12}
\end{aligned}
$$

Properties ([2])

Python: continuous uniform distribution

```
import random

# Generate a random number between 0 and 1 from a uniform continuous distribution
random_number = random.uniform(0, 1)
print(random_number)

import numpy as np

# Generate a random number between 0 and 1 from a uniform continuous distribution
random_number = np.random.uniform(0, 1)
print(random_number)

import numpy as np
import matplotlib.pyplot as plt

# Generate 1000 random numbers from a uniform continuous distribution: this code generates

random_numbers = np.random.uniform(0, 1, 1000)

# Plot histogram
plt.hist(random_numbers, bins=20, density=True, alpha=0.7, color='blue')
plt.title('Uniform Continuous Distribution')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```

Exponential distribution: $X \sim Exp(\lambda)$

- Often used to model time: waiting time, lifetime of electronic components, failure time, service time in a queueing system, etc.
- When the (discrete) *number* of events has Poisson distribution, the (continuous) *time between events* has Exponential distribution.
- Density:

$$f(x) = \lambda e^{-\lambda x} \quad x > 0. \tag{17}$$

- λ is the *parameter* of the Exponential distribution: if e.g. X represents the time, measured in minutes, then λ represents *frequency* and it is measured in min^{-1}. The meaning of λ is the same as the parameter of the Poisson distribution.
- E.g.: if a system breaks every half a minute, on average, then $\mathbb{E}[X] = 1/2$ and $\lambda = 2$ and we say that it breaks with a frequency (breaking rate) of 2 breaks per minute.

Exponential distribution - The *Memoryless* Property

- Exponential random variables lose memory: Suppose T is an Exp variable which represents some waiting time. Then: if we wait for t minutes, the future behaviour of T is not affected i.e. the fact that we waited t minutes is "forgotten" and it doesn't influence any future waiting time:

$$\mathbb{P}\left(T > t + x | T > t\right) = \mathbb{P}(T > x) \quad t, x > 0 \tag{18}$$

where t is the elapsed time and x is the remaining time. That is, independent of the event $T > t$, when the total waiting time exceeds t, the remaining waiting time still has Exp distribution with the same parameter:

$$\begin{aligned}\mathbb{P}\left(T > t + x | T > t\right) &= \frac{\mathbb{P}\left(\{T > t + x\} \cap \{T > t\}\right)}{\mathbb{P}(T > t)} \\ &= \frac{\mathbb{P}(T > t + x)}{\mathbb{P}(T > t)} = \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x}.\end{aligned}$$

Exponential distribution - The *Memoryless* Property

- No other type of continuous random variable X has the memoryless property. The discrete Geometric distribution has it though.
- In some sense, the Geometric distribution is a discrete version of the Exponential distribution.

Exponential distribution

Exponential distribution

$$
\begin{aligned}
\lambda &= \text{frequency parameter, the number of events per time unit} \\
f(x) &= \lambda e^{-\lambda x}, \quad x > 0 \\
\mathbf{E}(X) &= \frac{1}{\lambda} \\
\mathrm{Var}(X) &= \frac{1}{\lambda^2}
\end{aligned}
$$

Properties ([2])

Python: Exponential distribution

```
import numpy as np
import matplotlib.pyplot as plt

# Generate random numbers from a continuous exponential distribution \\
# with parameter 100: This code will generate random numbers from a  continuous \\
# exponential distribution with a scale parameter of 1/100, which corresponds\\
# to a mean of 100. It then plots the generated numbers. \\
# You can adjust the size parameter in np.random.exponential() \\
# to generate more or fewer random numbers.

random_numbers = np.random.exponential(scale=1/100, size=1000)

# Plot
plt.plot(random_numbers, color='green')
plt.title('Continuous Exponential Distribution')
plt.xlabel('Index')
plt.ylabel('Value')
plt.grid(True)
plt.show()
```

Gamma distribution $\sim \Gamma(\alpha, \lambda)$

- When a specific process takes place in α independent steps, and each step takes $Exp(\lambda)$, then the total time has Gamma distribution with parameters α and λ.
- Density

$$f(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0 \tag{19}$$

where Γ is the Gamma function

$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx, \quad t > 0$$

with

$$\Gamma(t+1) = t\Gamma(t) \quad \text{for any} \quad t > 0$$
$$\Gamma(t+1) = t! = 1 \cdot 2 \cdot \ldots \cdot t, \quad \text{for any integer} \quad t.$$

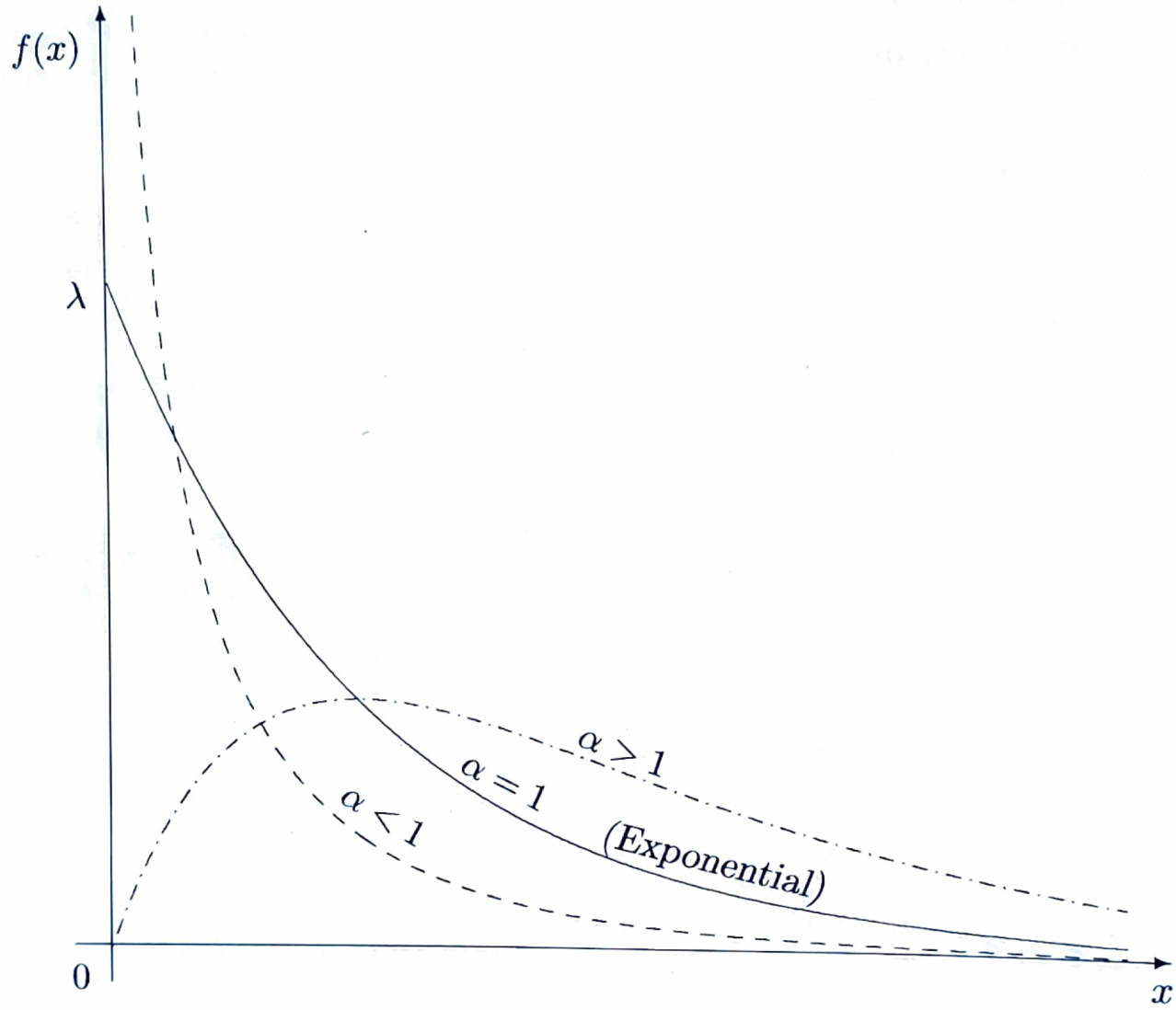
Gamma distribution $\sim \Gamma(\alpha, \lambda)$

- Gamma distribution can be used for modelling a sequential/multi-step process e.g. wait time in a queue, repair time of machines (especially relevant in maintenance operations where the repair time can vary due to factors such as complexity of the repair or availability of spare parts), duration of rainfall events (useful in hydrology and climate studies for predicting the occurrence and duration of rainfall events for purposes such as flood risk assessment and water resource management), etc.

- In a scenario with rare events the time between any two consecutive events is modelled using an Exp distribution, the time of the α^{th} event has Γ distribution, as it consists of α independent Exponential times.

- Different values for α generate different shapes for the Γ distribution $\Rightarrow \alpha$ is called shape parameter.

Gamma distribution



Gamma (continuous) densities with different shape parameters α. ([2])

Gamma distribution

Gamma distribution

$$
\begin{aligned}
\alpha &= \text{shape parameter} \\
\lambda &= \text{frequency parameter} \\
f(x) &= \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0 \\
\mathbf{E}(X) &= \frac{\alpha}{\lambda} \\
\mathrm{Var}(X) &= \frac{\alpha}{\lambda^2}
\end{aligned}
$$

Properties ([2])

Gamma distribution $\sim \Gamma(\alpha, \lambda)$

- Special cases:
  - When $\alpha = 1 \Rightarrow$ we obtain $Exp(\lambda)$ i.e. Gamma distribution becomes Exponential:
$$\Gamma(1, \lambda) = Exp(\lambda). \tag{20}$$
  - When $\lambda = \frac{1}{2}$ and for any $\alpha > 0$ we obtain the *chi-square distribution with* 2α *degress of freedom* which we will discuss next year.
$$\Gamma(\alpha, 1/2) = Chi - square(2\alpha). \tag{21}$$

Python: Gamma distribution

```
import numpy as np
import matplotlib.pyplot as plt

# Generate random numbers from a continuous gamma distribution \\
# with shape parameter k=2 and scale parameter theta=2

random_numbers = np.random.gamma(shape=2, scale=2, size=1000)

# Plot
plt.plot(random_numbers, color='blue')
plt.title('Continuous Gamma Distribution')
plt.xlabel('Index')
plt.ylabel('Value')
plt.grid(True)
plt.show()
```

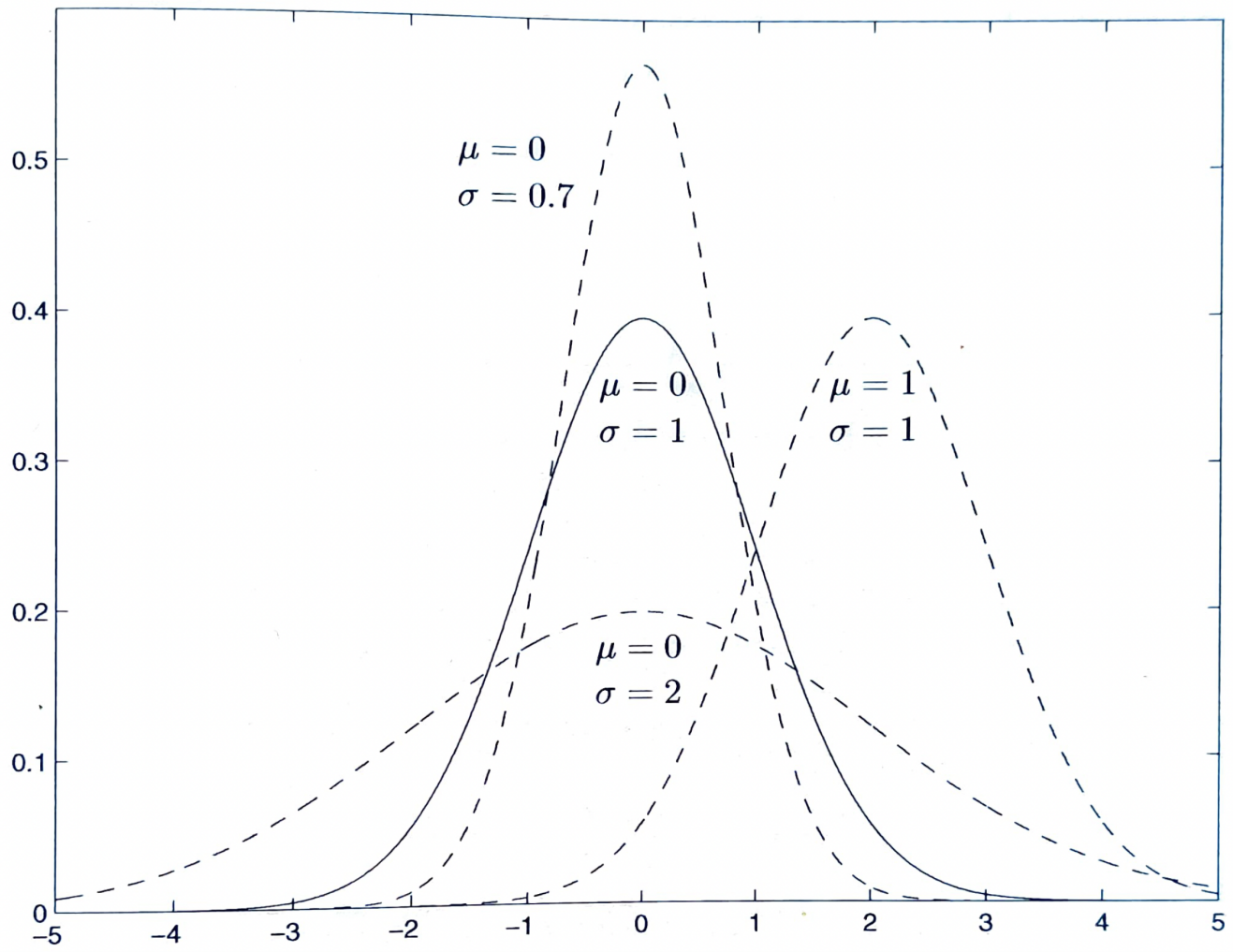
Normal distribution: $X \sim \mathcal{N}(\mu, \sigma^2)$

- Essential in Probability and Statistics - for many reasons.
  - Central Limit Theorem (CLT): the distribution of the sum (or average) of a large number of independent and identically distributed random variables approaches/has approximately a normal distribution, regardless of the underlying distribution of the individual variables.
  - Good model for physical quantities such as height, weight, temperature, pollution level, blood pressure, as well as errors and noise in various processes.
- Density

$$f(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \quad \text{for} \quad x \in (-\infty, \infty) \qquad (22)$$

  where $\mu = \mathbb{E}[X]$ and $\sigma^2 = V(X)$ as before.
- The normal density has a bell-shaped curve, symmetric and centered at μ. Its spread is controlled by σ. If we change μ, we shift the curve to the left/right, while if we change σ we make it more flat/concentrated.

Normal distribution: $X \sim \mathcal{N}(\mu, \sigma^2)$



Normal (continuous) densities with different scale parameters and locations. ([2])

Normal distribution

Normal distribution

$$
\begin{aligned}
\mu &= \text{expectation, location parameter} \\
\sigma &= \text{standard deviation, scale parameter} \\
f(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty \\
\mathbf{E}(X) &= \mu \\
\mathrm{Var}(X) &= \sigma^2
\end{aligned}
$$

Properties ([2])

Python: Normal distribution

```
import numpy as np
import matplotlib.pyplot as plt

# Generate random numbers from a continuous normal distribution\\
# with mean 0 and standard deviation 1
random_numbers = np.random.normal(loc=0, scale=1, size=1000)

# Plot
plt.plot(random_numbers, color='red')
plt.title('Continuous Normal Distribution')
plt.xlabel('Index')
plt.ylabel('Value')
plt.grid(True)
plt.show()
```

Moment generating function

Definition 26

The moment generating function of a random variable X is the function $M_X : D \subseteq \mathbb{R} \to \mathbb{R}$ defined by

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] \text{ for } t \in D$$

where D is the set of all $t \in \mathbb{R}$ for which the above expectation exists.

Moment generating function

Theorem 27

Let X and Y be random variables. Then the following properties hold:

1. $M_X(0) = 1$.
2. *If $Y = aX + b$ for fixed $a, b \in \mathbb{R}$, then $M_Y(t) = e^{tb} M_X(at)$ for all $b \in \mathbb{R}$ and $t, a \in \mathbb{R}$ for which $M_X(at)$ is defined.*

Theorem 28

If there exists $\delta > 0$ such that M_X is defined on $(-\delta, \delta)$, then for $t \in (-\delta, \delta)$

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}[X^k]$$

and for $n \in \mathbb{N}$

$$\mathbb{E}[X^n] = M_X^{(n)}(t)|_{t=0}$$

where $M_X^{(n)}(t)|_{t=0}$ is the derivative of order n of M_X at 0.

Moment generating function

Theorem 29

Let X and Y be random variables for which there exists $\delta > 0$ such that $M_X(t) = M_Y(t)$ for each $t \in (-\delta, \delta)$. Then $F_X = F_Y$, i.e. X and Y have the same distribution.

Moment generating function

Remark: Why is the moment generating function useful? There are basically two reasons for this

1. The moment generating function of X gives us all moments of X, see Theorem 28. That is why it is called the *moment generating function*.
2. The moment generating function uniquely determines the distribution (see Theorem 29): *There is an one-to-one correspondence between the probability distribution of a random variable and the moment generating function*, if this function is defined on an (open) interval containing 0. That is, the distribution of a random variable is uniquely determined by its moment generating function and the moment generating function of a random variable is (by definition) uniquely determined by the distribution of the random variable.

Moment generating function

Example: If the random variable X has a binomial distribution $\overline{Bino(n,p)}$, $n \in \mathbb{N}^*$, $p \in (0,1)$, then its moment generating function is given for each $t \in \mathbb{R}$ by

$$M_X(t) = \sum_{k=0}^{n} e^{tk} C_n^k p^k (1-p)^{n-k} = (e^t p + 1 - p)^n.$$

Moment generating function

Remark: The **characteristic function** of a random variable X is the function $\varphi_X : \mathbb{R} \to \mathbb{C}$ defined by

$$\varphi_X(t) = \mathbb{E}\Big[\exp\{itX\}\Big] = \mathbb{E}\big[\cos(tX)\big) + i\mathbb{E}\big[\sin(tX)\big] \text{ for } t \in \mathbb{R}.$$

Note that in comparison to the moment generating function, this is a complex valued function and it is defined for all $t \in \mathbb{R}$. There is a one-to-one correspondence between the distribution of a random variable and the characteristic function, i.e. the distribution of a random variable is uniquely determined by its characteristic function and the characteristic function of a random variable is (by definition) uniquely determined by the distribution of the random variable.

- The MGF is defined in terms of moments and exists for a range of values of t, while the characteristic function is always well-defined but may not exist for all distributions. However, both functions uniquely determine the probability distribution of a random variable if they exist.

Probabilistic inequalities

Theorem 30

Let X be a random variable with the expectation $\mathbb{E}[X]$ and variance $V(X)$. Assume $a > 0$. Then the following inequalities are true:

(1) ***Markov's inequality***

$$\mathbb{P}\Big(|X| \geq a\Big) \leq \frac{1}{a}\mathbb{E}\left[|X|\right]. \tag{23}$$

(2) ***Chebyshev's inequality***

$$\mathbb{P}\Big(|X - \mathbb{E}[X]| \geq a\Big) \leq \frac{1}{a^2}V(X). \tag{24}$$

LECTURES 8 and 9

Outline

- Sequences of Random Variables
- Weak Law of Large Numbers
- Strong Law of Large Numbers

Sequences of Random Variables

Definition 31

$(X_n)_{n\geq 1}$ *is a sequence of independent random variables, if* $\forall\ \{i_1, \ldots, i_k\} \subset \mathbb{N}$ *the random variables* $X_{i_1}, \ldots, X_{i_k}$ *are independent, i.e.*

$$\mathbb{P}(X_{i_1} \leq x_{i_1}, \ldots, X_{i_k} \leq x_{i_k}) = \mathbb{P}(X_{i_1} \leq x_{i_1}) \cdot \ldots \cdot \mathbb{P}(X_{i_k} \leq x_{i_k})$$

$\forall\ x_{i_1}, \ldots, x_{i_k} \in \mathbb{R}$.

$(X_n)_{n\geq 1}$ *is a sequence of pairwise independent random variables, if* $\forall\ i, j \in \mathbb{N}^*, i \neq j$ *the two random variables* X_i *and* X_j *are independent.*

Sequences of Random Variables

Definition 10

A sequence $(X_n)_{n\geq 1}$ of random variables converges almost surely (or almost everywhere, or with probability 1) to a random variable X if

$$\mathbb{P}\Big(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\Big) = 1.$$

This convergence is denoted by $X_n \overset{\text{a.s.}}{\to} X$.

Remark: The almost sure convergence $X_n \overset{\text{a.s.}}{\to} X$ requires that $\big(X_n(\omega)\big)_n$ converges pointwise to $X(\omega)$ for each $\omega \in \Omega$, *except* a "small event" with zero probability. If $X_n \overset{a.s.}{\to} X$ then the event

$$M = \{\omega \in \Omega : (X_n(\omega))_n \ \textit{does not converge to} \ X(\omega)\}$$

has probability $\mathbb{P}(M) = 0$.

Sequences of Random Variables

Example: On the sample space $(\Omega, \mathcal{F}, \mathbb{P})$ let $A \in \mathcal{F}$ such that $\mathbb{P}(A) = 0.4$ and $\mathbb{P}(\bar{A}) = 0.6$:

$$X_n(\omega) = \begin{cases} 1 + \frac{1}{n}, & \text{for } \omega \in A \\ -\frac{1}{n}, & \text{for } \omega \in \bar{A}. \end{cases} \implies \mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = ???\}) = 1.$$

We define

$$X(\omega) = \begin{cases} 1, & \text{for } \omega \in A \\ 0, & \text{for } \omega \in \bar{A}. \end{cases}$$

$$\implies \mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = \mathbb{P}(A) + \mathbb{P}(\bar{A}) = 1.$$

We obtain $X_n \overset{\text{a.s.}}{\to} X$.

Sequences of Random Variables

Definition 32

A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables converges in probability to a random variable X, denoted by $X_n \xrightarrow{\mathbb{P}} X$, if

$$\lim_{n\to\infty} \mathbb{P}\Big(|X_n - X| \leq \varepsilon\Big) = 1 \quad \textit{for every} \quad \varepsilon > 0\,.$$

Remark: It holds that

$$X_n \xrightarrow{\mathbb{P}} X \iff \lim_{n\to\infty} \mathbb{P}\Big(|X_n - X| > \varepsilon\Big) = 0 \quad \text{for every} \quad \varepsilon > 0.$$

Sequences of Random Variables

Example: Let $(X_n)_{n\geq 1}$ be a sequence of random variables such that for each $n \geq 1$ we have $\mathbb{P}(X_n = 1) = p_n$ and $\mathbb{P}(X_n = 0) = 1 - p_n$, where $p_n \in [0, 1]$. We prove the following equivalences:

$$X_n \xrightarrow{\mathbb{P}} 0 \iff \lim_{n\to\infty} p_n = 0.$$

We have

$$X_n \xrightarrow{\mathbb{P}} 0 \iff \lim_{n\to\infty} \mathbb{P}(|X_n| > \epsilon) = 0 \text{ for all } \epsilon > 0.$$

But,

$$\mathbb{P}(|X_n| > \epsilon) = \begin{cases} \mathbb{P}(X_n = 1) = p_n & \text{if } \epsilon < 1 \\ \mathbb{P}(\emptyset) = 0 & \text{if } \epsilon \geq 1. \end{cases}$$

Hence, $X_n \xrightarrow{\mathbb{P}} 0 \iff \lim_{n\to\infty} p_n = 0.$

Sequences of Random Variables

Definition 33

A sequence $(X_n)_{n\geq 1}$ of random variables converges in mean square to a random variable X if

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^2] = 0.$$

This convergence is denoted by $X_n \overset{L^2}{\to} X$.

Sequences of Random Variables

Example: Let $(X_n)_{n\geq 1}$ be a sequence of pairwise independent random variables such that for each $n \geq 1$ we have $\mathbb{E}[X_n] = m$ and $V(X_n) = \sigma^2$, where $m \in \mathbb{R}$ and $\sigma > 0$. For the sequence of random variables $(Y_n)_{n\geq 1}$ defined by $Y_n := \frac{1}{n}(X_1 + ... + X_n)$ we prove that $Y_n \xrightarrow{L^2} m$.

Proof: We use the additivity of the expectation and the independence property of the random variables $X_1 - m, X_2 - m, ..., X_n - m$ to write

$$\mathbb{E}[(Y_n - m)^2] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - m)\right)^2\right]$$

$$= \frac{1}{n^2}\Big(\sum_{i=1}^{n}\mathbb{E}[(X_i - m)^2] + 2\sum_{1\le i<j\le n}\mathbb{E}\big[(X_i - m)(X_j - m)\big]\Big)$$

$$= \frac{1}{n^2}\Big(\sum_{i=1}^{n}\mathbb{E}[(X_i - m)^2]\Big) = \frac{1}{n^2}\Big(V(X_1) + V(X_2) + ... + V(X_n)\Big)$$

$$= \frac{\sigma^2}{n} \to 0, \text{ as } n \to \infty.$$

Hence $Y_n \overset{L^2}{\to} m$.

Sequences of Random Variables

Definition 34

A sequence $(X_n)_{n\geq 1}$ of random variables converges in distribution to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

in each continuity point x of F_X. This convergence is denoted by $X_n \xrightarrow{d} X$.

Sequences of Random Variables

Example: Consider a sequence $(X_n)_{n\geq 1}$ of random variables such that each $X_n \sim Unif[-n, n]$. Then (exercise!) the distribution function F_{X_n} of X_n is given by

$$F_{X_n}(x) = \begin{cases} 0 & \text{if } x < -n \\ \dfrac{x+n}{2n} & \text{if } -n \leq x < n \\ 1 & \text{if } x \geq n. \end{cases}$$

We see that

$$\lim_{n\to\infty} F_{X_n}(x) = \frac{1}{2} \quad \text{for all } x \in \mathbb{R}.$$

But the limiting function is *not* a cumulative distribution function. Thus, a sequence of distribution functions can converge at all points but the limiting function may not be a distribution function. We conclude that $(X_n)_{n\geq 1}$ *does not converge in distribution*.

Sequences of Random Variables

Theorem 35

Let $(X_n)_{n\geq 1}$ be a sequence of random variables, and let X be a random variable. Then the following implications are true:

(1) $X_n \overset{a.s.}{\to} X \Longrightarrow X_n \overset{\mathbb{P}}{\to} X.$

(2) $X_n \overset{L^2}{\to} X \Longrightarrow X_n \overset{\mathbb{P}}{\to} X.$

(3) $X_n \overset{\mathbb{P}}{\to} X \Longrightarrow X_n \overset{d}{\to} X.$

Law of Large Numbers

- The Law of Large Numbers (LLN) states that if the same experiment/study is repeated independently a large number of times, the average of the results of the trials must be close to the expected value. The result becomes closer to the expected value as the number of trials is increased.

- It provides a theoretical basis for making predictions and inferences based on large samples. This allows us to make more accurate predictions and inferences about a population, based on the sample data.

- Insurance companies use the law of large numbers to estimate the losses a certain group of insured persons may have in the future and to predict the risks. The law of large numbers states that as the number of policyholders increases, the more confident the insurance company is that its prediction will prove true. Therefore, they attempt to acquire a large number of similar policyholders who all contribute to a fund which will pay the losses.

Law of Large Numbers

Definition 11

A sequence $(X_n)_{n\geq 1}$ of random variables such that $\mathbb{E}[|X_n|] < \infty$ for all $n \in \mathbb{N}$ obeys the weak law of large numbers (**WLLN**) if

$$\frac{1}{n}\sum_{k=1}^{n}\Big(X_k - \mathbb{E}[X_k]\Big) \xrightarrow{\mathbb{P}} 0.$$

Law of Large Numbers

Theorem 36

Let $(X_n)_{n\geq 1}$ be a sequence of pairwise independent random variables satisfying the condition

$$V(X_n) \leq L, \quad \textit{for all} \quad n \in \mathbb{N}^*,$$

where $L > 0$ is a constant. Then $(X_n)_{n\geq 1}$ obeys the WLLN.

Law of Large Numbers

Proof.

Let

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) \quad \text{for each } n \in \mathbb{N}.$$

By the Chebyshev inequality we have for each $\epsilon > 0$

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{k=1}^{n}\Big(X_k - \mathbb{E}[X_k]\Big)\Big| \geq \epsilon\Big) = \mathbb{P}\Big(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon\Big)$$

$$\leq \frac{1}{\epsilon^2}V(\overline{X}_n) = \frac{1}{\epsilon^2 n^2}V(X_1 + \cdots + X_n) = \frac{1}{\epsilon^2 n^2}\left(V(X_1) + \cdots + V(X_n)\right) \leq \frac{L}{n},$$

since $X_1, \ldots, X_n$ are pairwise independent random variables. Letting $n \to \infty$, we conclude that $(X_n)_{n\geq 1}$ obeys the WLLN. □

Law of Large Numbers

Example: Let $(X_n)_{n\geq 1}$ be a sequence of pairwise independent random variables, where for every $n \geq 1$ we have

$$\mathbb{P}(X_n = 0) = 1 - p_n, \quad \mathbb{P}(X_n = 1) = p_n \text{ with } p_n \in (0, 1).$$

Then $V(X_n) = p_n(1 - p_n) \leq \frac{1}{4}$, $\forall n \geq 1$, and $(X_n)_{n\geq 1}$ obeys the WLLN.

Law of Large Numbers

Definition 37

A sequence $(X_n)_{n\geq 1}$ *of random variables such that* $\mathbb{E}|X_n| < \infty$ *for all* $n \in \mathbb{N}$ *obeys the strong law of large numbers* (***SLLN***) *if*

$$\frac{1}{n}\sum_{k=1}^{n}\Big(X_k - \mathbb{E}[X_k]\Big) \overset{a.s.}{\to} 0.$$

Remark: If $(X_n)_{n\geq 1}$ satisfies the SLLN, then by Theorem 35 $(X_n)_{n\geq 1}$ satisfies also the WLLN (a.s. convergence implies convergence in probability).

Law of Large Numbers

Theorem 38

If $(X_n)_{n\geq 1}$ is a sequence of independent random variables such that $\sum_{n=1}^{\infty} \frac{1}{n^2} V(X_n) < \infty$, then

$$\frac{1}{n} \sum_{k=1}^{n} \Big(X_k - \mathbb{E}[X_k] \Big) \overset{a.s.}{\to} 0,$$

i.e. $(X_n)_{n\in\mathbb{N}}$ obeys the SLLN.

Law of Large Numbers

Theorem 39

Let $(X_n)_{n\geq 1}$ be a sequence of independent identically distributed random variables such that $E[X_n] = m$ for all $n \in \mathbb{N}$. Then

$$\frac{1}{n}\sum_{k=1}^{n} X_k \overset{a.s.}{\to} m$$

i.e. $(X_n)_{n\in\mathbb{N}}$ obeys the SLLN.

Law of Large Numbers

Example: Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables, where $X_n \sim Bernoulli(p)$, i.e.

$$\mathbb{P}(X_n = 0) = 1 - p, \mathbb{P}(X_n = 1) = p, \text{ with } p \in [0, 1].$$

Then $(X_n)_{n\in\mathbb{N}}$ obeys the SLLN, i.e. $\frac{1}{n}(X_1 + ... + X_n) \overset{a.s.}{\to} p$.

Law of Large Numbers

Theorem 40

Consider the event A which appears within a certain experiment. SLLN: After repeating the experiment independently n times (n sufficiently large), under the same conditions, the relative frequency $h_n(A)$ of the event A approximates the theoretical probability $\mathbb{P}(A)$:

$$h_n(A) \xrightarrow{a.s.} \mathbb{P}(A) \text{ if } n \to \infty.$$

Remark: In simulations: $h_n(A) \approx \mathbb{P}(A)$, if n is sufficiently large.

Definition 41

Let A be a random event appearing in an experiment; the experiment is repeated n times (under the same given conditions) and denote by k_n how many times the event A appears; the relative frequency of the event A is the number

$$h_n(A) = \frac{k_n}{n}$$

and k_n is the absolute frequency of the event A.

- After repeating an experiment n times (n sufficiently large), under the same conditions, the relative frequency $h_n(A)$ of the event A is approximately equal to the probability $\mathbb{P}(A)$

$$h_n(A) \approx \mathbb{P}(A), \text{ if } n \to \infty.$$

$\Rightarrow$ In the long run, the probability of an event can be viewed as a *proportion* of times this event happens i.e. its relative frequency.

Law of Large Numbers

Proof.

We follow the ideas from Example 208 for the sequence of independent random variables $(X_n)_{n\geq 1}$, where

$$X_n = \begin{cases} 1, & \text{if } A \text{ appears in the } n\text{-th repeating of the experiment} \\ 0, & \text{if } \bar{A} \text{ appears in the } n\text{-th repeating of the experiment} \end{cases}$$

$$\Longrightarrow X_n \sim \begin{pmatrix} 0 & 1 \\ 1-\mathbb{P}(A) & \mathbb{P}(A) \end{pmatrix} \Rightarrow X_n \sim Bernoulli\big(\mathbb{P}(A)\big)$$

$$\Rightarrow \mathbb{E}[X_n] = 0 \cdot (1-\mathbb{P}(A)) + 1 \cdot \mathbb{P}(A) = \mathbb{P}(A) \ \forall n \in \mathbb{N}^*.$$

By Theorem 207 this implies that $\frac{1}{n}(X_1 + ... + X_n) \xrightarrow{a.s.} \mathbb{P}(A)$.
But $\frac{1}{n}(X_1 + ... + X_n) = h_n(A)$ (the relative frequency of the event A)
$\Longrightarrow h_n(A) \xrightarrow{a.s.} \mathbb{P}(A)$. □

LECTURE 10

Outline

- Weak and Strong Law of Large Numbers - Recap
- Central Limit Theorem

Sums of random variables

- Sums of random variables of the form

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

  appear in many applications.

- What's the behaviour of S_n for large n?

Sums of random variables

Let's first look at the following exercise:

Exercise: The function $numpy.random.rand$ returns in Python, for each independent call, a random value from the interval $[0, 1]$, according to the uniform distribution $Unif[0, 1]$. Let $n \in \mathbb{N}^*$ and $p \in (0, 1)$. If the function is called n times, what is the expectation and variance of the number of values less than p?

Code here.

Sums of random variables

Answer:
Let's call a *success* the event that a value returned by $numpy.random.rand$ is less than p. Let X be number of values less than p. Since a value generated by $numpy.random.rand$ is less than p with probability $\int_0^p 1dx = p$, we have $X \sim Bino(n, p)$.

We will show that $\mathbb{E}[X] = np$ and $V(X) = np(1-p)$.

For $i \in \{1, \ldots, n\}$ let $X_i \sim Bernoulli(p)$ be such that $X_i = 1$, if the ith trial is a success, and $X_i = 0$, if the ith trial is not a success (in particular, $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = 0) = 1 - p$). Then $X = X_1 + \ldots + X_n \sim Bino(n, p)$, so

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = p + ... + p = np.$$

Since the random variables $X_1, \ldots, X_n$ are independent,

$$\begin{aligned} V(X) &= V(X_1 + \cdots + X_n) = V(X_1) + \cdots + V(X_n) \\ &= np(1-p) = np(1-p). \end{aligned}$$

Random walk

- A *random walk* is often used to model the behaviour of systems where each step is unpredictable, such as the movement of particles in a fluid or the fluctuations of stock prices over time.
- A simple random walk in one dimension can be defined as follows:

$$S_n = S_{n-1} + X_n$$

where:

  - S_n is the position of the walker after n steps.
  - S_{n-1} is the position of the walker after $n-1$ steps.
  - X_n is a random variable representing the step taken at step n.

- For a simple random walk in one dimension with step size $+1$ or -1 with probabilities p and $1-p$ respectively:
- Expected Value:

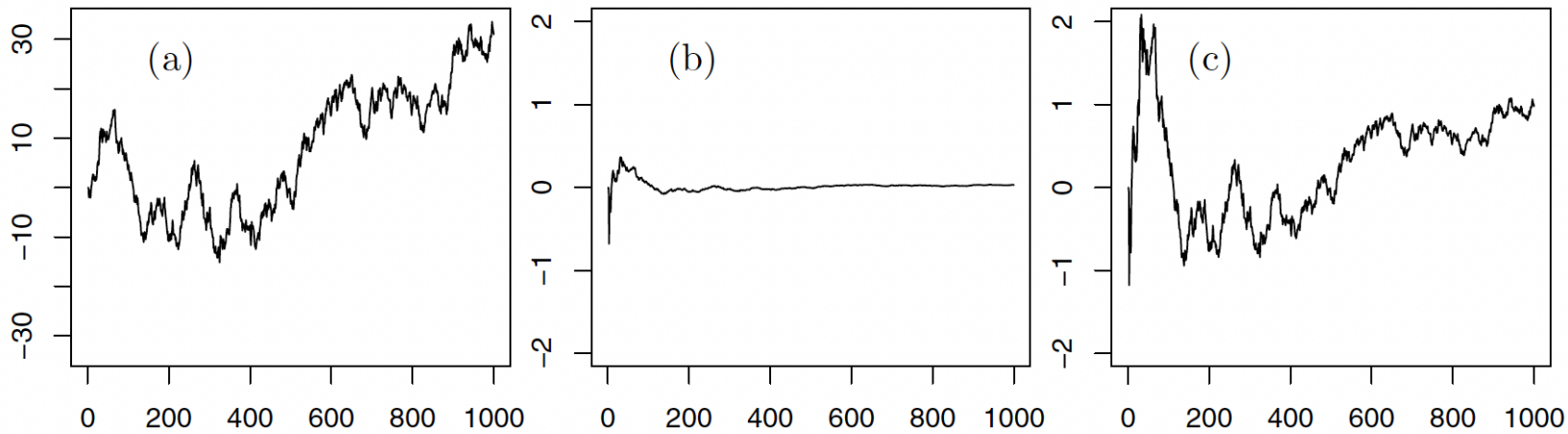
$$\mathbb{E}[S_n] = np$$

- Variance:

$$\text{Var}(S_n) = np(1-p)$$

.

Central Limit Theorem



(a) Random walk S_n, (b) S_n/n, (c) $S_n/\sqrt{n}$ ([2]).

Central Limit Theorem

- The sum S_n diverges, which is not surprising, since

$$V(S_n) = n\sigma^2 \to \infty \quad \text{when} \quad n \to \infty.$$

  That is, the variability of S_n is not bounded when $n \to \infty$.

- The average S_n/n converges:

$$V\left(\frac{S_n}{n}\right) = \frac{1}{n^2}V(S_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \to 0.$$

- However, $S_n/\sqrt{n}$ neither converges nor diverges: it doesn't converge to 0, but it doesn't entirely leave 0 either $\to$ it behaves like a random variable. For large n, this random variable behaves like a normal distribution $\Rightarrow$ The Central Limit Theorem.

Central Limit Theorem

Theorem 42

[CLT; Lindeberg-Lévy Theorem] *Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed random variables such that $\mu = \mathbb{E}[X_n]$ and $\sigma^2 = V(X_n) > 0 \ \forall \ n \geq 1$. Then*

$$\frac{(X_1 + ... + X_n) - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z, \quad \text{where } Z \sim N(0,1),$$

this means

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{(X_1 + ... + X_n) - n\mu}{\sigma\sqrt{n}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = F_{N(0,1)}(x)$$

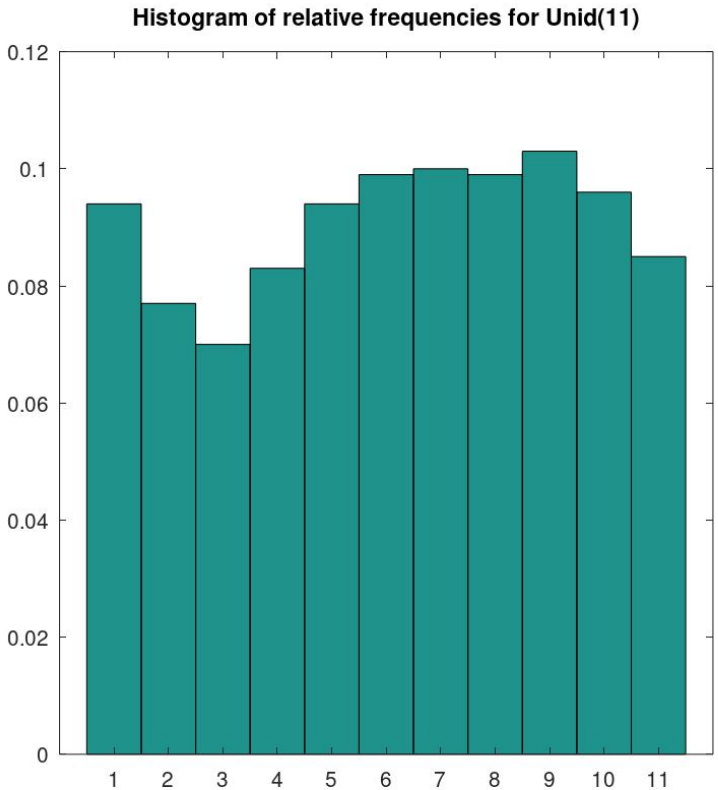
$\forall x \in \mathbb{R}$ *where $F_{N(0,1)}$ denotes the cumulative distribution function of the standard normal distribution $N(0,1)$.*

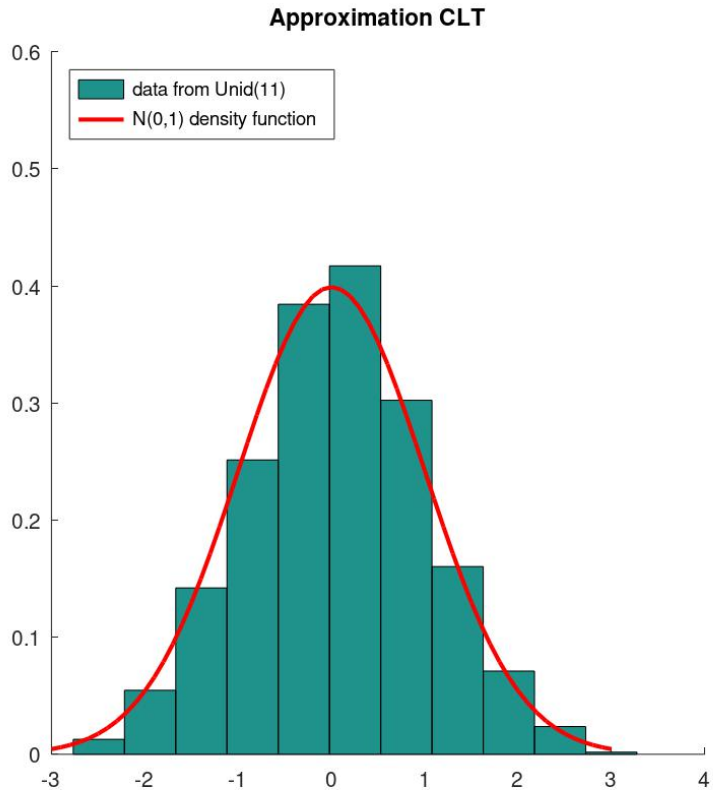
Central Limit Theorem

- The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the variables' distribution

  $\Rightarrow$ it can be applied to random variables $X_1, X_2, \ldots, X_n$ with **any distribution type**: as long as n is large enough, we can use the Normal distribution to compute probabilities for S_n.

Central Limit Theorem





$n = 400, X_1, ..., X_{400} \sim Unid(11)$, $\mu = 6, \sigma^2 = 10$, and the CLT (Theorem 42)

Central Limit Theorem

- The histogram of a random variable is a graphical representation of the distribution of values that the random variable can take. It shows how frequently different values occur within a given range.
- We can create a histogram in Python using $matplotlib.pyplot.hist()$.

Central Limit Theorem

- The value $F_{N(0,1)}(x)$ can be computed in Python using $norm.cdf(x, 0, 1)$ (from scipy.stats import norm).
- $\frac{(X_1+...+X_n)-n\mu}{\sigma\sqrt{n}} \sim N(0,1)$ for n sufficiently large (usually $n > 30$)
- $\forall\ a, b \in \mathbb{R}, a < b$ it holds that

$$\mathbb{P}\Big(a < \frac{(X_1 + ... + X_n) - n\mu}{\sigma\sqrt{n}} < b\Big) \approx F_{N(0,1)}(b) - F_{N(0,1)}(a).$$

for n sufficiently large ($n > 30$).

Central Limit Theorem

Example: Let $(X_n)_{1 \leq n \leq 100}$ be independent random variables following the $Bernoulli(0.5)$ distribution. Using the CLT (Theorem 42) estimate the probability $\mathbb{P}(35 < X_1 + ... + X_{100} < 65)$.

Central Limit Theorem

Answer: We have $\mathbb{E}[X_n] = 0.5$, $\sigma = \sqrt{V(X_n)} = 0.5$ and

$$\mathbb{P}(35 < X_1 + ... + X_{100} < 65) = \mathbb{P}\left(-3 < \frac{(X_1 + ... + X_{100}) - 50}{0.5 \cdot 10} < 3\right)$$
$$= F_{N(0,1)}(3) - F_{N(0,1)}(-3) \approx 0.9973.$$

$$\Longrightarrow \mathbb{P}(35 < X_1 + ... + X_{100} < 65) \approx 0.9973.$$

Central Limit Theorem

- Binomial variables are a special case of $S_n = X_1 + X_2 + \ldots + X_n$ where each $X_i \sim Bernoulli(p)$. We have the following theorem:

Theorem 43 (Moivre-Laplace)

We consider a sequence of independent Bernoulli trials, such that the probability that an event A occurs in each trial of the experiment is $p \in (0,1)$. Denote by A_j the event that A occurred in the j-th trial and let $X_j = \mathbb{1}_{A_j}$ for all $j \in \mathbb{N}$. Then $S_n = X_1 + \cdots + X_n \sim Bino(n,p)$ and S_n represents the number of occurrences of A in n trials. Moreover,

$$\lim_{n\to\infty} \mathbb{P}\left(a < \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{t^2}{2}}\, dt \tag{25}$$

where $a, b \in \mathbb{R}, a < b$.

Central Limit Theorem

Proof.

We have $m_n = \mathbb{E}(X_n) = p$ and $\sigma^2 = V(X_n) = pq$ for each $n \in \mathbb{N}$ and $(X_n)_{n\in\mathbb{N}}$ is a sequence of independent random variables (because $(A_n)_{n\geq 1}$ is a sequence of independent events).
Each random variable X_j ($j \in \mathbb{N}$) can take the value 0 or 1. We have $X_j = 1$ if the event A occurred in the jth trial and $X_j = 0$ if A did not occur in the jth trial.
Then $S_n = X_1 + \cdots + X_n$ shows how often A occurred in n trials (it is the absolute frequency of A: $k_n(A)$) and it has binomial distribution with parameters n and p: $Bino(n, p)$. Theorem 42 implies that (25) is true. □

Central Limit Theorem

Remark:

- By Theorem 43 we deduce that the probability that S_n (the number of occurrences of A in n trials) belongs to the interval (a, b) is approximated as follows

$$\mathbb{P}(a < S_n < b) \approx F_{N(0,1)}\left(\frac{b-np}{\sqrt{npq}}\right) - F_{N(0,1)}\left(\frac{a-np}{\sqrt{npq}}\right).$$

- In a sequence of independent trials of an experiment the probability that an event A occurs in each trial is p and k_n denotes the number of occurrences of the event A in n trials (absolute frequency).
- How well the relative frequency $h_n(A) = \dfrac{k_n}{n}$ of A is approximated by the probability p is expressed in

$$\mathbb{P}\left(|h_n(A) - p| < \epsilon\right) \approx F_{N(0,1)}\left(\frac{\epsilon n}{\sqrt{np(1-p)}}\right) - F_{N(0,1)}\left(\frac{-\epsilon n}{\sqrt{np(1-p)}}\right).$$

Central Limit Theorem

Theorem 44

Let $(X_n)_n$ be a sequence of random variables such that $X_n \sim Bino(n, p_n)$ such that $p_n = \frac{\lambda}{n}$ $(\lambda > 0)$.

Then $X_n \xrightarrow{d} X$, where $X \sim Poiss(\lambda)$.

Central Limit Theorem

Proof: Recall

$$X_n \sim Bino(n, p_n) \Longrightarrow \mathbb{P}(X_n = k) = C_n^k p_n^k (1-p_n)^{n-k}, k \in \{0, 1, 2, ..., n\}$$

and

$$X \sim Poiss(\lambda) \Longrightarrow \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} k \in \{0, 1, 2, ...\}.$$

We compute for $k \in \{0, 1, 2, ...\}$

$$\lim_{n\to\infty} C_n^k p_n^k (1-p_n)^{n-k} = \lim_{n\to\infty} \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n-k+1}{n} \cdot \ldots \cdot \frac{n-1}{n} \cdot \frac{n}{n} \cdot \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k} = \frac{\lambda^k}{k!} e^{-\lambda} = \mathbb{P}(X = k).$$

Central Limit Theorem

For $x \in \mathbb{R}$,

$$F_{X_n}(x) = \mathbb{P}(X_n \leq x) = \begin{cases} 0, \text{ if } x < 0 \\ \sum\limits_{0 \leq k \leq x} \mathbb{P}(X_n = k) \text{ if } \ x \geq 0, \end{cases}$$

and

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0, \text{ if } x < 0 \\ \sum\limits_{0 \leq k \leq x} \mathbb{P}(X = k) \text{ if } x \geq 0. \end{cases}$$

Hence,

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

in each continuity point x of F_X. Thus, we have obtained $X_n \xrightarrow{d} X$.

Central Limit Theorem

Let $k \in \mathbb{N}$ with $k < 500$. What is the probability that in a group with 500 persons k persons have birthday on January 1 (assume that a year has 365 days)?
Denote by X the random variable that shows the number of persons having birthday on January 1 and we want to find $\mathbb{P}(X = k)$. We can find this probability in two ways:

Central Limit Theorem

1. Using binomial distribution: among $n = 500$ trials to obtain k times "success" ("success" means that a person has birthday on January 1), where $\mathbb{P}(\text{"}success\text{"}) = p = \frac{1}{365}$. Then $\mathbb{P}(X = k) = C_{500}^{k} p^{k} (1-p)^{n-k}$.

2. Using Poisson approximation (see Theorem 44): we have $\lambda = np = \frac{500}{365} \approx 1.3698$. Then $\mathbb{P}(X = k) = \frac{\lambda^{k}}{k!} e^{-\lambda}$ and these probabilities can be calculated in e.g. Python.

LECTURES 11 and 12

Outline

- Computer Simulations (theory)
- Monte Carlo Methods
- Python Simulations: use this link.

References:
1. M. Baron, *Probability and Statistics for Computer Scientists* (2019).
2. S. Chongchitnan, *Exploring University Mathematics with Python* (2023).

Computer Simulations

- Computer simulations refer to a regeneration of a process by writing a suitable computer program and observing its results.
- Monte Carlo methods are based on computer simulations involving random numbers $\rightarrow$ used for computing probabilities, expected values, and other distribution characteristics.
- The main purpose of simulations is to estimate quantities for which the direct computation is complicated, risky, consuming, expensive, or impossible.
- Examples: evaluating the performance and associated risks of complex devices or machines before they are built and launched (remember the first example in "Lectures 4 & 5".)

Computer Simulations

- Remember that probability can be defined as a long-run proportion:

$$h_n(A) \approx \mathbb{P}(A), \text{ if } n \to \infty$$

where

$$h_n(A) = \frac{k_n}{n}$$

and k_n tells us how many times the event A appears in an experiment which is repeated n times.

- Computers can simulate a long run using *random number generators*, allowing us to estimate probabilities based on observed frequencies.
- Applications in: forecasting, percolation, queueing, Markov Chain Monte Carlo (MCMC).

Computer Simulations

Monte Carlo Methods

- *Monte Carlo Methods* inherit their name from Europe's most famous Monte Carlo casino. Mathematicians used Monte Carlo methods to estimate probabilities and to devise optimal *gambling strategies*.

  $\Rightarrow$ algorithms are designed to generate random variables and vectors with a desired distribution: random variables with uniform distribution are typically generated first, and then these are transformed into random variables with the original requested distribution.

- Python Simulations: use this link.

Monte Carlo Casino in Monaco. Source: Wikipedia

Monte Carlo Methods

- Statistics packages $\rightarrow$ built-in procedures to generate random variables from most common continuous and discrete distributions.
- Computer languages $\rightarrow$ *random number generators* $\rightarrow$ returns only uniformly distributed independent random variables $U_1, U_2, \ldots \in (0, 1) \rightarrow$ used to then generate RVs X with more complicated distributions (with cdf $F(x)$).
- *Pseudo-random number generator* $\rightarrow$ a long list of numbers
  - the user specifies a *random number seed* which points to the location from which the list will be read
  - same seed $\rightarrow$ same random numbers generated

Computer Simulations

Generating Bernoulli and Geometric RV from Uniform RV

- Simulate a Bernoulli trial with probability of success p. For a Standard Uniform variable U, define

$$X = \begin{cases} 1 & \text{if} \quad U < p \\ 0 & \text{if} \quad U \geq p \end{cases}$$

- We obtain a "success" if $X = 1$ and a "failure" if $X = 0$. Using the Uniform distribution of U, we find that

$$\mathbb{P}(\text{ success }) = \mathbb{P}(U < p) = p.$$

  $\Rightarrow$ we have generated a Bernoulli trial, and X has Bernoulli distribution with probability p.

- A "while" loop of Bernoulli trials which we run until the first success occurs, will generate a Geometric random variable.

Computer Simulations

Generating Bernoulli and Geometric RV from Uniform RV

- For Bernoulli:

```
import numpy as np
# Probability parameter for Bernoulli distribution
p = 0.5  # Example probability
# Generate a standard uniform random variable U
U = np.random.rand()
# Convert U into a Bernoulli random variable X
X = int(U < p)
print(f"U: {U}, X: {X}")
```

- For Geometric:

```
import numpy as np
# Probability parameter for Bernoulli distribution
p = 0.5  # Example probability
# Need at least one trial
X = 1
U = np.random.rand()
# Continue while there are failures
while U > p:
    X += 1
    U = np.random.rand()
# Stop at the 1st success
print(f"X: {X}")
```

Computer Simulations

Generating an abstract discrete RV from Uniform RV

Algorithm:

1. Divide the interval $[0, 1]$ into subintervals:

$$\begin{aligned} A_0 &= [0, p_0) \\ A_1 &= [p_0, p_0 + p_1) \\ A_2 &= [p_0 + p_1, p_0 + p_1 + p_2), \qquad \text{etc.} \end{aligned}$$

   The sub-intervals A_i will have length p_i. We can have a finite or an infinite number of them, depending on the possible values of X.

2. Obtain a Standard Uniform random variable from a random number generator.

3. If U belongs to A_i, let $X = x_i$.
   From the Uniform distribution, it follows that

$$\mathbb{P}(X = x_i) = \mathbb{P}(U \in A_i) = p_i.$$

   $\Rightarrow$ the generated variable X has the desired distribution.

Computer Simulations

Generating a Poisson RV

A Poisson variable takes values $x_0 = 0, x_1 = 1, x_2 = 2, \ldots$ with probabilities

$$p_i = \mathbb{P}(X = x_i) = e^{-\lambda} \frac{\lambda^i}{i!} \text{ for } i = 0, 1, 2, \ldots$$

Following the algorithm, we generate a Uniform random number U and find the set A_i containing U, such that

$$p_0 + \ldots + p_{i-1} \leq U < p_0 + \ldots + p_{i-1} + p_i.$$

In terms of the cdf:

$$F(i-1) \leq U < F(i).$$

Computer Simulations

Generating a Poisson RV

```
import numpy as np
from math import exp, factorial
# Parameter: choose any positive lambda
lambda_val = 5  # Example lambda value
# Generated Uniform variable
U = np.random.rand()
# Initial value, F(0)
i = 0
F = exp(-lambda_val)
# The loop ends when U < F(i)
while U >= F:
    F += exp(-lambda_val) * (lambda_val ** i) / factorial(i)
    i += 1
# Result: we generated a Poisson variable X
X = i
print(f"Generated Poisson variable X: {X}")
```

Theorem

Let X be a continuous random variable with cdf $F_X(x)$. Define a random variable $U = F_X(X)$. Then the distribution of U is $\mathrm{Uniform}(0, 1)$.

Proof: Note that $0 \leq F(x) \leq 1$ for all x, therefore, values of U lie in $[0, 1]$. Then, for any $u \in [0, 1]$, we can find the cdf of U:

$$\begin{aligned} F_U(u) &= \mathbb{P}(U \leq u) \\ &= \mathbb{P}\left(F_X(X) \leq u\right) \\ &= \mathbb{P}\left(X \leq F_X^{-1}(u)\right) && \text{(solve the inequality for } X\text{)} \\ &= F_X\left(F_X^{-1}(u)\right) && \text{(by definition of a cdf)} \\ &= u && \left(F_X \text{ and } F_X^{-1} \text{ cancel each other}\right) \end{aligned}$$

If F_X is not invertible, let $F_X^{-1}(u)$ denote the smallest x such that $F_X(x) = u$. We observe that U has cdf $F_U(u) = u$ and density $f_U(u) = F_U'(u) = 1$ for $0 \leq u \leq 1$. This is the Uniform $(0, 1)$ density $\Rightarrow$ U has Uniform $(0, 1)$ distribution.

Algorithm:

1. Obtain a Standard Uniform random variable from a random number generator.
2. Compute $X = F^{-1}(U)$. In other words, solve the equation $F(X) = U$ for X.

Computer Simulations

Generate an exponential RV

We generate an Exponential variable with parameter λ:

- Generate a Uniform random variable U.
- Recall that the Exponential cdf is $F(x) = 1 - e^{-\lambda x}$.
- Solve the equation
$$1 - e^{-\lambda X} = U.$$
- We obtain
$$X = -\frac{1}{\lambda} \ln(1 - U).$$

- Note that $(1 - U)$ has the same distribution as U, Standard Uniform. Therefore, we can replace U by $(1 - U)$, and variable
$$X_1 = -\frac{1}{\lambda} \ln(U)$$
although different from X, will also have the Exponential (λ) distribution.

Computer Simulations

Generating Continuous RVs

```
import numpy as np

# Parameter: choose any positive lambda
lambda_val = 5  # Example lambda value

# Generate a Uniform random variable U
U = np.random.rand()

# Solve the equation 1 - exp(-lambda * X) = U to obtain X
X = -np.log(1 - U) / lambda_val

print(f"Generated Exponential random variable X: {X}")
```

Monte Carlo Methods

- We know now how to generate random variables from any given distribution $\Rightarrow$ we can put the algorithm in a loop, to generate many variables i.e. a *long run*.
- Then we can estimate probabilities using the long-run proportions, we can estimate expectations using long-run averages, etc.

- For a random variable X, the probability $p = \mathbb{P}(X \in A)$ is estimated by

$$\widehat{p} = \widehat{\mathbb{P}}(X \in A) = \frac{\text{number of } X_1, \ldots, X_N \in A}{N}$$

  where N is the size of a Monte Carlo experiment, $X_1, \ldots, X_N$ are generated random variables with the same distribution as X, and a "hat" represents the estimator.

- In general, the estimator of any quantity a is denoted by $\hat{a}$.

Monte Carlo Methods

- Likewise, we can estimate means, standard deviations, or other distribution characteristics: we generate a Monte Carlo sequence of random variables $X_1, \ldots, X_N$ and compute the necessary long-run averages. The mean $\mathbb{E}[X]$ is estimated by the average

$$\bar{X} = \frac{1}{N}\left(X_1 + \ldots + X_N\right).$$

- If the distribution of $X_1, \ldots, X_N$ has mean μ and standard deviation σ,

$$\mathbb{E}[\bar{X}] = \frac{1}{N}\left(\mathbb{E}[X_1] + \ldots + \mathbb{E}[X_N]\right) = \frac{1}{N}(N\mu) = \mu, \text{ and}$$
$$\operatorname{Var}(\bar{X}) = \frac{1}{N^2}\left(\operatorname{Var} X_1 + \ldots + \operatorname{Var} X_N\right) = \frac{1}{N^2}\left(N\sigma^2\right) = \frac{\sigma^2}{N}.$$

Monte Carlo Methods

- We generate a long sequence of random variables $X_1, X_2, \ldots X_N$ from the *target distribution* and then we can *estimate* different quantitites of interest: probabilities by long-run proportions, expectations by long-run averages, etc.
- Fundamental theorems that underpin the Monte Carlo methodology: The Law(s) of Large Numbers and The Central Limit Theorem.

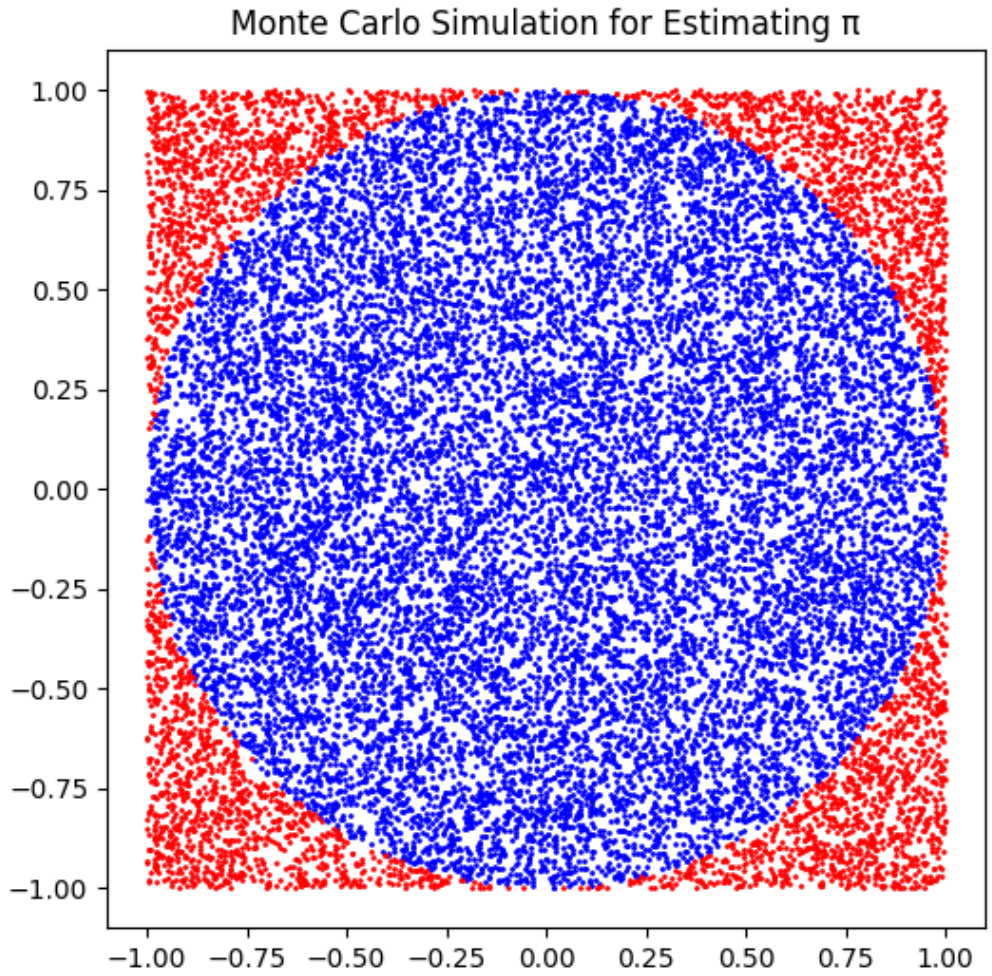
Estimating π using Monte Carlo Simulations

- Inscribe a circle of radius 1 within a square with side length 2.
- The area of the circle is $\pi \times 1^2 = \pi$.
- The area of the square is $2 \times 2 = 4$.
- Ratio of areas: $\frac{\text{Area of the Circle}}{\text{Area of the Square}} = \frac{\pi}{4}$.
- Generate uniformly ditributed random points (x, y) where x and y are between -1 and 1.
- For each point: check if it is inside the circle: $x^2 + y^2 \leq 1$.
- The ratio of points inside the circle to the total number of points approximates $\frac{\pi}{4}$.
- Multiply this ratio by 4 to estimate π.

Estimating π using Monte Carlo Simulations

```
import random
import matplotlib.pyplot as plt

def visualize_pi_estimation(num_samples):
    inside_x = []
    inside_y = []
    outside_x = []
    outside_y = []

    for _ in range(num_samples):
        x = random.uniform(-1, 1)
        y = random.uniform(-1, 1)

        if x**2 + y**2 <= 1:
            inside_x.append(x)
            inside_y.append(y)
        else:
            outside_x.append(x)
            outside_y.append(y)

    plt.figure(figsize=(6,6))
    plt.scatter(inside_x, inside_y, color='blue', s=1)
    plt.scatter(outside_x, outside_y, color='red', s=1)
    plt.gca().set_aspect('equal', adjustable='box')
    plt.title('Monte Carlo Simulation for Estimating pi')
    plt.show()

# Example usage
visualize_pi_estimation(10000)
```

Estimating π using Monte Carlo Simulations



Results for 20 000 samples. The accuracy improves with the number of samples: the more points we generate, the closer the estimate will be to the actual value of π.

Estimating π using Monte Carlo Simulations

Why does it work?

- The Law of Large Numbers states that as the number of trials increases, the average of the results will get closer to the expected value. In our context, as the number of random points increases, the estimated value of π will converge to the actual value of π.
- Define an indicator variable X_i as:

$$X_i = \begin{cases} 1 & \text{if } (x_i^2 + y_i^2 \leq 1) \\ 0 & \text{otherwise} \end{cases}$$

- Then

$$\mathbb{E}[X_i] = \mathbb{P}(X_i = 1) = \frac{\text{Area of the Circle}}{\text{Area of the Square}} = \frac{\pi}{4}.$$

- The Weak Law of Large Numbers ensures that for a large number of samples, the probability that our estimate deviates significantly from the true value of π is very small: for a sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n$ with expected value μ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| \geq \epsilon\right) = 0.$$

Estimating π using Monte Carlo Simulations

Why does it work?

- For us $\mu = \frac{\pi}{4}$. That is

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \frac{\pi}{4}\right| \geq \epsilon\right) = 0.$$

- The Strong Law of Large Numbers guarantees that with a sufficiently large number of samples, our estimate will converge to the true value of π almost surely:
- Given i.i.d. random variables $X_1, X_2, \ldots, X_n$ with expected value μ and $\bar{X}$ as before

$$\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

- For us $\mu = \frac{\pi}{4}$ and then

$$\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = \frac{\pi}{4}\right) = 1.$$

Theoretical topics for the exam

1. Definitions for (□)
   - probability space
   - (unconditional) probability
   - conditional probability
   - independent event
2. Total Probability Rule
3. Bayes' Rule
4. Definition and example for (□)
   - discrete random variable
   - continuous random variable
5. Definitions for (□)
   - probability mass function (pmf)
   - cumulative distribution function (cdf)
   - density function

(□) denotes a *minimal* requirement.

Theoretical topics for the exam

6. Definitions for
   - expectation / expected value (□)
   - variance (□)
   - random vectors
   - joint (cumulative) distribution function
   - covariance and correlation coefficient
7. Markov's inequality and Chebyshev's inequality
8. Sequences of random variables:
   - a.s. convergence - definition
   - convergence in probability - definition
   - convergence in L^2/mean square sense - definition
   - convergence in distribution - definition
   - Link between different types of convergence (Theorem 35 / diagram)
9. Laws of Large Numbers:
   - Weak Law of Large Numbers: Def. 11 + Theorem 36 (with proof).
   - Strong Law of Large Numbers: Def. 37 + Theorem 38.
10. Central Limit Theorem (Theorem 42 and Theorem 44 (with proof)).

(□) denotes a *minimal* requirement.

Exam structure

- 4 parts $\rightarrow$ 70 marks total

  1. Short theoretical question (**one** item from the list above)
  2. Problem involving conditional probabilities/Bayes' rule/total probability rule.
  3. Problem involving distributions (discrete and/or continuous).
  4. Problem involving types of convergence or the weak/strong law of large numbers.

References

1. H. Lisei, *Probability Theory - Lecture Notes* (2022-2023)
2. M. Baron, *Probability and Statistics for Computer Scientists* (2019)
3. H. Lisei, *Probability Theory* (2004)
4. Pictures from *google.com*
5. S. Chongchitnan, *Exploring University Mathematics with Python*. Python codes: here.
6. Bright side of Maths: here.