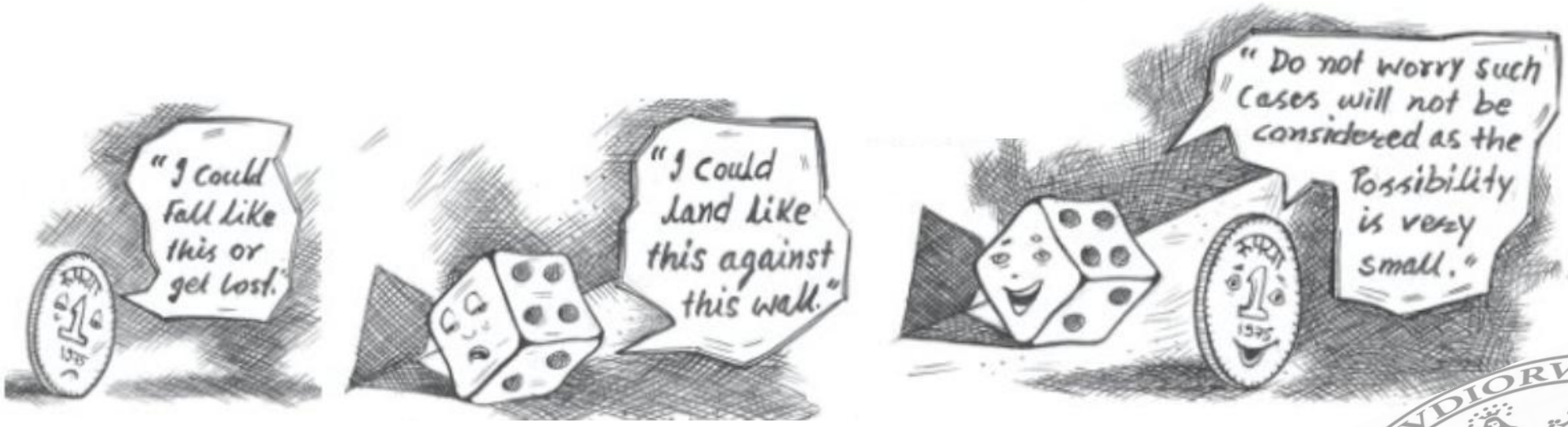
UNIVERSITY OF CAGLIARI

DIEE - Department of Electrical and Electronic Engineering

STOCHASTIC MODELS

-

Probability Theory



alessandro.pilloni@unica.it

Summary

Part 1 - Introduction to probability

Part 2 – Random variables

- Discrete random variables
- Continuous random variables
- Mean, variance and other moments

Preliminary Definition:

- *Experiment*: any activity with an observable uncertain result (ex. flipping a coin, rolling a die, count the number of fishes fished in 1 hour…)
- *Outcome/Realization*: result of the experiment accounting what was observed (ex. observed face of the die, #fishes fished in 1 hour...)
- *Trial*: a repetition of the same experiment
- *Sample Space* (or *Universe*): The set S of all possible outcomes (elementary events) of a given experiment. It can be:
  - discrete: countable (finite or infinite)
  - continuous: uncountable (infinite)
- *Event*: a subset A of S ($A \subseteq S$)

Notice that the impossible event is $A = \varnothing$, while the certain event is $A = S$

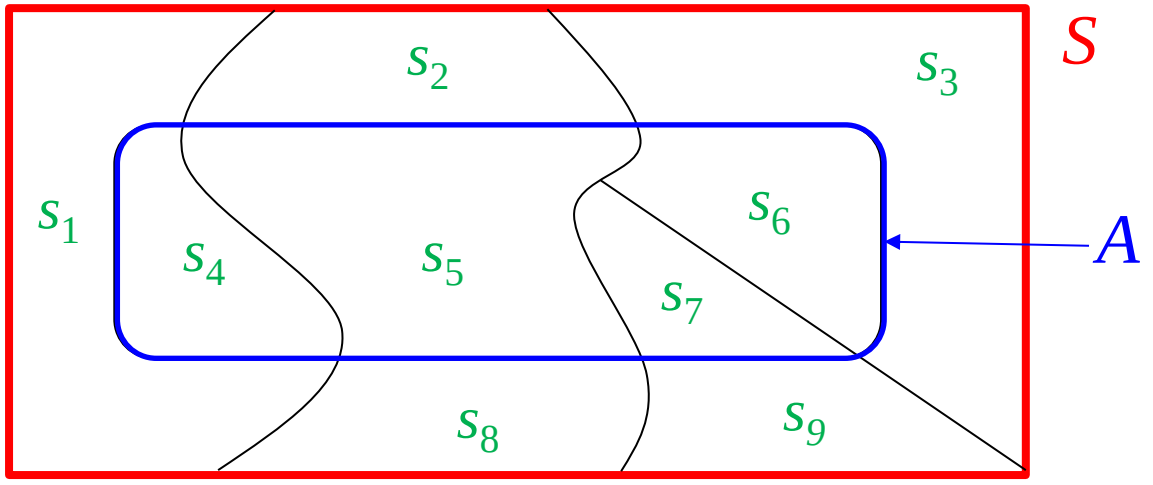
Basics on Venn diagram

- "*Set Theory*" helps in converting our language into maths
- Venn diagram and Set theory operators and relationships, such as

(see ∪, ∩, +, / and De Morgan Laws)

helps the identification of all the outcomes of an experiment.

- A **box** (i.e. a **closed set**) can be used to denote the sample space



Elementary events: $s_1, s_2, \ldots, s_9$

Event: $A = \{ s_4, s_5, s_6, s_7 \}$

Sample space: $S = \{ s_1, s_2, \ldots, s_9 \}$

Examples:

1) Experiment: Extract 2 balls from an urn with **n black** (“b”) and **n white** (“w”) balls

Sample space: $S = \{ww, wb, bw, bb\}$ (ex. finite discrete space)

Event: the two balls have same color $A = \{ww, bb\} \subset S$

2) Experiment: Flip a coin until a head (H) appears

Sample space: $S = \{1, 2, 3, \ldots\}$ (ex. countably infinite discrete space)

coin tosses until a head H appears

Event: wait 3 or more flips $A = \{3, 4, \ldots\} \equiv S \setminus \{1, 2\} \subset S$

(An event can account of infinite many realizations)

3) Experiment: Roll of a dice

Sample space: $S = \{1,2,3,4,5,6\}$ (finite discrete space)

Examples of events:

Event A: observe an even number $A = \{2,4,6\} \subset S$

Event B: observe a number > 6 (here impossible)

Event C: observe a number < 7 (here is a certain event)

4) Experiment: Arrival time of customers at a queue

Sample space: $S = \{t \in \mathbb{R} \mid t \geq 0\}$ (continuous space)

Arrival time

Sometimes determine the sample space requires a bit of reasoning

5) A guy is used to go to work by train.

- He is used to arrive at the platform at between 9.01 and 9.04
- He may catch two trains: that of the 9.02 or that of the 9.04 AM.
- Suppose we are interested to the **waiting time in minute** (W) at the station.

By plotting W as a function of the arrival time, one derives:

Sample space is: $S_W = \{t \in [0,2]\}$
(continuous space)

Event example: wait no more than 1min

$A = \{t \in [0,1]\}$ (uncountable event)

Events always satisfy the following axiomes:

1) if A is an event, its complement ($S \setminus A$) is also an event

2) the union of events $A_1, A_2, \ldots \subseteq S$ it's still an event, i.e.,

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

Definitions

- Two events A_1, A_2 are said to be mutually exclusive if

  $A_1 \cap A_2 = \varnothing$ (their intersection is the Empty set)

- A set of events $\{A_1, A_2, \ldots, A_k\}$ is said to be exhaustive if

  $A_1 \cup A_2 \cup \ldots \cup A_k = S$ (their union is the Sample Set)

Probability (for discrete Sample space)

Consider a discrete sample space (finite or infinite) of elementary events $S = \{ s_1, s_2 , \ldots \}$

- Probability function: $Pr : S \rightarrow [0,1]$

Associates/maps "→" each event $s \in S$ to number between 0 and 1, such that

$$\sum_{s \in S} Pr(s) = 1$$

- Follows that the probability of an event $A \subseteq S$ satisfies:

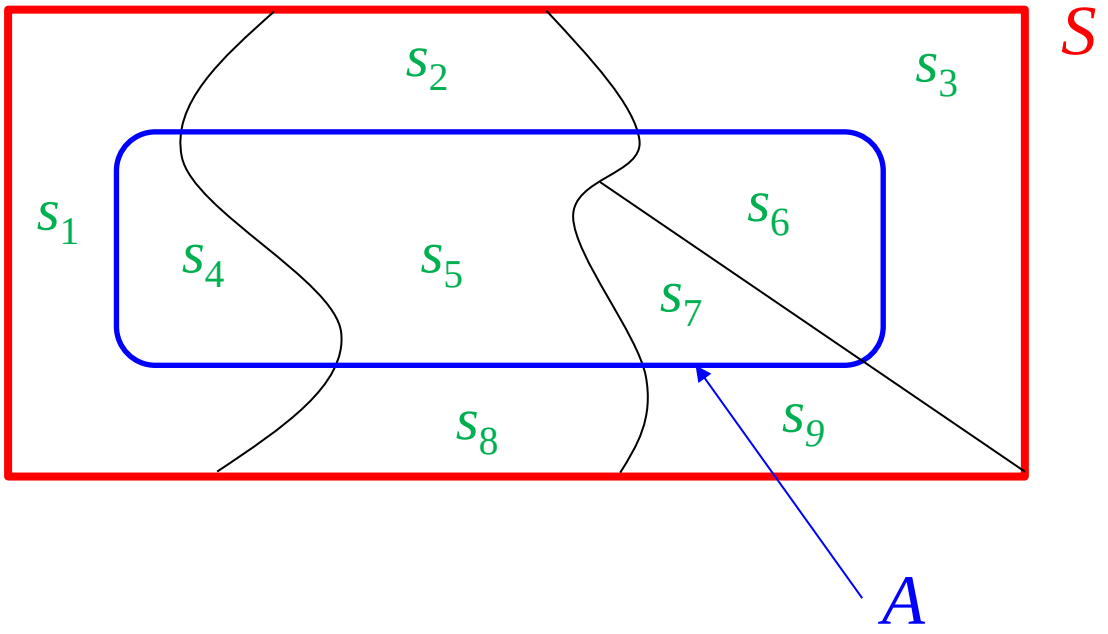
$$Pr(A) = \sum_{s \in A} \Pr(s) \in [0,1]$$

Probability and Venn Diagram

$$Pr(S) = Pr(s_1) + Pr(s_2) + \cdots + Pr(s_9) = 1$$

$$Pr(s_i) = \frac{Area\ s_i}{Area\ S}$$

Normalization factor



$$Pr(A) = Pr(s_4) + Pr(s_5) + Pr(s_6) + Pr(s_7)$$

Elementary events: $s_1, s_2, \ldots, s_9$

Sample space: $S = \{ s_1, s_2, \ldots, s_9 \}$

Event: $A = \{ s_4, s_5, s_6, s_7 \}$

Physical meaning of probability:

Suppose we can repeat a (random) experiment N times.

Given an event A, its probability $Pr(A)$ represents the ratio when N tends to infinity between

- number of trials N_A in which A occurs and
- total number of trials N

$$Pr(A) = \lim_{N \to \infty} \frac{N_A}{N}$$

Pierre-Simon Laplace (1749 –1827)

P.S. Laplace (1812) "Théorie analytique des probabilités"

Example: Fair VS Unfair coins

$$Pr(A) = \lim_{N \to \infty} \frac{N_A}{N}$$

Coin	Total Flips	Heads	Tails
0	100	53	47
1	100	55	45
2	100	49	51
3	100	41	59
4	100	39	61
5	100	27	73
6	100	0	100

Impossible event $A = \varnothing \rightarrow Pr(\varnothing) = 0$

Certain event $A = S \rightarrow Pr(S) = 1$

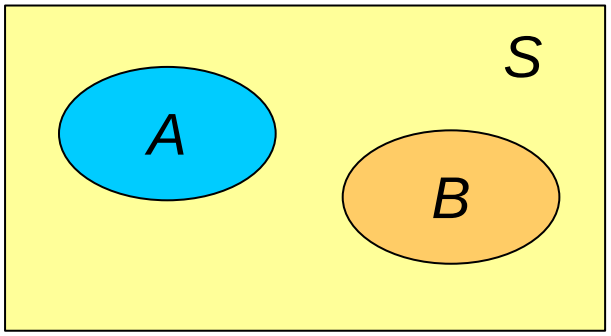
Addition's Rule:

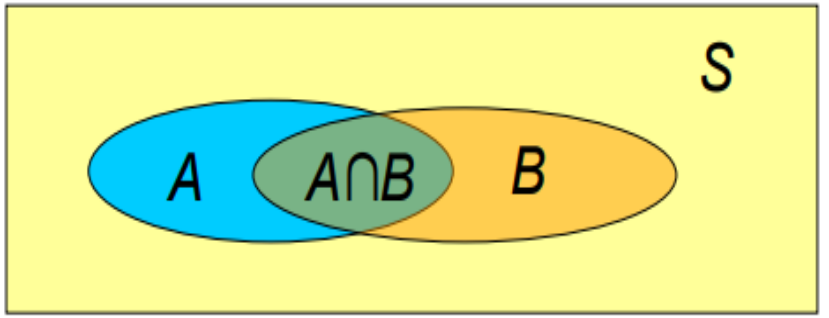
- Events A and B are disjoint (mutually exclusive) ($A \cap B = \varnothing$)

$$Pr(A \cup B) = Pr(A) + Pr(B)$$

- Events A e B are not disjoint ($A \cap B \neq \varnothing$)

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$





Example 1: Extract two balls (with replacement) from an urn containing $2n$ balls (n black, n white)

S ={ww,wb,bw,bb}

s	*Pr(s)*
ww	1/4
wb	1/4
bw	1/4
bb	1/4

Event: two balls of the same color

A = {ww,bb} $Pr(A) = Pr(\text{ww}) + Pr(\text{bb}) = 0.5$

Without replacement:

$$\Pr(ww) = \frac{n}{2n} \cdot \frac{(n-1)}{(2n-1)}$$

$$\Pr(wb) = \frac{n}{2n} \cdot \frac{n}{(2n-1)}$$

...

...

s	*Pr(s)*
ww	(n-1)/[2(2n-1)]
wb	n/[2(2n-1)]
bw	n/[2(2n-1)]
bb	(n-1)/[2(2n-1)]

Example 2: Number of coin flips until get H (head)

$S = \{1,2,3,\ldots,k,\ldots\}$

s	*Pr*(*s*)
1	1/2
2	1/4
3	1/8
…	…
k	$1/2^k$
…	…

H

TH

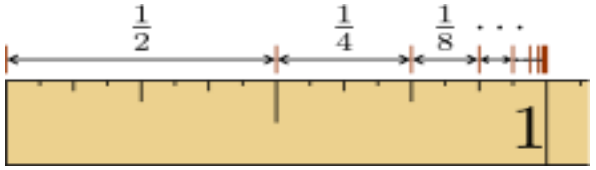
TTH

T…TH

k-1 times

Note that:

$$\sum_{k=1}^{\infty} \underbrace{\left(\frac{1}{2}\right)}_{\rho}^{k} = 1$$



This is a geometric series with a common ratio $\rho = \frac{1}{2} < 1$

(thus it converges)

Event: H succeed in three flips or less

$A = \{1,2,3\}$ $\quad Pr(A) = Pr(1) + Pr(2) + Pr(3) = 7/8$

Example 3: Event of picking a face card (K, Q, J) **or** a red card (♦,♥) from a deck of 52 cards

A = {pick a face card}

B = {pick a red card}

$Pr(A)=12/52$

♠	A	2	3	4	5	6	7	8	9	10	J	Q	K
♣	A	2	3	4	5	6	7	8	9	10	J	Q	K
♦	A	2	3	4	5	6	7	8	9	10	J	Q	K
♥	A	2	3	4	5	6	7	8	9	10	J	Q	K

$Pr(B)=26/52$

$Pr(A \cap B)=6/52$

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B) = 32/52$$

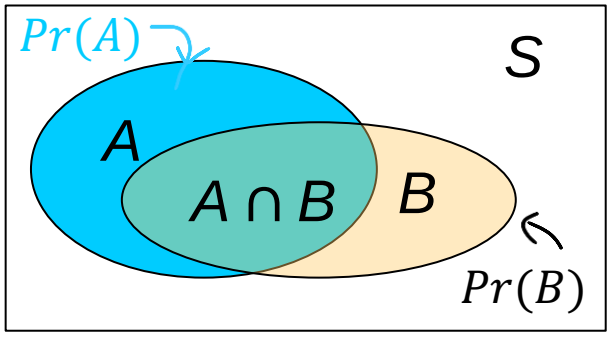
Conditional probability

Given two random events A e $B \subseteq S$, with $Pr(B) > 0$ the conditional probability of A given B is defined as the ratio between

- probability of the joint occurrence of events A and B $(A \cap B)$

and

- probability of occurrence of B



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

Explanation: If event B has occurred, the new sample space is reduced to B. Thus, the possible outcomes for A are restricted to those in which A and B both occurs with respect to the outcomes for B.

Example 4: Somebody (secretly) rolls 2 dice.

Compute the probability that the roll of D1 is 2, given the information that the sum of the two rolls is no greater than 5.

A={D1 = 2}

$Pr(A) = \frac{6}{36} = \frac{1}{6}$

B={D1+D2<=5}

$Pr(B) = \frac{10}{36}$

$Pr([D1 = 2] \cap [D1+D2 \leq 5]) = \frac{3}{36}$

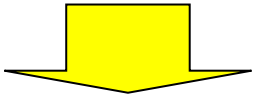
+		D2					
		1	2	3	4	5	6
D1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

$$Pr(D1 = 2 \mid D1+D2 \leq 5) = \frac{Pr([D1 = 2] \cap [D1+D2 \leq 5])}{Pr(D1+D2 \leq 5)} = \frac{\frac{3}{36}}{\frac{10}{36}} = \frac{3}{10}$$

Stochastic independence

Two events A and B are stochastically independent if the knowledge A has occurred does not affect the probability B will occur (and vice-versa), i.e.,

$$Pr(A|B)=Pr(A) \quad \text{or equivalently} \quad Pr(B|A)=Pr(B)$$



$$Pr(A \cap B) = Pr(A|B) \cdot Pr(B) = Pr(A) \cdot Pr(B)$$

Example 5: Roll of a dice: $S = \{1,2,3,4,5,6\}$

Event A : roll a number < 4 $\Rightarrow A = \{1,2,3\}$

Event B : roll an even number $\Rightarrow B = \{2,4,6\}$

Are A and B stochastically independent?

$Pr(A) = Pr(B) = 1/2$

$A \cap B = \{2\} \quad \Rightarrow \quad Pr(A \cap B) = 1/6$

If B occurs, the probability that also has A occured is

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{1/6}{1/2} = \frac{1}{3} \neq \frac{1}{2} = Pr(A)$$

NOT Stochastically Independent

Exercise: Consider the experiment of flipping two coins C1 and C2

Compute the following probabilities

Case 1: $Pr(\text{C1=H} \cap \text{C2=H})$ (i.e. only {HH})

Case 2: $Pr(\text{C1=H} \cup \text{C2=H})$ (i.e. {HH, HT, TH})

then, provide an interpretation of what is obtained.

Solution:

$$Pr(\text{C1=H} \cap \text{C2=H}) = Pr(\text{C2=H} \mid \text{C1=H})\, Pr(\text{C1=H}) = 1/4$$

$$Pr(\text{C1=H} \cup \text{C2=H})$$

$$= Pr(\text{C2=H}) + Pr(\text{C1=H}) - Pr(\text{C1=H} \cap \text{C2=H}) = 3/4$$

Note: $\cap$ means AND (or joint) whereas $\cup$ means OR, thus Case 2 do not consider only the case where {TT}.

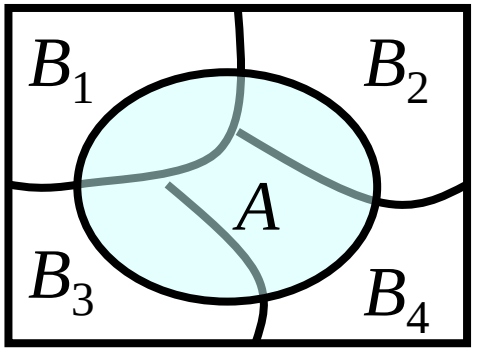
Law of Total probability

Let $B_1, B_2, \ldots, B_k$ be a finite or countably infinite partition of a sample space S

$$\bigcup_{i=1}^{k} B_i = S \quad \text{with} \quad B_i \cap B_j = \emptyset \quad \forall i \neq j$$

Then for any event A in S holds:

$$Pr(A) = \sum_{i=1}^{k} Pr(A|B_i) \cdot Pr(B_i)$$



Proof:

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_k)$$

$$Pr(A) = \sum_{i=1}^{k} \Pr(A \cap B_i) = \sum_{i=1}^{k} Pr(A|B_i) \cdot Pr(B_i)$$

Interpretation: It expresses the total probability of an outcome that can be realized via several **distinct events**, *hence the name*.

Example 6a: A factory's output is produced by

3 machines M_1, M_2,M_3

each accounting for 20%, 30%, 50% of the whole ouput.

The fraction of defective (D) items produced by each machine is:

$$\Pr(D|M_1) = \frac{5}{100} \ , \ \Pr(D|M_2) = \frac{3}{100} \ , \ \Pr(D|M_3) = \frac{1}{100}$$

What is the probability that a produced item is defective?

According to the Law of Total Probability :

$$\begin{aligned}\Pr(D) &= \sum_i \Pr(D|M_i)\Pr(M_i) \\ &= \Pr(D|M_1)\Pr(M_1) + \Pr(D|M_2)\Pr(M_2) + \Pr(D|M_3)\Pr(M_3) \\ &= \frac{5}{100} \times \frac{20}{100} + \frac{3}{100} \times \frac{30}{100} + \frac{1}{100} \times \frac{50}{100} = \frac{240}{100^2}\end{aligned}$$

Bayes' theorem

Allows to determine posterior inferences, i.e. allows to know the probability of something "unobservable" given an "observed" event.

Remember : conditional probability

$$Pr(A \cap B) = Pr(A|B) \cdot Pr(B) = Pr(B|A) \cdot Pr(A)$$

Theorem (Bayes theorem):

Reverse conditional probability

Prior info
(the beliefs we have before observing some data)

$$Pr(A|B) = \frac{Pr(B|A) \cdot Pr(A)}{Pr(B)}$$

Posterior info
(the probability obtained after the data has been taken into account)

Normalization factor
(is related with an observed event)

Example 6b: The output of a factory is produced by 3 machines M_1, M_2,M_3 each generating 20%, 30%, 50% of the whole ouput.

The fraction of defective (D) items per machine output is:

$$Pr(D|M_1)=5\%, \quad Pr(D|M_2)=3\%, \quad Pr(D|M_3)=1\%$$

thus $Pr(D)=240/100^2$.

What is the probability that, chosen a defective item, it was produced by M_3?

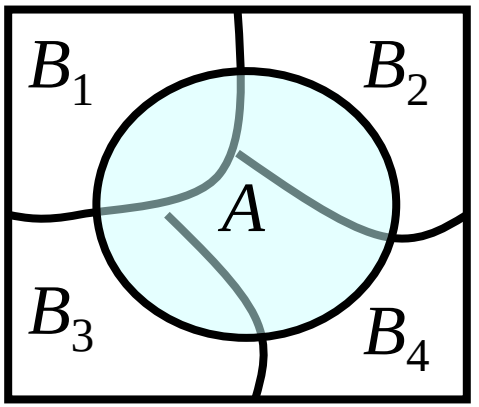
According to Bayes' theorem:

$$Pr(M_3 \mid D) = \frac{Pr(D \mid M_3) \cdot Pr(M_3)}{Pr(D)} = \frac{\frac{1}{100} \cdot \frac{50}{100}}{\frac{240}{100^2}} = \frac{50}{240} \approx 21\%$$

Bayes' theorem (total form)

Remember : law of total probability

$$Pr(A) = \sum_{i=1}^{k} Pr(A|B_i) \cdot Pr(B_i)$$



Considering the conditional probability definition

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

Reverse conditional prob.

it follows that

$$Pr(B_j|A) = \frac{Pr(A \cap B_j)}{Pr(A)} = \frac{Pr(A|B_j) \cdot P(B_j)}{\sum_{i=1}^{k} Pr(A|B_i) \cdot Pr(B_i)}$$

Prior info (the beliefs we have before observing some data)

Posterior info

Normalization factor

Note: It is fundamental for classification purposes in spam filtering, network intrusion detection, and most of recommendation systems by probabilistically classifying data.
These applications leverage prior info (e.g. some words in an email) to provide raccomentations.

Sensitivity and Specificity in Classification problems

Example: A professor camed up with a test to detect if a student cheats (event C).

Suppose: $Pr(C)$= 20% . Student positive to the test are denoted by (+), otherwise (-)

The Professor's test has the following performance:

- Sensitivity of 90%: gives a measure on how well a test can identify true positives

$$Pr(+|C) = 0.95$$

- Specificity of 95%: gives a measure on how well a test can identify true negative

$$Pr(-|notC) = 0.9$$

- What's the probability that a student testing positive + is cheating (event *C*)?

$$Pr(C|+) = \frac{Pr(+|C) \cdot Pr(C)}{Pr(+)} = \frac{Pr(+|C) \cdot Pr(C)}{Pr(+|C) \cdot Pr(C) + Pr(+|notC) \cdot Pr(notC)}$$

$$= \frac{0.95 \cdot 0.2}{0.95 \cdot 0.2 + (1 - 0.9) \cdot (1 - 0.2)} = 0.82$$

It this a good test? For this class YES!!!

What if $Pr(C) = 0.05?$ $\Rightarrow \Pr(C|+) = 0.48$

For this praiseworthy class it is not!!!

A good test should always take care on the population under study $Pr(C)$ and priorities, your Screening (high sens.) VS diagnostic (high spec.)

Sensitivity and Specificity in Classification problems

Example: A professor camed up with a test to detect if a student cheats (event C).

Suppose: $Pr(C)$= 20% . Student positive to the test are denoted by (+), otherwise (-)

The Professor's test has the following performance:

- Sensitivity of 90%: gives a measure on how well a test can identify true positives

$$Pr(+|C) = 0.95$$

- Specificity of 95%: gives a measure on how well a test can identify true negative

$$Pr(-|notC) = 0.9$$

- What's the probability that a student testing positive + is cheating (event *C*)?

$$Pr(C|+) = \frac{Pr(+|C) \cdot Pr(C)}{Pr(+)} = \frac{Pr(+|C) \cdot Pr(C)}{Pr(+|C) \cdot Pr(C) + Pr(+|notC) \cdot Pr(notC)}$$

$$= \frac{0.95 \cdot 0.2}{0.95 \cdot 0.2 + (1 - 0.9) \cdot (1 - 0.2)} = 0.82$$

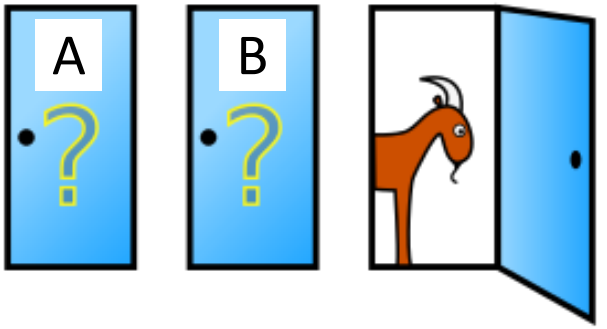
It this a good test? For this class YES!!!

What if $Pr(C) = 0.05?$ $\Rightarrow \Pr(C|+) = 0.48$

For this praiseworthy class it is not!!!

A good test should always take care on the population under study $Pr(C)$ and priorities, your Screening (high sens.) VS diagnostic (high spec.)

Example 7: Monty Hall problem.



In a TV game show, there are 3 doors. Behind one door there is a car; behind the others, goats. To win the car you have to choose the right door.

You pick a door, say "A", thus $\Pr(\mathrm{A}) = 1$, whereas

$$Pr(Car_A) = \frac{1}{3} \quad \Rightarrow \quad Pr(Car_A|A) = \frac{1}{3} \quad \text{(due to independence)}$$

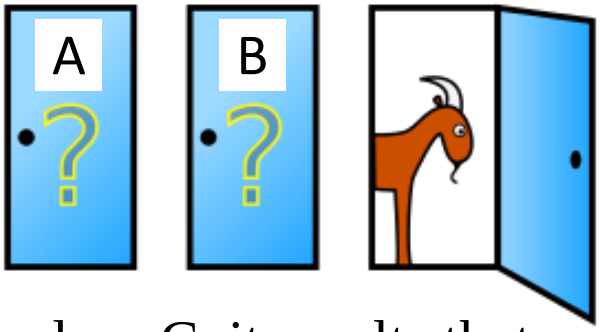
The Host, who knows what's behind each door, opens another door, say "C", showing one of the two goats.

He then says: "Do you want to change your choice and pick door B?"

Is it to your advantage to change your original choice?

In other words, is $Pr(Car_B|H_c, A)$ greater than $Pr(Car_A|A) = \frac{1}{3}$?

Example 7: Monty Hall problem.



Solution: Let H_C be the event the Host open the door C, it results that

$$\begin{aligned}\Pr(H_C) =& \Pr(H_C|Car_A,A)\Pr(Car_A,A)\\ &+\Pr(H_C|Car_B,A)\Pr(Car_B,A)\\ &+\Pr(H_C|Car_C,A)\Pr(Car_C,A) \quad = \frac{1}{2}\cdot\frac{1}{3}+1\cdot\frac{1}{3}+0\cdot\frac{1}{3}=\frac{1}{2}\end{aligned}$$

It follows that:

$$\Pr(Car_A|H_C,A) = \frac{\Pr(H_C|Car_A,A)\cdot\Pr(Car_A,A)}{\Pr(H_C)} = \frac{\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

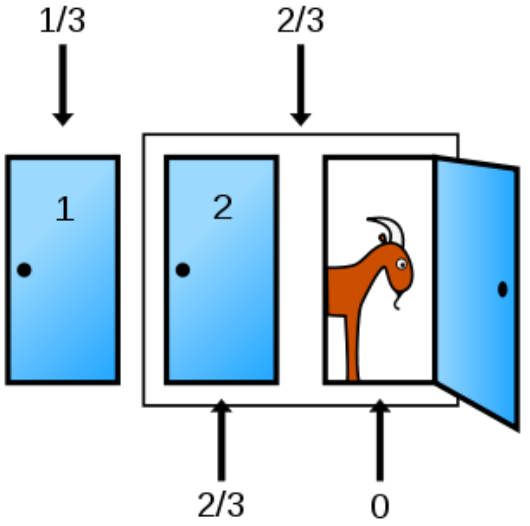
$$\Pr(Car_B|H_C,A) = \frac{\Pr(H_C|Car_B,A)\cdot\Pr(Car_B,A)}{\Pr(H_C)} = \frac{1\times\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

$$\Pr(Car_C|H_C,A) = 0$$

Thus, it is better to change your door!

Example 7: Monty Hall problem simulator: No change

```
clear all
n_trials=10000;
success_with_change=0;
success_no_change=0
for i=1:n_trials
  doors_t0=[1 2 3];
  host_choices=doors_t0;
  doors_t1=doors_t0;
  car_in_door=doors_t0(unidrnd(3));      % random choice from 1 to 3
  pick_up_door_t0=doors_t0(unidrnd(3)); % random choice from 1 to 3
  % continue in the next slide
```



Guess what was behind your door?

Well… 2/3 of the time, if you don't change, you will find a…

```
  % use-case: you change
  if pick_up_door_t0==car_in_door
    host_choices(host_choices==pick_up_door_t0)=[];
    host_open=host_choices(unidrnd(2));
    pick_up_door_t1=host_choices(host_choices~=host_open);
  elseif pick_up_door_t0~=car_in_door
    host_choices(host_choices==car_in_door)=[];
    host_open=host_choices(host_choices~=pick_up_door_t0);
    doors_t1(doors_t1==host_open)=[];
    pick_up_door_t1=doors_t1(doors_t1~=pick_up_door_t0);
  end
  if pick_up_door_t1==car_in_door
    success_with_change=success_with_change+1;
  end
end
pr_success_no_change=success_no_change/n_trials
pr_success_with_change=success_with_change/n_trials
```

Guess what was behind your door?

Well… 2/3 of the time, if you don't change, you will find a…

Summary

Part 1 - Introduction to probability

Part 2 – Random variables

- Discrete random variables
- Continuous random variables
- Mean, variance and other moments

Random variables

A **random variable *X*** takes value from the given Sample Space

$$S_X = \{x_1, x_2, x_3, \ldots\}$$

according to a specified probability "measure"

- When S_X is countable one has a discrete random variable denoted

$$X = (S_X, p_X) \quad \text{with} \quad p_X(x_i) = Pr(X = x_i) : S_X \to [0,1]$$

and the probability measure is called probability mass function (pmf) of X.

- When S_X is uncountable one has a continuous random variable denoted

$$X = (S_X, f_X) \quad \text{with} \quad f_X : S_X \to \mathbb{R}_{\geq 0}$$

and the probability measure is called probability density function (pdf) of X.

Summary

Part 1 - Introduction to probability

Part 2 – Random variables

- Discrete random variables
- Continuous random variables
- Mean, variance and other moments

Discrete random variables

A discrete random variable takes value in a countable set S_X.

We consider a numerical sample space

$$S_X \subseteq \mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$$

A discrete random variable is denoted as

$$X = (S_X, p_X) \quad \text{and} \quad p_X(x) = Pr(X = x) : S_X \to [0,1]$$

where $p_X(k) = Pr(X = k)$ is called probability mass function (**pmf**) of X.

The probability mass function satisfies $\sum_{k \in S_X} p_X(k) = 1$

When clear, one may write (S, p) instead of (S_X, p_X) thus simplifying the notation.

Experiment: Coin flips to get head $X = (S, p)$

Here $S = \{1, 2, 3, \dots\}$ and its $p : S \to [0,1]$ is shown in table.

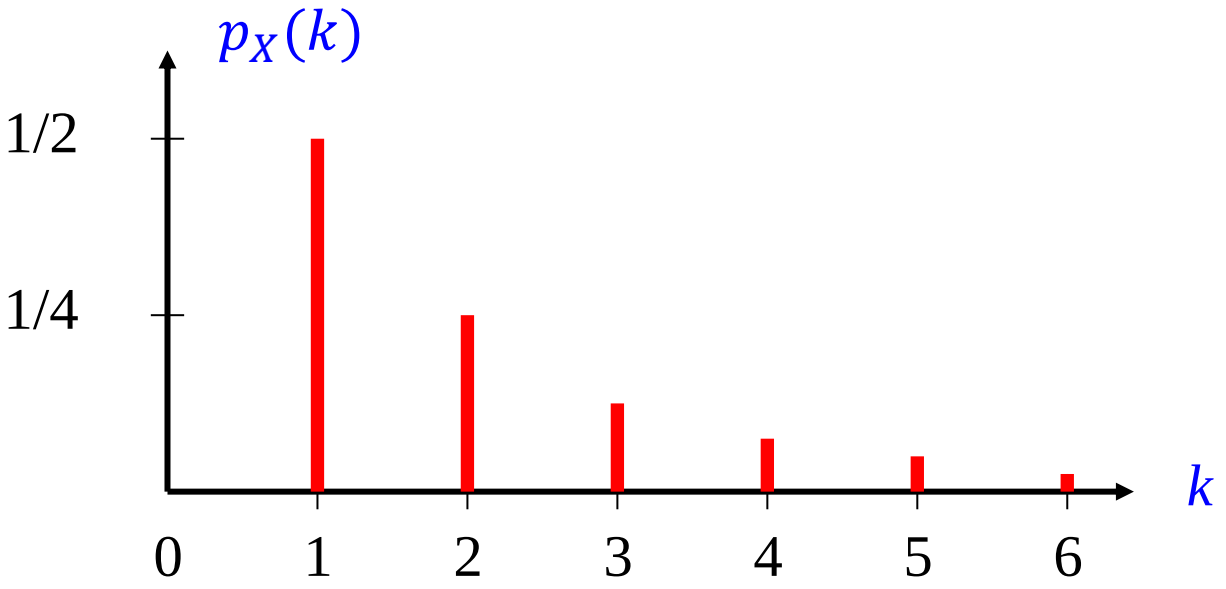
1° toss H T

2° toss HH HT TH TT

3° toss TTH over 8 simple permutation

…

k° toss T…TH over 2^k simple permutation



k	$p_X(k)$
1	1/2
2	1/4
3	1/8
4	1/16
⋮	⋮
k	$1/2^k$
⋮	⋮

* Simple permutation are also called «permutation with repetition»

37

Experiment: Flip 2 coins *C1* and *C2*.

The number of heads is a random variable $X = (S, p)$ with:

- The sample space is $S = \{0, 1, 2\}$ (simple combinations)
- The PMF p_X: $S \to [0,1]$ is given by

$$p_X(0) \equiv \Pr(C_1 = T, C_2 = T)$$
$$= \Pr(C_2 = T | C_1 = T)\Pr(C_1 = T)$$
$$= \Pr(C_2 = T)\Pr(C_1 = T) = \frac{1}{4}$$

σ	*Pr*(σ)	*k*	*p*(*k*)
TT	1/4	0	1/4
TH	1/4	1	1/2
HT	1/4		
HH	1/4	2	1/4

Note: the experiment has a non-numerical sample space

Σ = { TT, TH, HT, HH } *(here T = tail; H = head)*

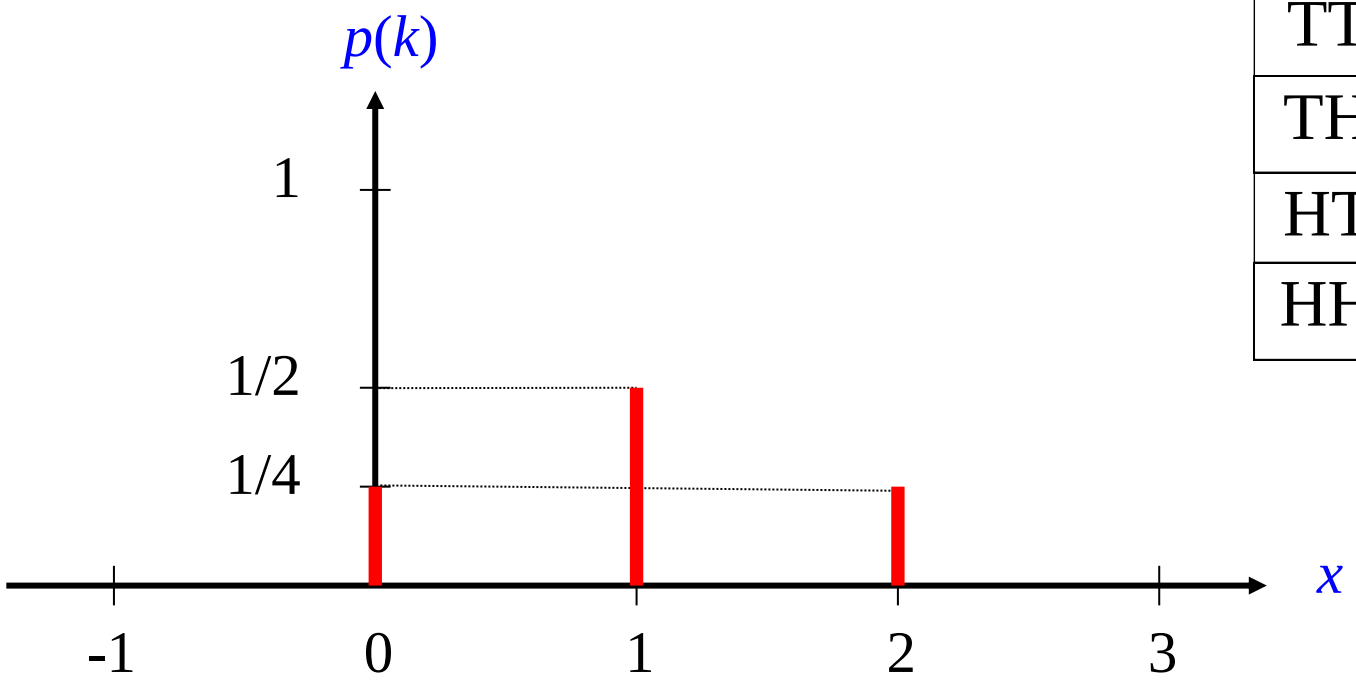
but in the r.v. we consider a numerical space $S = \{0, 1, 2\}$

* Wrt to permutation in combinations the order doen't matter thus HT and TH are counted as a single event in the Sample-space cardinality.

A discrete random variable and its PMF can be easily represented by a histogram.

Ex: for the previous random variable

σ	Pr(s)	k	$p(k)$
TT	1/4	0	1/4
TH	1/4	1	1/2
HT	1/4		
HH	1/4	2	1/4

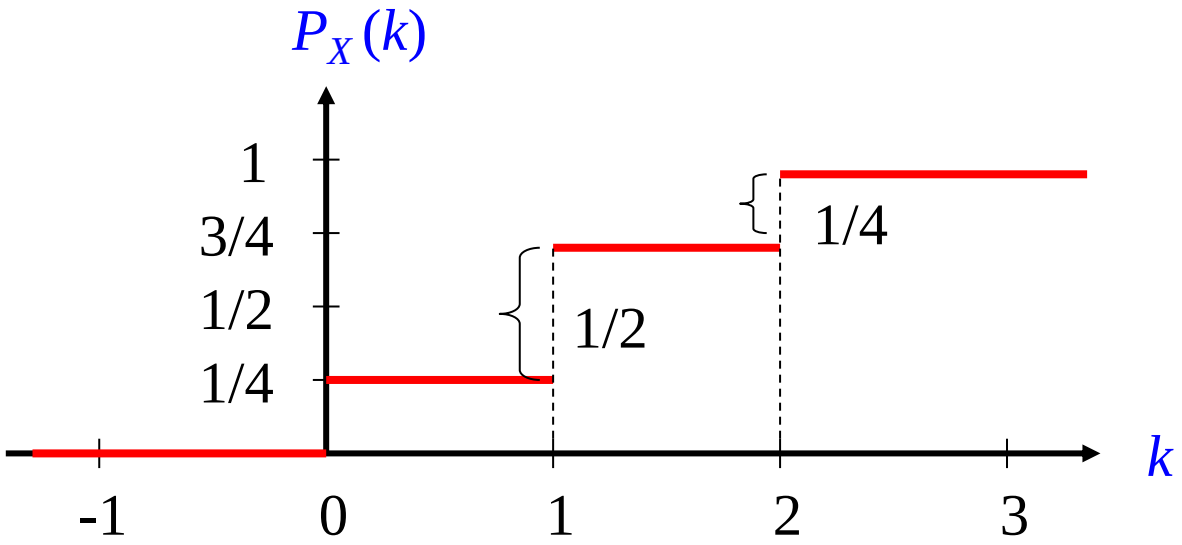


Cumulative distribution function (cdf) of a discrete r. v.

Denotes the probability that X assumes values less than or equal to k:

$$P_X(k) = Pr(X \le k) = \sum_{s \in S_X, s \le k} p_X(s) = \sum_{s=-\infty}^{k} p_X(s)$$

Example: number of heads tossing 2 coins



$$P_X(1) = Pr(X \le 1)$$

$$= Pr(X = 0) + Pr(X = 1) = 3/4$$

Complementary cumulative probability function of a discrete r. v.

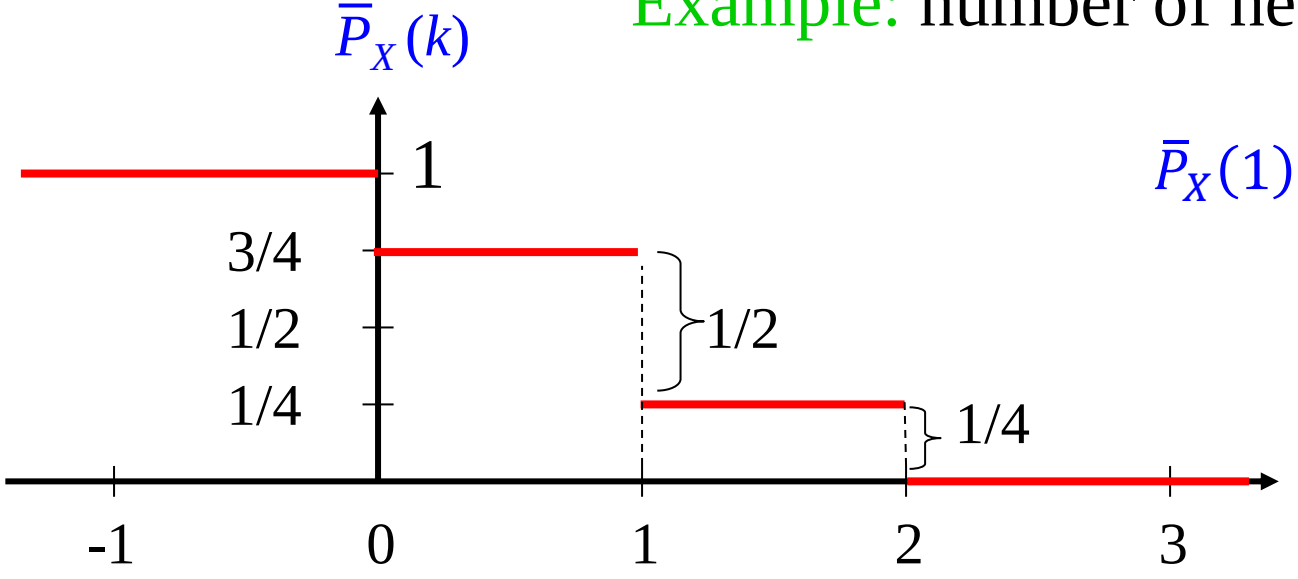
Denotes the probability that X assumes a value greater than x:

$$\bar{P}_X(k) = Pr(X > k) = \sum_{s \in S_X,\, s > k} p_X(s) = \sum_{s=k+1}^{+\infty} p_X(s)$$

Clearly: $\bar{P}_X(k) = Pr(X > k) = 1 - P_X(k) = 1 - Pr(X \leq k)$

i.e., the two cumulative functions contain the same information.

Example: number of heads tossing 2 coins



$\bar{P}_X(1) = Pr(X > 1)$

$=1-Pr(X \leq 1)=1/4$

Joint probability function (discrete random variables)

Example: Let $X_1=(S_1,p_1)$ and $X_2,=(S_2,p_2)$ be two r.v. associated two consecutive coin's flips the Joint probability of X_1, X_2 maps $S_1 \times S_2$ into $[0,1]$ as follows

$$p_{X_1,X_2}(x_1,x_2) = Pr(X_1 = x_1, X_2 = x_2) : S_1 \times S_2 \rightarrow [0,1]$$

where $\times$ denotes the Cartesian product (i.e, all simple permutation in S_1 and S_2)

$$S_1 \times S_2 = \{H,T\} \times \{H,T\} = \{(H,H),(H,T),(T,H),(T,T)\}$$

Note: From $p_{X_1,X_2}(x_1,x_2)$ we can derive each $p_{X_i}(x_i)\ \forall\ k_i \in S_i$ by removing the dependence of the other r.v.

p_{X_2} \ p_{X_1}	H	T
H	1/4	1/4
T	1/4	1/4

$$p_{X_2}(H) = \sum_{\forall\, x_1 \in S_1} p_{X_1,X_2}(x_1,H) = \frac{1}{2}$$

$$p_{X_2}(T) = \sum_{\forall\, x_1 \in S_1} p_{X_1,X_2}(x_1,T) = \frac{1}{2}$$

Note: The operation aimed at obtaining p_{X_1} (or p_{X_2}) from the joint pmf p_{X_1,X_2} is called «marginalization». It consists to an application of the Law of Tot. Prob.

Joint probability function (discrete random variables)

Given n discrete r.v. X_i,=(S_i,p_i) for i=1, …, n, their joint probability function is

where $p_{X_1,X_2,\cdots,X_n}: S_1 \times S_2 \times \cdots \times S_n \rightarrow [0,1]$

$$p_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = Pr(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$

Given the joint probability function, we can always compute the probability function of each single random variable as follows (with no loss of generality we consider X_1):

$$p_{X_1}(k_1) = \sum_{\forall\ (x_2,\cdots,x_n) \in S_2 \times \cdots \times S_n} p_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots,\ x_n)$$

On the contrary, the single probability functions of each r.v. do not have all information needed to compute the joint probabilities

Exercise **Tossing a coin** (0 = tail; 1 = head)

Case A: the coin is tossed every hour (at the hour)

Case B: the coin is tossed every 24h (at midnight)

Between two tossings the coin is not moved. Compute $p_{X_1,X_2}(x_1, x_2)$

Define two random variables (r.v.)

$X_1=(S_1,p_1)$: coin value at 7:30am

$X_2=(S_2,p_2)$: coin value at 9:30pm

$$\begin{cases} S_1 = S_2 = \{0,1\} \\ p_1(k) = p_2\ (k) = 0.5, \forall\ k \in \{0,1\} \end{cases}$$

Joint probabilities

Case A: $p_{X_1,X_2}(0,0) = p_{X_1,X_2}(0,1) = p_{X_1,X_2}(1,0) = p_{X_1,X_2}(1,1) = 0.25$

Case B: $p_{X_1,X_2}(0,0) = p_{X_1,X_2}(1,1) = 0.5;\ p_{X_1,X_2}(0,1) = p_{X_1,X_2}(1,0) = 0$

Conditional Probability for random variables

The notion of conditional probability defined for events also applies to two random variables $X = (S_X, p_X)$ and $Y = (S_Y, p_Y)$

$$Pr(X = x|Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)} \Rightarrow p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Since by definition of joint probability: $p_X(x) = \sum_{\forall y \in S_Y} p_{X,Y}(x,y)$

we get the total probability law for random variables

$$p_X(x) = \sum_{\forall y \in S_Y} p_{X,Y}(x,y) = \sum_{\forall y \in S_Y} p_{X|Y}(x|y) \cdot p_Y(y)$$

Independence of Random Variables

We say that two random variables $X=(S_X, p_X)$ and $Y=(S_Y, p_Y)$ are stochastically independent if the events $\{X = x\}$ and $\{Y = y\}$ are independent for every x and y.

Thus,

$$p_{X|Y}(x|y) = p_X(x)$$

or equivalently

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y) \qquad \forall x, y$$

Sum of (independent) random variables

Consider independent random variables $X=(S_X, p_X)$ and $Y=(S_Y, p_Y)$

Their sum (often denoted $W=X+Y$) is another random variable

$$(S_W, p_W)$$

such that

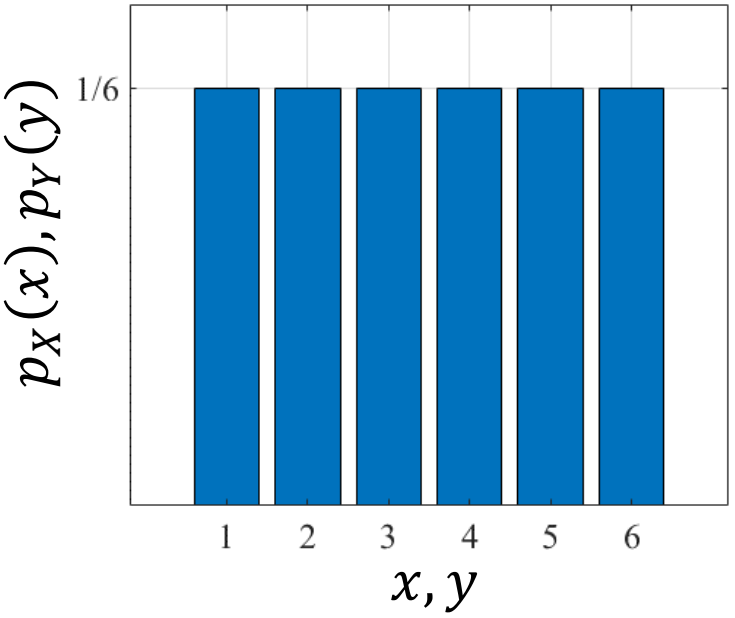
Here $$S_W = S_X + S_Y \stackrel{\text{def}}{=} \{ w = x + y \mid x \in S_X, y \in S_Y \}$$

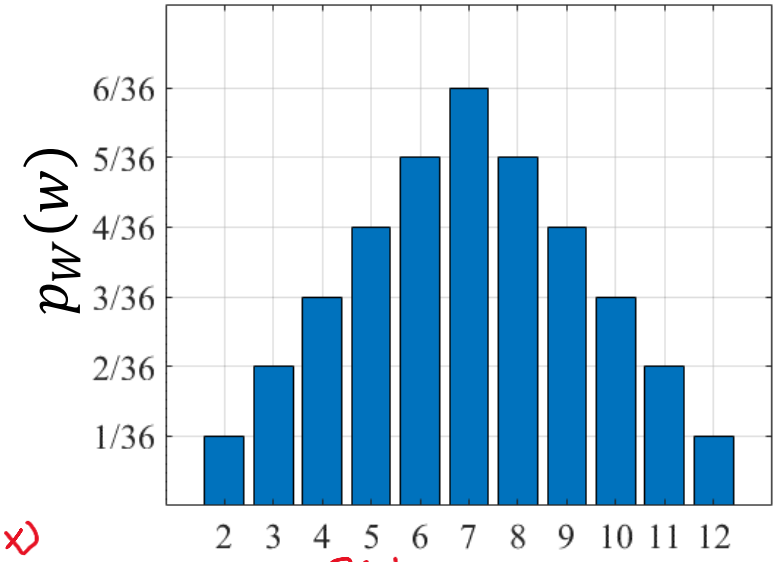
The probability function of the sum is given by the convolution of the probability function of two random variables, i.e.,:

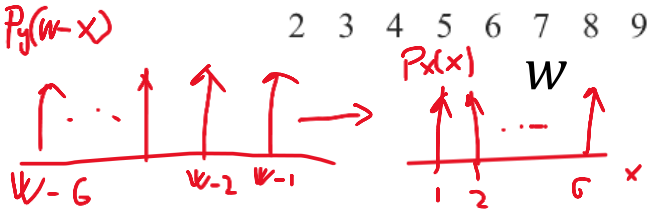
$$p_W(w) = \sum_{x \in S_X} p_X(x) \cdot p_Y(w - x) = \sum_{y \in S_Y} p_X(w - y) \cdot p_Y(y)$$

Example 8: Roll of two dice $W=(S_W, p_W)$

$X=(S_X, p_X)$ and $Y=(S_Y, p_Y)$: r.v. associated to the single rolls







$S_X = S_Y = \{1, 2, \ldots, 6\}$

$S_W = S_X + S_Y = \{2, 3, \ldots, 12\}$

$$p_W(w) = \sum_{x \in S_X} p_X(x) \cdot p_Y(w-x)$$

```
>> pX=ones(1,6)/6
>> pY=ones(1,6)/6
>> conv(pX,pY)
```

$p_W(2) = p_X(1) \cdot p_Y(1) = 1/6 \cdot 1/6 = 1/32$

$p_W(3) = p_X(1) \cdot p_Y(2) + p_X(2) \cdot p_Y(1)$
$= (1/6 \cdot 1/6) + (1/6 \cdot 1/6) = 2/36$

…

$p_W(12) = p_X(6) \cdot p_Y(6) = 1/6 \cdot 1/6 = 1/32$

Selected Discrete Random Variables

Non-parametric distribution $X = (S, p)$

A random variable with sample space $S = \{x_1, x_2, \ldots, x_n\}$ which can take one of the n values with arbitrary non-zero probability:

$$p(x_i) = Pr(X=x_i), \qquad i=1,2,\ldots,n.$$

It is called non-parametric when $p(x)$ cannot be written by a mathematical function which depends on given parameters.

To describe this random variable one must list all possible values of $p(x_i)$, e.g., giving a table.

Bernoulli distribution $X = (S, p) \sim \text{Ber}(\pi)$

Parameter: $\pi \in [0, 1]$

Sample space $S = \{0, 1\}$

Probability function : $p(k) = \Pr(X = k) = \begin{cases} 1 - \pi & \text{if } k = 0 \\ \pi & \text{if } k = 1 \end{cases}$

Interpretation:
It describes an experiment which has only two possible outcomes:
"success" (X=1) and "failure" (X=0).

These two outcomes are mutually exclusive and exhaustive events.

We denote π the probability of "success" .

Geometric distribution $X = (S, p) \sim \text{Geo}(\pi)$

Parameter: $\pi \in [0, 1]$

Sample space: $S = \mathbb{N} \quad = \{1, 2, 3, \ldots\}$

Probability function : $p(k) = \Pr(X = k) = \pi \cdot (1 - \pi)^{k-1}$

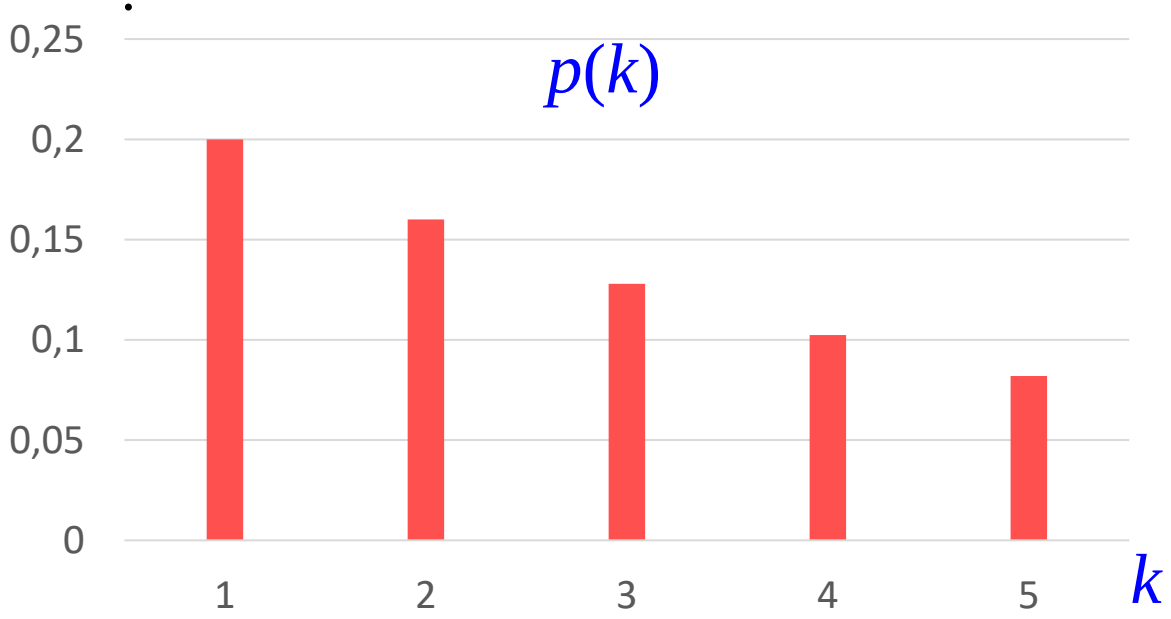
Interpretation: describes the number of independent Bernoulli trials (each with probability of success π) required until a first success.

Event $X=k$ occurs when we have "$k-1$" consecutive failures [prob $= (1 - \pi)^{k-1}$] and then one success [prob $= \pi$]

It is easy to proof that the complementary cumulative probability function of the geometric random variable is

$$Pr(X>k) = (1-\pi)^k, \quad k = 0,1,2,3,\ldots$$

k	$p(k) = \pi\,(1-\pi)^{k-1} \quad (\pi = 0.2)$
1	0.2
2	$0.2 \cdot 0.8 = 0.16$
3	$0.2 \cdot 0.8^2 = 0.128$
4	$0.2 \cdot 0.8^3 = 0.124$
5	$0.2 \cdot 0.8^4 = 0.082$
.	.



Memorylessness property

A discrete random variable X is memoryless if

$$Pr(X > m + n | X > m) = Pr(X > n), \qquad \forall m, n \in \mathbb{N}$$

A geometric random variable is memoryless because it is based on independent Bernoulli trials.

Therefore having m failures do not affect the probability that the next n trials will be failures as well.

Geometric random variables are the only discrete random variables that are memoryless.

Example:

A packet sent on a channel is correctly received 80% of the time.

a) Describe this transmission process with a random variable.

b) Which probability distribution describes the number of trials needed to correctly send a packet?

c) How many transmission k do we need to guarantee

$$\Pr(successfully\ receive\ the\ packet) > 0.98$$

Sol:

a) If 0 means «not correctly received» and 1:«correctly received» the single transmission is Ber(0.8), while the the number of trials is Geo(0.8)

b) This probability is equivalent to find

c) k such that $(:)$ $P_X(k) = Pr(X \leq k) = \sum_{i=1}^{k} 0.8^i 0.2^{i-1} \leq 0.98$

$$\Rightarrow k \geq 3 \Rightarrow P_X(3) = 0.8 \cdot 0.2^{1-1} + 0.8 \cdot 0.2^{2-1} + 0.8 \cdot 0.2^{3-1} = 0.992$$

Binomial distribution $X = (S, p) \sim \text{Bin}(n, \pi)$

Parameters: $n \in \mathbb{N}_{>0} = \{1, 2, 3, \ldots\}$; $\pi \in [0, 1]$

Sample space: $S = \{0, 1, 2, \ldots, n\}$

Probability function : $p(k) = \binom{n}{k} (1 - \pi)^{n-k} \cdot \pi^k$

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

binomial coeff.

Interpretation: describes the number of successes in n independent Bernoulli trials (each with probability of success π)

Note that $\text{Bin}(1, \pi) = \text{Ber}(\pi)$, i.e., a binomial random variable with n=1 reduces to a Bernoulli random variable.

Example:

A data-source alternates between two states "on", in which transmits streams at 10Mbps, and "off" which it is idle.

Let $\pi = 0.3$ be the **proportion of time** the source is active.

Assume we have n=10 independent identical sources

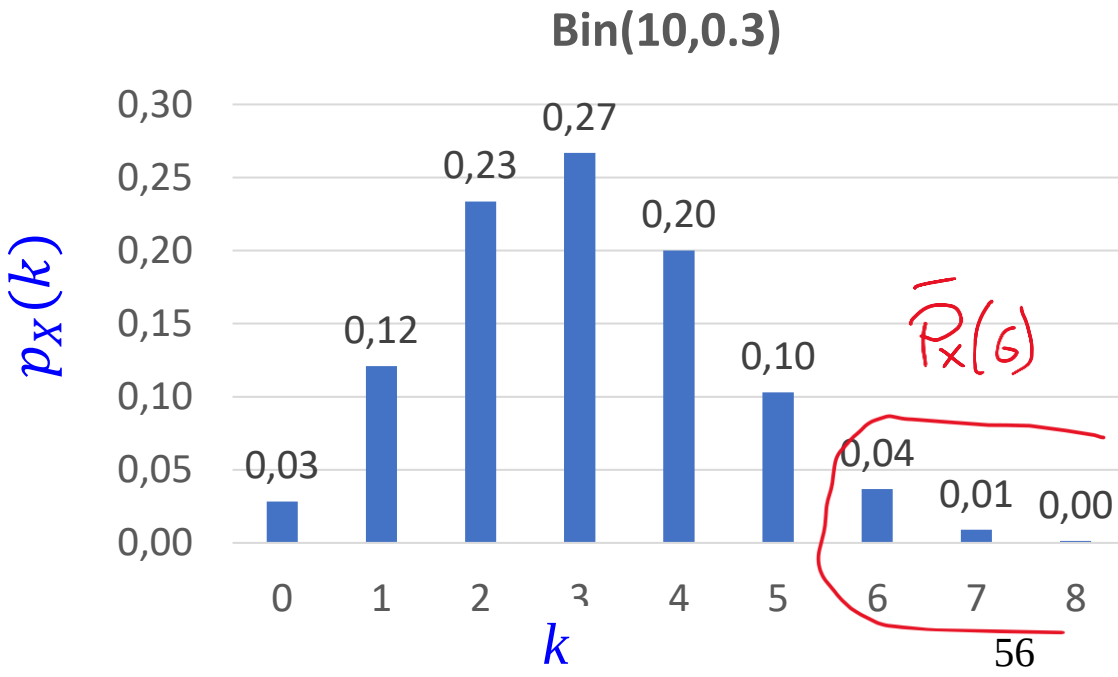
Let X be the r.v. counting the #sources simultaneously active.

Compute the proportion of time 6 or more, over 10 sources are "on"

Sol: X~Bin(10,0.3)

$$\bar{P}(6) = 1 - P_X(5) \approx 0.05 \equiv 5\%$$

This implies bandwidth of $50Mbps$ $(10M \cdot 5)$ is enough for the 95% of the time.

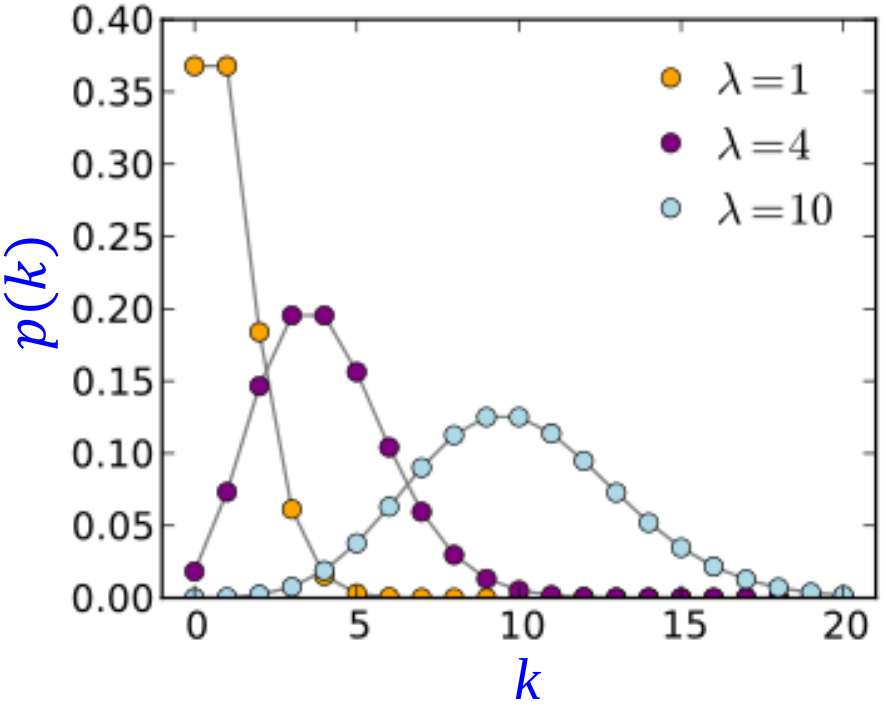


Poisson distribution $X = (S, p) \sim \text{Pois}(\lambda)$

Parameters: $\lambda \in \mathbb{R}_+$

Sample space: $S = \mathbb{N}_{\geq 0} = \{0, 1, 2, 3, \dots\}$

Probability function : $p(k) = e^{-\lambda} \frac{\lambda^k}{k!}$



Interpretation: Pois(λ) denotes the probability of having a certain number of independent events occurring in a given time interval T, knowing that the **average** number of events occurring in T is λ.

Application

A good model for the number of independent calls arriving at a telephone exchange or Internet service provider in a short period of time, a few seconds, minute, or hour.

It works well whenever the #customers n (or ***flows***) is **large**, and the **probability** π of a customer making a call is **small**.

Example

- Messages to be transmitted wait in a buffer before the transmission
- In 1 sec, on average, 3 messages arrive in the buffer.
- A transmitter, with a period of 1 sec, clears the buffer and send the messages using 5 channels (only one message per channel).
- If the buffer contains $k > 5$ messages, last k-5 messages are discarded

Example (cont'd)

Let the arrival process of messages in the buffer be our r.v. of interest.

a) How is it distributed?

b) Consider a time window of 10 seconds. Which is the proportion of time during which packets are discarded (i.e. **1** OR **more messages**)?

Sol:

a) The arrival process of of messages in the buffer is Pois(λ) = Pois(3)

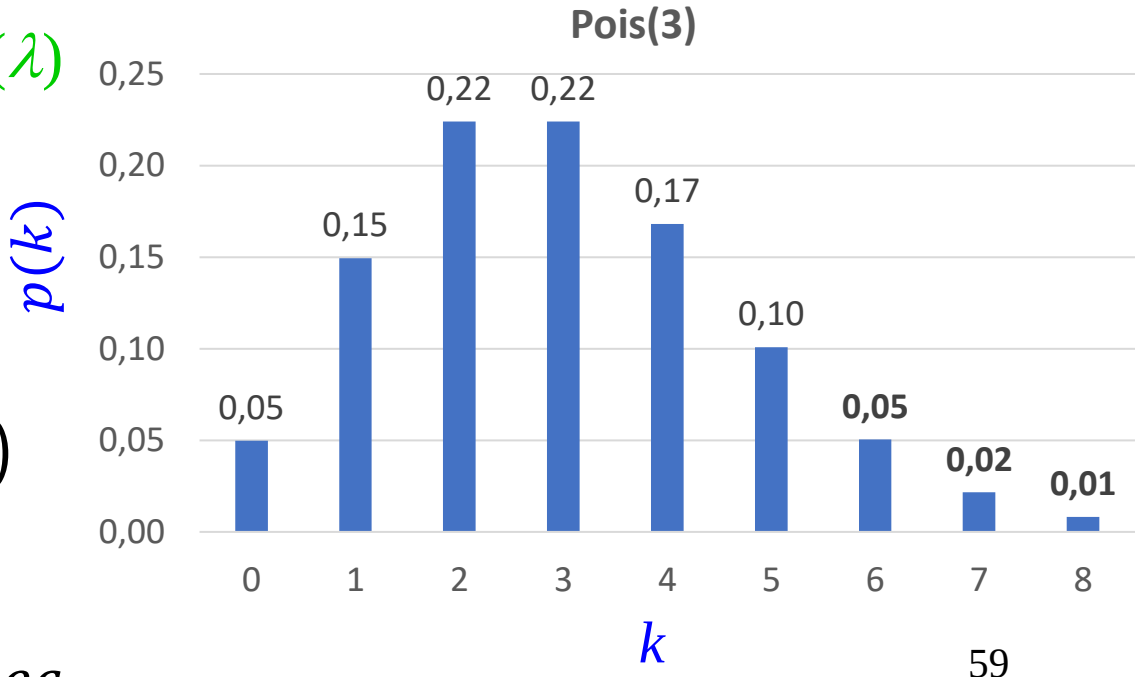
b) $\Pr(discard) \equiv \bar{P}(6)$

$= 1 - P(5)$

$= 1 - \big(p(1) + p(2) + \cdots p(5)\big)$

≈ 0.0839

c) time $= 10sec \cdot 0.0839 \approx 0.8sec$



On Matlab:

```
lambda=3                        % arrival rate : 3 mex/sec
px=poisspdf(0:8,lambda); % Compute the poisson pmf
Pr_disc=1-poisscdf(5,3)  % Compute the discharge Pr.

% Experiment:
exp_duration=2*3600 % 2 hours
arrivals=poissrnd(lambda,exp_duration,1);% Arr. for 2h.
mean_of_arrivals=mean(arrivals); % .. close to 3

Pr_disc_exp=sum(arrivals>5)/length(arrivals)
[Pr_disc, Pr_disc_exp]
```

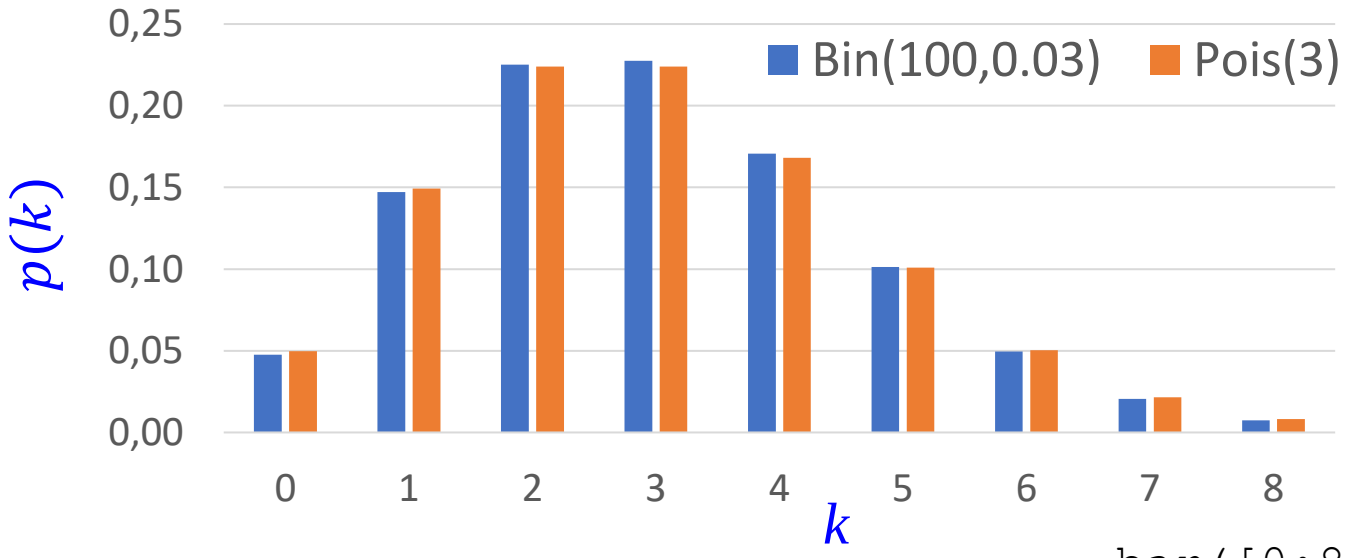
```
ans =

    0.0839    0.0872
```

Important note

Pois(λ) **well approximates** Bin(n,π) **whenever** $n \geq 100$ and $\lambda = n \cdot \pi \leq 10$

Ex: Consider Bin(n,π) = Bin(100,0.03)

Take $\lambda = n \cdot \pi = 3$ and compare with Pois(λ) = Pois(3)




```
bar([0:8],poisspdf(0:8,3)),
hold on
bar([0:8],binopdf(0:8,100,0.03))
```

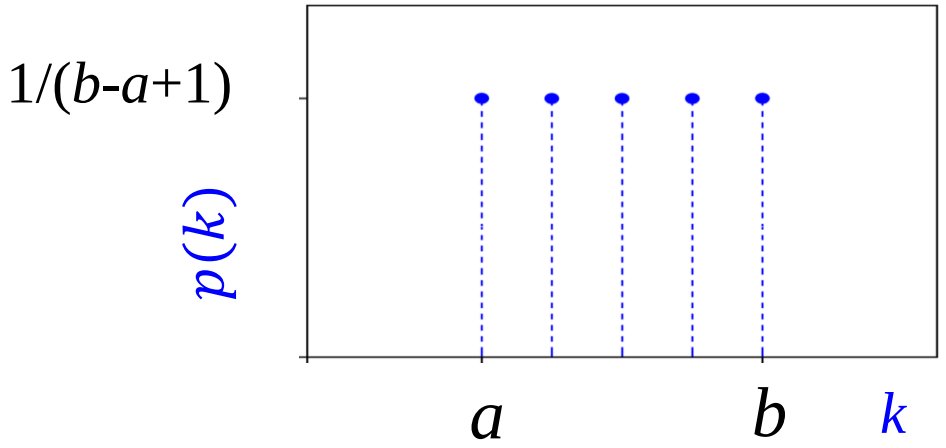
Discrete uniform distribution $X = (S, p) \sim \text{Uni}(a,b)$

Parameters: $a, b \in \mathbb{Z}, a < b$

Sample space: $S = \{a, a+1, \ldots, b\}$

Probability function : $p(k) = \frac{1}{b-a+1}$

Interpretation: It is a probability distribution describing a finite number of values which are equally likely to be observed



Example: the roll of a (fair) dice is a uniform r.v. Uni(1,6)

UNIVERSITY OF CAGLIARI

DIEE - Department of Electrical and Electronic Engineering

Disclaimer

ENGLISH

- **It is forbidden to copy, edit, and reproduce the contents and images of the lessons in any form**
- **It is forbidden to disseminate, redistribute and publish the contents and images and videos of this lecture in any way and means not expressly authorized by the author or by Unica**

ITALIANO

- **E' vietata la copia, la rielaborazione e riproduzione dei contenuti ed immagini presenti nelle lezioni in qualsiasi forma**
- **E' vietata la diffusione, la redistribuzione e pubblicazione dei contenuti, ed immagini, e registrazioni delle lezioni con qualsiasi modalità e mezzo non autorizzati espressamente dall'autore o da Unica**

Summary

Part 1 - Introduction to probability

Part 2 – Random variables

- Discrete random variables
- Continuous random variables
- Mean, variance and other moments

Continuous random variables

A continuous random variable X takes value in an uncountable set S_X.

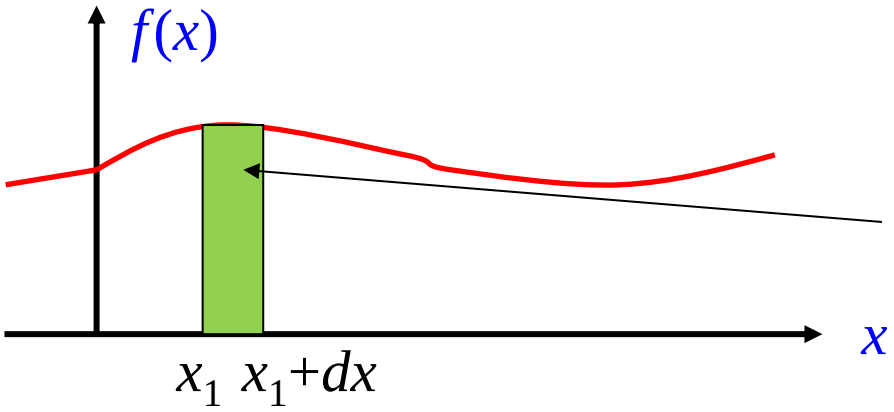
We consider a numerical sample space $S_X \subseteq \mathbb{R}$

A continuous random variable is denoted as

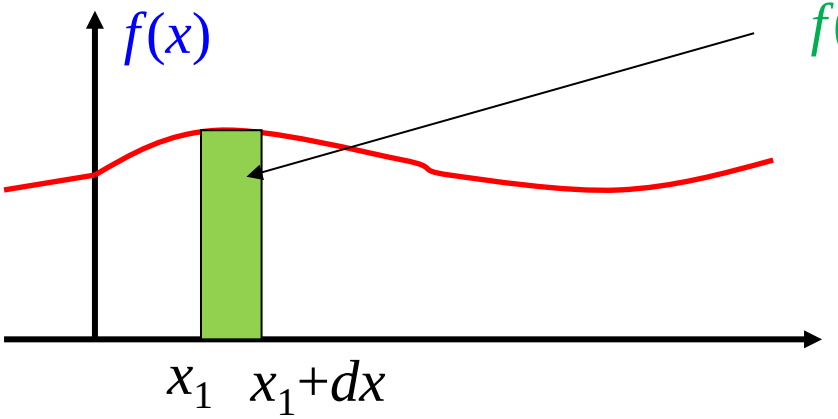
$$X = (S_X, f_X) \quad \text{where } f_X : S_X \to \mathbb{R}_{\geq 0}$$

is called probability density function (pdf) of X.

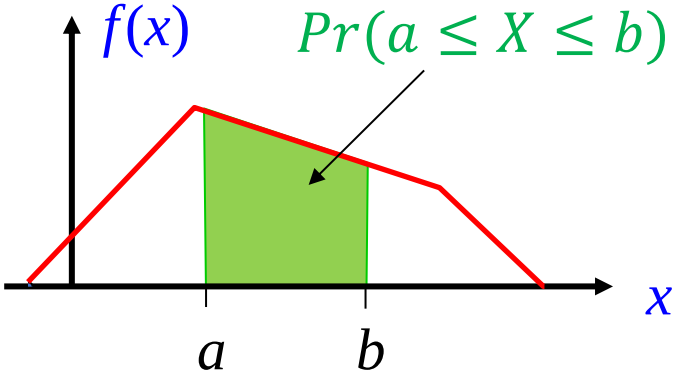
When clear, one may write (S, f) instead of (S_X, f_X).



$$f(x_1)\mathrm{d}x = Pr(X \in [x_1, x_1+dx])$$



$f(x_1)dx = Pr(X \in [x_1, x_1+dx])$



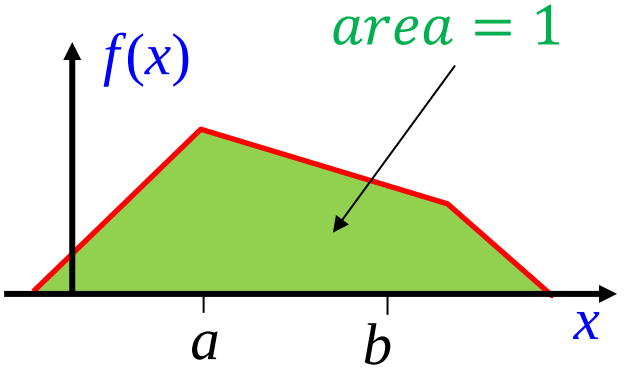
Clearly it holds:

a) $$Pr(a \le X \le b) = \int_a^b f(x)dx$$

b) $$Pr(X = a) = \int_a^a f(x)dx = 0$$

Since there are infinite many possible values to begin with, then $Pr(X = a)$=0

c) $$\int_{x \in S} f(x)dx = \int_{-\infty}^{+\infty} f(x)dx = 1$$



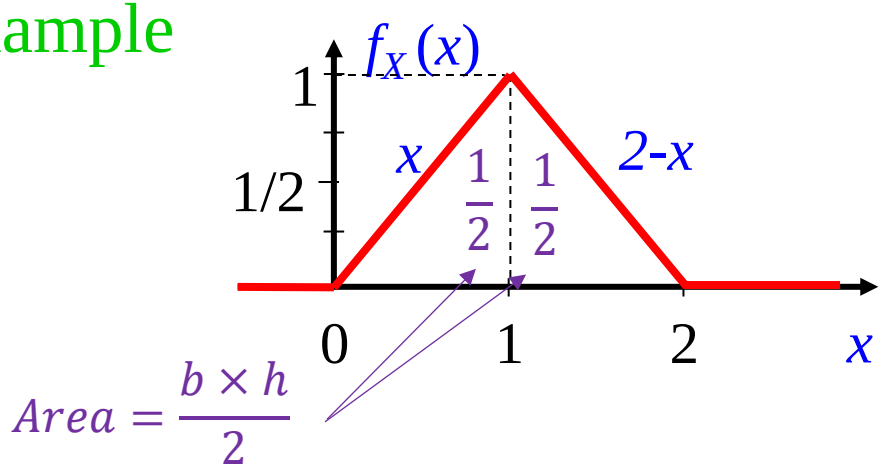
Cumulative distribution function of a continuous r.v. $X=(\mathbb{R}, f_X)$

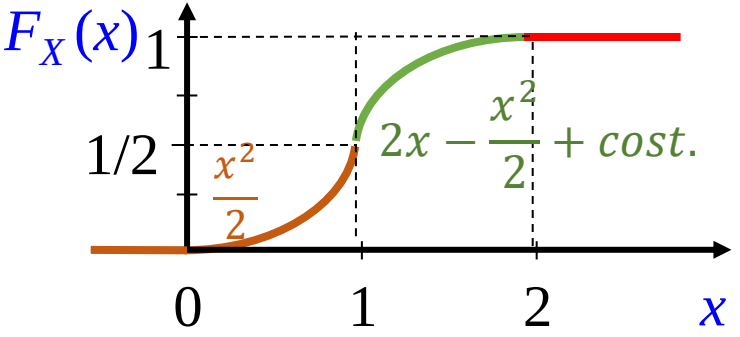
The (cumulative) distribution function $F_X(x)$ denotes the probability that X takes a value less than or equal to x:

$$F_X(x) = Pr(X \leq x) = \int_{-\infty}^{x} f_X(x)\,dx$$

Important: By differntiating $F_X(x)$ we can obtain the pdf $f_X(x) = \frac{dF_X(x)}{dx}$

Example





$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{2} & 0 < x < 1 \\ 2x - \frac{x^2}{2} + \text{const.} & 1 < x < 2 \\ 1 & x \geq 2 \end{cases} \quad \Longleftarrow \quad F_X(1) = \tfrac{1}{2} \implies \text{const.} = -1) \qquad \text{(for continuity)}$$

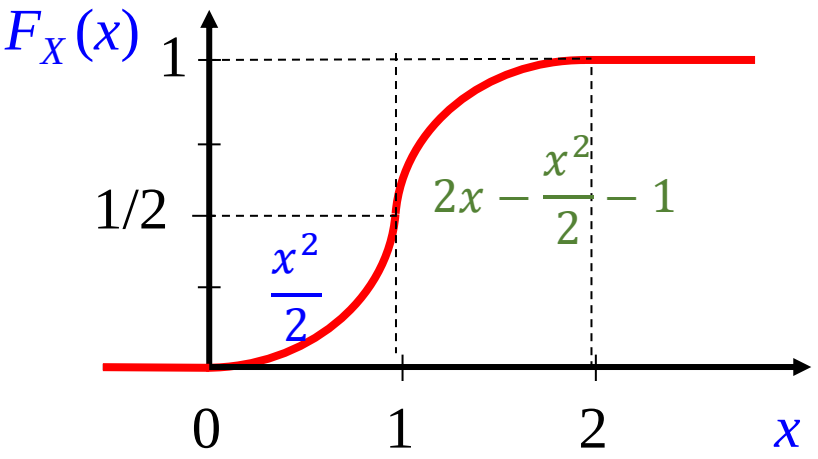
Complementary cumulative distribution function of a continuous r. v.

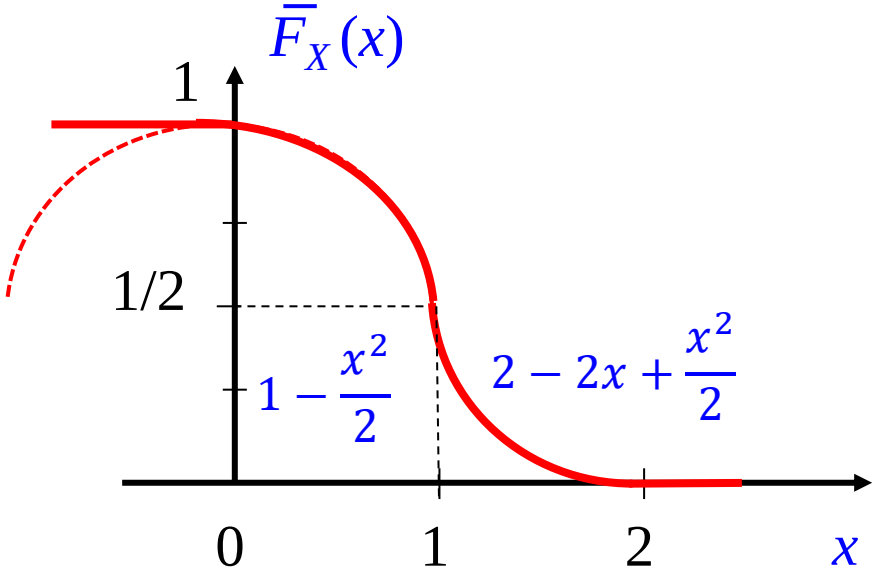
Denotes the probability that X takes a value greater than x:

$$\bar{F}_X(x) = Pr(X > x) = \int_x^{\infty} f_X(x)\,dx$$

Clearly: $\bar{F}_X(x) = 1 - F_X(x)$ (same information)

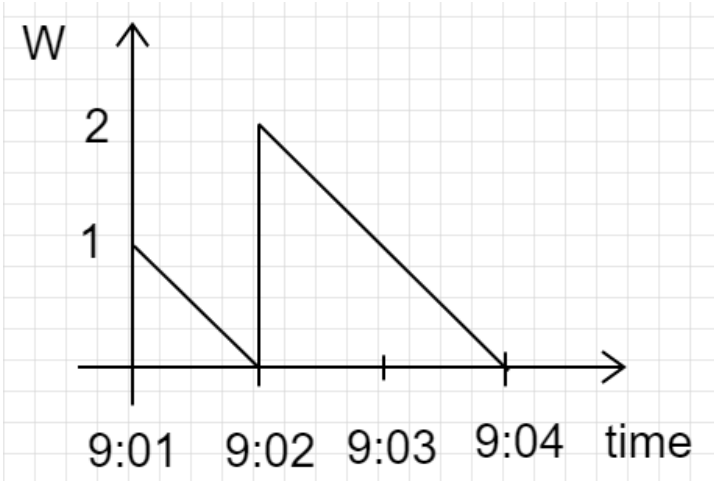
Example

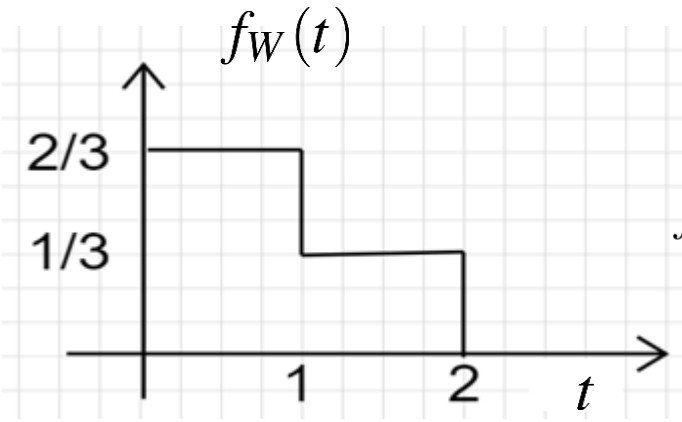




Example: A guy is used to go to work by train.

- He can catch two trains, that of the 9.02 or that of the 9.04 AM.
- He is used to arrive at the platform at between the 9.01 and 9.04
- The waiting time at the station is a r.v.: $W(S_w, f_W)$ with $S_W = \{t \in [0,2]\}$
- Calculate $f_W(t)$.



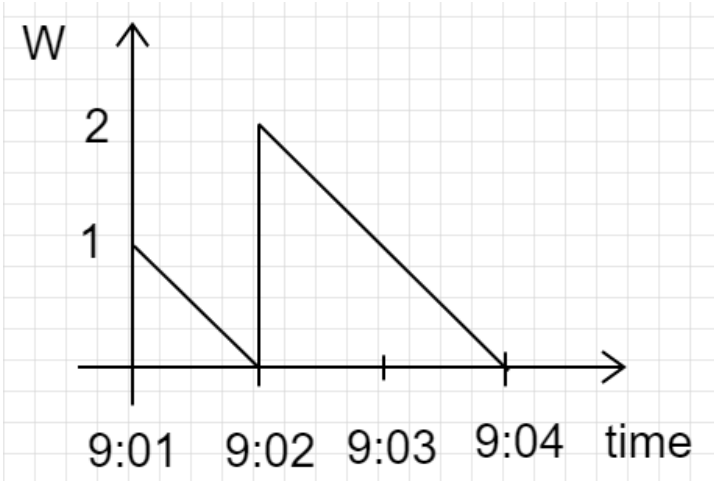


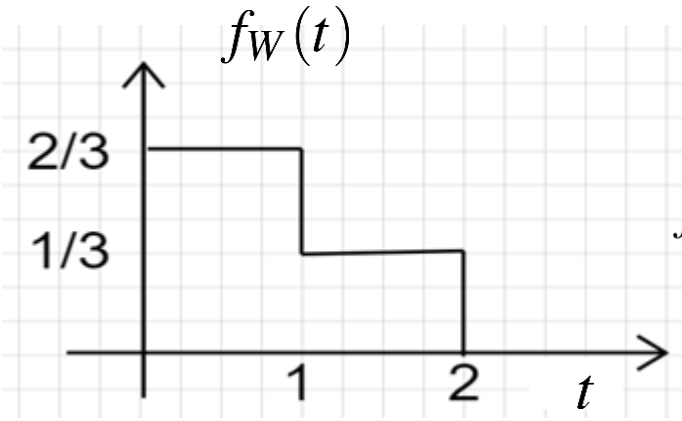
$$f_W(t) = \begin{cases} \frac{2}{3} & 0 \le t \le 1 \\ \frac{1}{3} & 1 \le t \le 2 \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow \begin{cases} \Pr(0 \le W \le t) = \frac{2t}{3} & 0 \le t \le 1 \Rightarrow F_W(1) = \frac{2}{3} \\ \Pr(1 \le W \le t) = \frac{t}{3} + cost & 1 \le t \le 2 \\ 0 & \text{otherwise} \end{cases} \Rightarrow F_W(t) = \Pr(W \le t) \begin{cases} 0 & t < 0 \\ \frac{2}{3}t & 0 \le t \le 1 \\ \frac{t}{3} + \frac{1}{3} & 1 \le t \le 2 \\ 1 & t > 2 \end{cases}$$

Example: A guy is used to go to work by train.

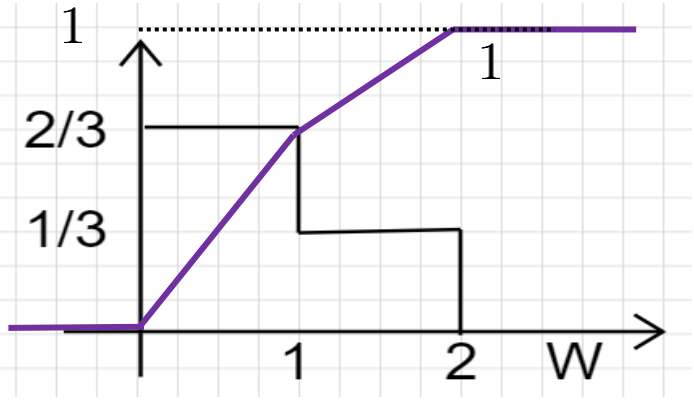
- He can catch two trains, that of the 9.02 or that of the 9.04 AM.
- He is used to arrive at the platform at between the 9.01 and 9.04
- The waiting time at the station is a r.v.: $W(S_w, f_W)$ with $S_W = \{t \in [0,2]\}$
- Calculate $f_W(t)$.





$$f_W(t) = \begin{cases} \frac{2}{3} & 0 \le t \le 1 \\ \frac{1}{3} & 1 \le t \le 2 \\ 0 & \text{otherwise} \end{cases}$$

$$F_W(t) = \begin{cases} 0 & t < 0 \\ \frac{2}{3}t & 0 < t < 1 \\ \frac{t}{3} + \frac{1}{3} & 1 < t < 2 \\ 1 & t \ge 2 \end{cases}$$



Joint PDF and CDF (continuous random var.)

Given n continuous r.v. $X_i=(S_i,f_i)$ for $i=1, \ldots, n$, their joint (cumulative) distribution function is defined as

$$F_{X_1,X_2,\cdots,X_n}: S_1 \times S_2 \times \cdots \times S_n \rightarrow [0,1]$$

$$F_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = Pr(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n)$$

Joint PDF

$$= \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f_{X_1,\ldots,X_n}(x_1, \ldots x_n) dx_1 \ldots dx_n$$

Given the joint PDF we can always compute the PDF or the CDF of each single r.v as follows (not viceversa). This operation is called marginalization.

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy \equiv \int_{-\infty}^{+\infty} f_{X|Y}(x|y) \cdot f_Y(y)dy$$

Marginal PDF of X

Conditional PDF

Marginal PDF of Y

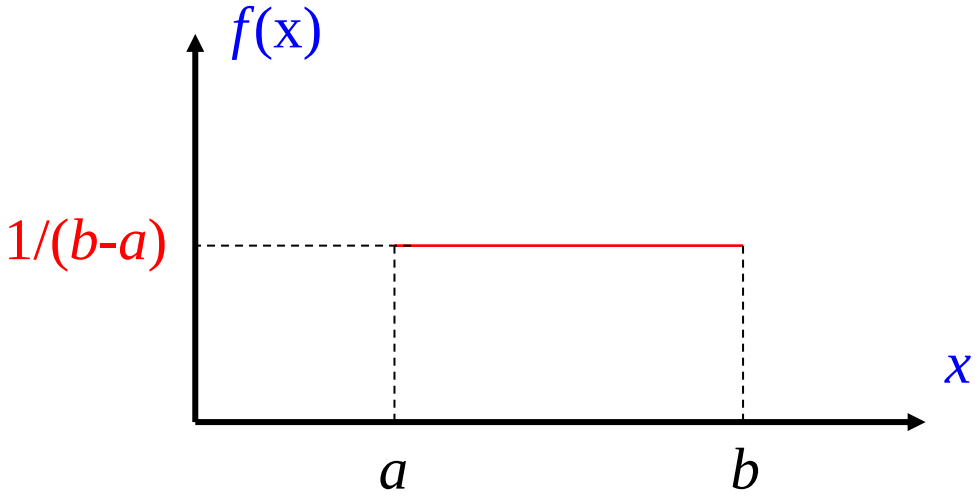
Selection of significant continuous random variables

Continuous uniform distribution $X = (S, f) \sim$ Uni_c(a,b)

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

Sample space: $S = [a, b]$

Probability density function : $f(x) = \frac{1}{b-a}$



Interpretation: The outcome is real value between a minimum (a) and a maximum (b). All possible values equally likely to occur.

Important property: For a **standard uniform distribution** it holds that

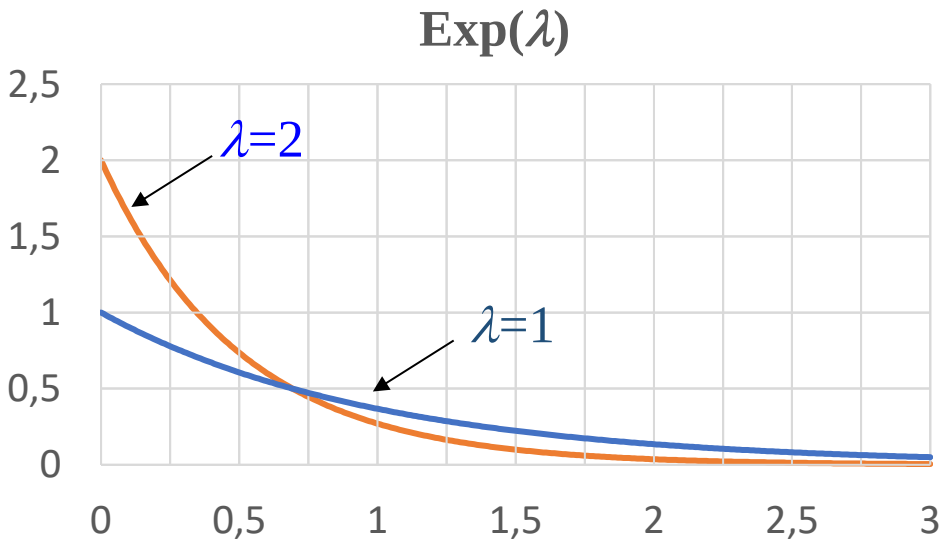
$$X \sim Uni\_c(0,1) \Rightarrow F_X(x) = x$$

Exponential distribution $X = (S, f) \sim \text{Exp}(\lambda)$

Parameter: $\lambda \in \mathbb{R}_{>0}$ (rate)

Sample space: $S = \mathbb{R}_{\geq 0}$

Probability density function: $f(x) = \lambda e^{-\lambda x}$



Feature: smaller values of X are more likely that higher values

Ex: Assume x is the cost of a purchase. One buys many cheap things (small x) but few expensive ones (large x)

Cumulative distribution function:

$$\Pr(X \leq x) = F_X(x) = \int_0^x \lambda e^{-\lambda s} ds = 1 - e^{-\lambda x} \quad , \quad \bar{F}_X(x) = e^{-\lambda x}$$

Interpretation as inter-event times

The exponential distribution is often used to model inter-event times, i.e., the time between the occurrence of two consecutive events (e.g., arrival of customers at a queue, arrival of calls at phone exchange, service times, etc).

Physical meaning of the rate parameter : $\lambda = \frac{\#events}{time}$

Example

A carwash on the average can serve 6 cars/hour.
Let the car service time be modelled by a r.v. $\boldsymbol{X \sim Exp(\lambda)}$

- Characterize the r.v. parameter and sample space
- What's the probability a client is served in less than x = 10 minutes?

Sol: $x \in [0, \infty)$ whereas λ = 6 cars/hour = 1 car/10 min

- The service time pdf is:

$$f(x) = \lambda e^{-\lambda x} = \frac{1}{10} e^{-\frac{x}{10}}$$

$$\Pr(X \leq 10) = \int_0^{10} \lambda e^{-\lambda s} ds = -e^{-\lambda s}|_0^{10} = 1 - e^{-1} = 0.6321$$

The exponential distribution is memoryless [MOST IMPORTANT]

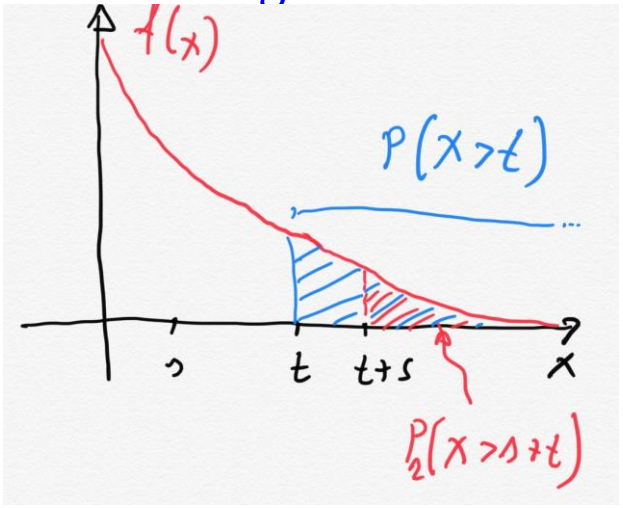
Let "s" be the time passed from the previous event (occurred at time "t")

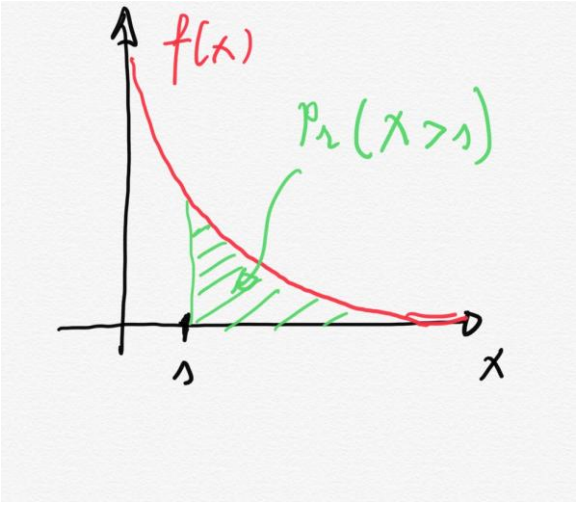
When inter-event times are exponentially distributed then events are independent, thus knowing how long has passed since the previous event occurrence does not give any additional info to better estimate when the next event will occur:

$$Pr\,(X>s+t \mid X>t) = Pr\,(X>s) \qquad \text{for all } t \geq 0,\ s \geq 0$$

The probability that next event will occur in less than s instants from now given that the previous event occurred more that t instants ago.

The probability that next event will occur in less than s instants after the previous event



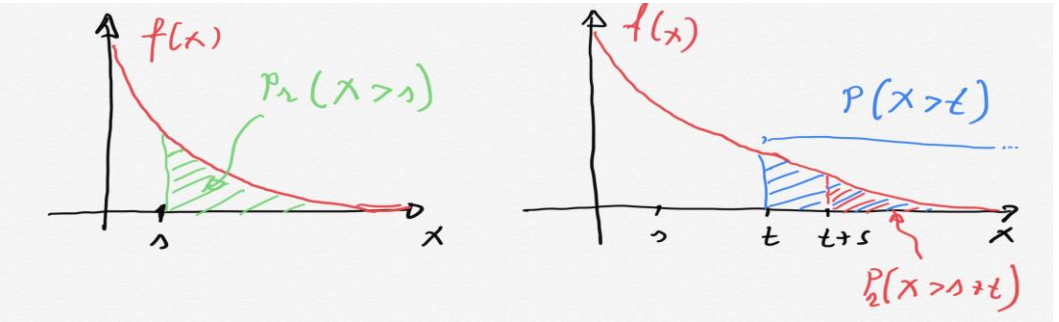


Proof of memoryless $Pr(X>s+t \mid X>t) = Pr(X>s)$

First note that in the exponential distribution for all x ≥ 0 holds:

CDF $$Pr(X \le x) = \int_0^x \lambda e^{-\lambda s} ds = e^{-\lambda s}\Big|_0^x = 1 - e^{-\lambda x}$$

Complementary CDF $$Pr(X > x) = 1 - F(x) = e^{-\lambda x}$$



Now:

$$(X > s+t) \cap (X > t) \equiv (X > s+t)$$

def comp. CDF

$$Pr(X>s+t \mid X>t) = \frac{Pr(X>s+t,\ X>t)}{Pr(X>t)} = \frac{Pr(X>s+t)}{Pr(X>t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = Pr(X>s)$$

thm conditional probability

Is human aging process memoryless?

NO! If our aging process were memoryless, the probability of reaching the age of 100 years given that one is 80 years old would be equal to the probability that a newborn baby reaches the age of 20.

Of course human aging is not memoryless.

On the other hand:

- the inter-arrivals time of server requests
- the inter-event time between calls at a telephone exchange
- Lifetime of batteries in devices, or of electronic components
- the intensity, duration, and number of rainfall/earthquakes
- Inter-session time in communication protocols
- Etc…

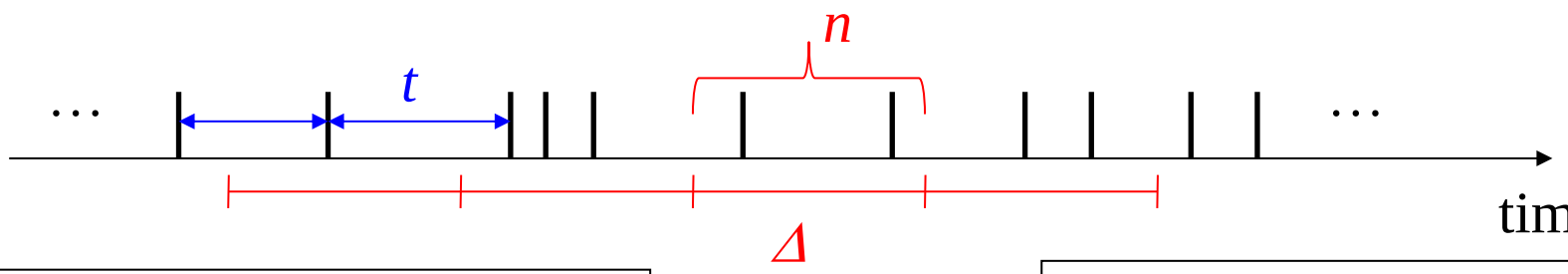
are well approximated by exponential random variable.

Note: the exponential random variable is the only memoryless continuous random variable.

Exponential VS Poisson (IMPORTANT!!!)

A Poisson distribution counts the number N of indipendent events of an arrival process with a given rate λ.

IMPORTANT: Since events are independent, the inter-arrival times are exponentially distributed with the same rate λ



The number of arrivals n in a unit of time is a discrete random variable

$N = \text{Pois}(\lambda)$

with pmf $p_N(n) = \frac{\lambda^n}{n!} e^{-\lambda}$

The time t between two consecutive arrivals is a continuous random variable

$X = \text{Exp}(\lambda)$

with pdf $f_X(t) = \lambda e^{-\lambda t}$

Scaling property: The number of arrivals in a generic time interval of length Δ is still a Poisson r.v. with rate $\lambda' = \Delta \cdot \lambda$

$N' \sim Pois(\lambda \cdot \Delta)$ thus with a pmf $p_{N'}(n) = e^{-\lambda \cdot \Delta} \frac{(\lambda \cdot \Delta)^n}{n!}$

Exponential VS Poisson (cont.)

- Proof: Let N = Pois(λ) be number of events occurred in a Poisson process

$$p_N(N(t)=n)=\Pr(N(t)=n)\begin{cases}\frac{(\lambda t)^n}{n!}\cdot e^{-\lambda t} & \lambda\in\mathbb{R}^+, n\in\mathbb{N}_0\\ 0 & \text{otherwise}\end{cases}$$

- Assume no events are occurred in $[0,t]$.
- Let T be the r.v. associated with the next arrival
- The probability of not observing an arrival in the future corresponds to

$$\Pr(T>t)=\Pr(N(t)=0)=\frac{(\lambda t)^0}{0!}e^{-\lambda t}=e^{-\lambda t}$$

$$\Rightarrow \quad F_T(t)=\Pr(T\le t)=1-e^{-\lambda t}$$

$$\Rightarrow \quad f_T(t)=\frac{dF_T(t)}{dt}=\lambda\cdot \mathrm{e}^{-\lambda t}$$

Thus $T\sim Exp(\lambda)$ and $1/\lambda$ is denotes the mean arrival time

Theorem: min of exponential random variables

Consider n IID (indipendent and identically distributed) random variables $X_1, X_2, \ldots, X_n$ such that

$$X_i \sim \text{Exp}(\lambda_i).$$

and let : $X = \min\{ X_1, X_2, \ldots, X_n \}$

Then: $X \sim \text{Exp}(\lambda)$ with $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$.

$$min\{X_1, X_2\} > t$$
$$\equiv$$
$$Neither\, X_1 nor\, X_2\, occur$$
$$X_1 > t\,, X_2 > t$$

Proof (case n=2) :

$$Pr(X > t) = Pr(min\{X_1, X_2\} > t) = Pr(X_1 > t, X_2 > t)$$

$$= e^{-\lambda_1 t} \cdot e^{-\lambda_2 t} = e^{-(\lambda_1+\lambda_2)t} = e^{-\lambda t}$$

X has complementary distribution $e^{-\lambda t}$ ⟷ $X \sim \text{Exp}(\lambda)$

Theorem: competition between Exponential r.v.

Consider *2* IID (indipendent and identically distributed) random variables X_1, X_2 such that

$$X_i \sim Exp(\lambda_i)$$

The probability X_1 occurs before X_2 is:

$$Pr(X_1 < X_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$$

Proof :

$$f_{X_1,X_2}(x_1, x_2) = \lambda_1 e^{-\lambda_1 x_1} \cdot \lambda_2 e^{-\lambda_2 x_2} \text{ (due to independence)}$$

We want X_2 may takes any value [0,∞] while X_1 takes any value [0, X_2]

$$Pr(X_1 < X_2) = \int_0^{\infty} \int_0^{x_2} \lambda_1 e^{-\lambda_1 x_1} \cdot \lambda_2 e^{-\lambda_2 x_2} \, dx_1 dx_2 = \int_0^{\infty} \lambda_2 e^{-\lambda_2 x_2} \overbrace{\left(\int_0^{x_2} \cdot \lambda_1 e^{-\lambda_1 x_1} \, dx_1 \right)}^{1-e^{-\lambda_1 x_2}} dx_2$$

$$= \int_0^{\infty} \lambda_2 e^{-\lambda_2 x_2} dx_2 - \int_0^{\infty} \lambda_2 e^{-(\lambda_1+\lambda_2)x_2} \, dx_2 = 1 - \frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

Example

Consider *3* IID (indipendent and identically distributed) random variables X_1, X_2 and X_3 such that

$$X_i \sim Exp(\lambda_i)$$

Calculate

$$Pr(X_1 < X_2 < X_3)$$

Although this probability could be computed by solving a triple integral there is a much easier way.

First note that

$$\Pr(X_1 < X_2 < X_3) \equiv \Pr(X_1 < \min\{X_2, X_3\} \cap X_2 < X_3)$$

Then, due to independence

$$= \Pr(X_1 < \min\{X_2, X_3\}) \cdot \Pr(X_2 < X_3)$$

Thus

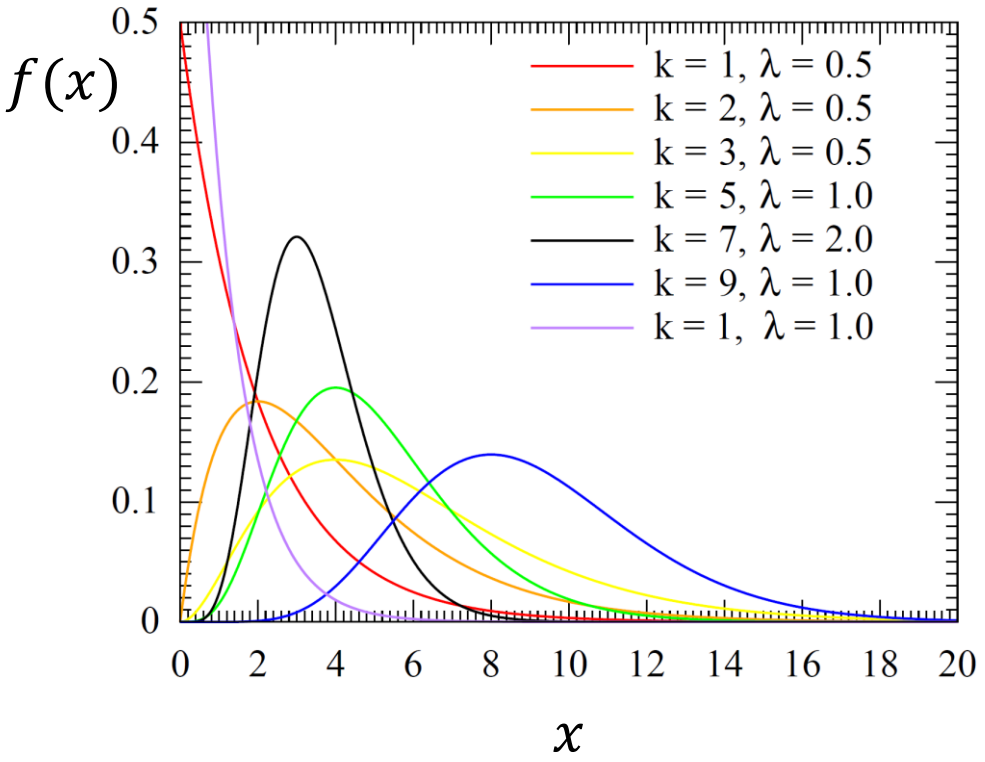
$$= \frac{\lambda_1}{\lambda_1 + (\lambda_2 + \lambda_3)} \cdot \frac{\lambda_2}{\lambda_2 + \lambda_3}$$

Erlang distribution $X = (S, f) \sim \mathbf{Erl}(k, \lambda)$

Parameters: $k \in \mathbb{N}_{>0}$ (shape); $\lambda \in \mathbb{R}_{>0}$ (rate)

Sample space: $S = \mathbb{R}_{\geq 0}$

Probability density function: $p(x) = \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x}$



Interpretation: generalizes an exponential distribution

Erl(1, λ) = Exp(λ)

i.e., Erlang with shape k=1 is an exponential

Erl(k, λ) describe the inter-event time associated to the occurrence of **k consecutive events**

Erlang as sum of exponential random variables

Let X_i, i=1,2,…,k, be k independent and identically distributed (IID) exponential random variables

$$X_i \sim \text{Exp}(\lambda)$$

Then, the random variable given by their sum $X = \sum_{i=1}^{k} X_i$

has an Erlang distribution $X \sim \text{Erl}(k,\lambda)$.

Ex 1. The time to perform an operation is a random variable

$$X \sim \text{Exp}(3.5).$$

Producing a high quality part requires 2 identical operations.

Thus the time to produce it is a random variable

$$\text{Erl}(2, 3.5).$$

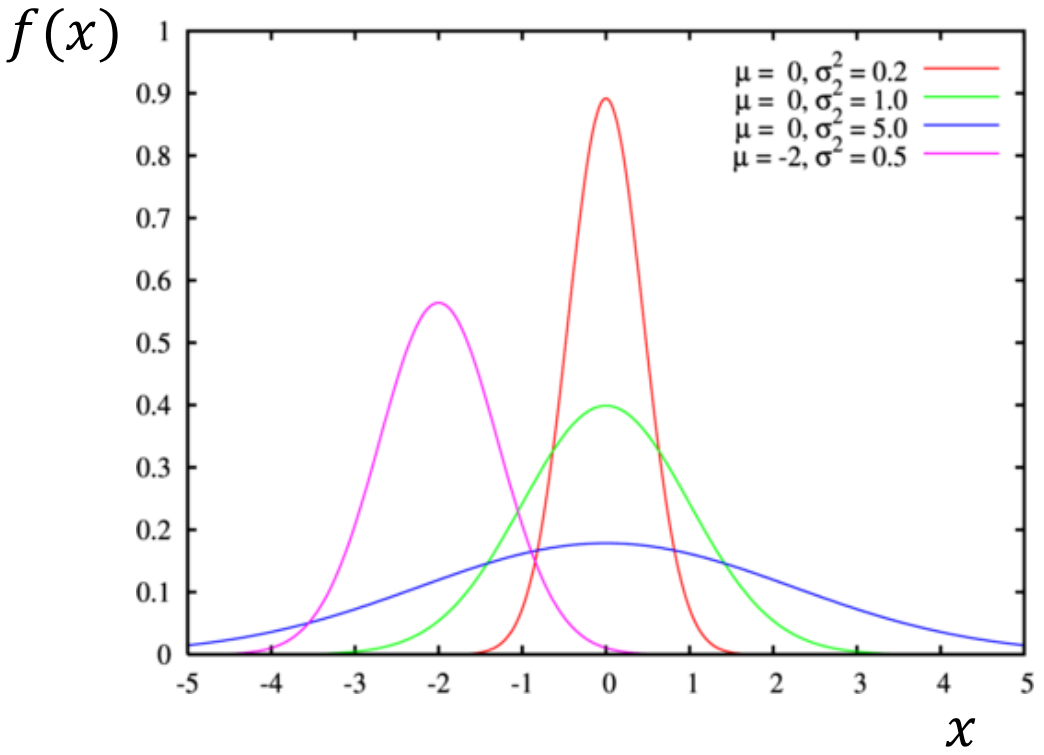
Ex 2. Packets arrive at Switch X with a rate λ=3.5 then, are regularly forwarded to Switch A and B alternatively.
The inter-arrival time of packets at each Switch is Erl(2, 3.5)

Normal (or Gaussian) distribution $X = (S, f) \sim N(\mu, \sigma^2)$

Parameters: $\mu \in \mathbb{R}$ (mean); $\sigma \in \mathbb{R}_{>0}$ (standard deviation)

Sample space: $S = \mathbb{R}$

Probability density function: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$

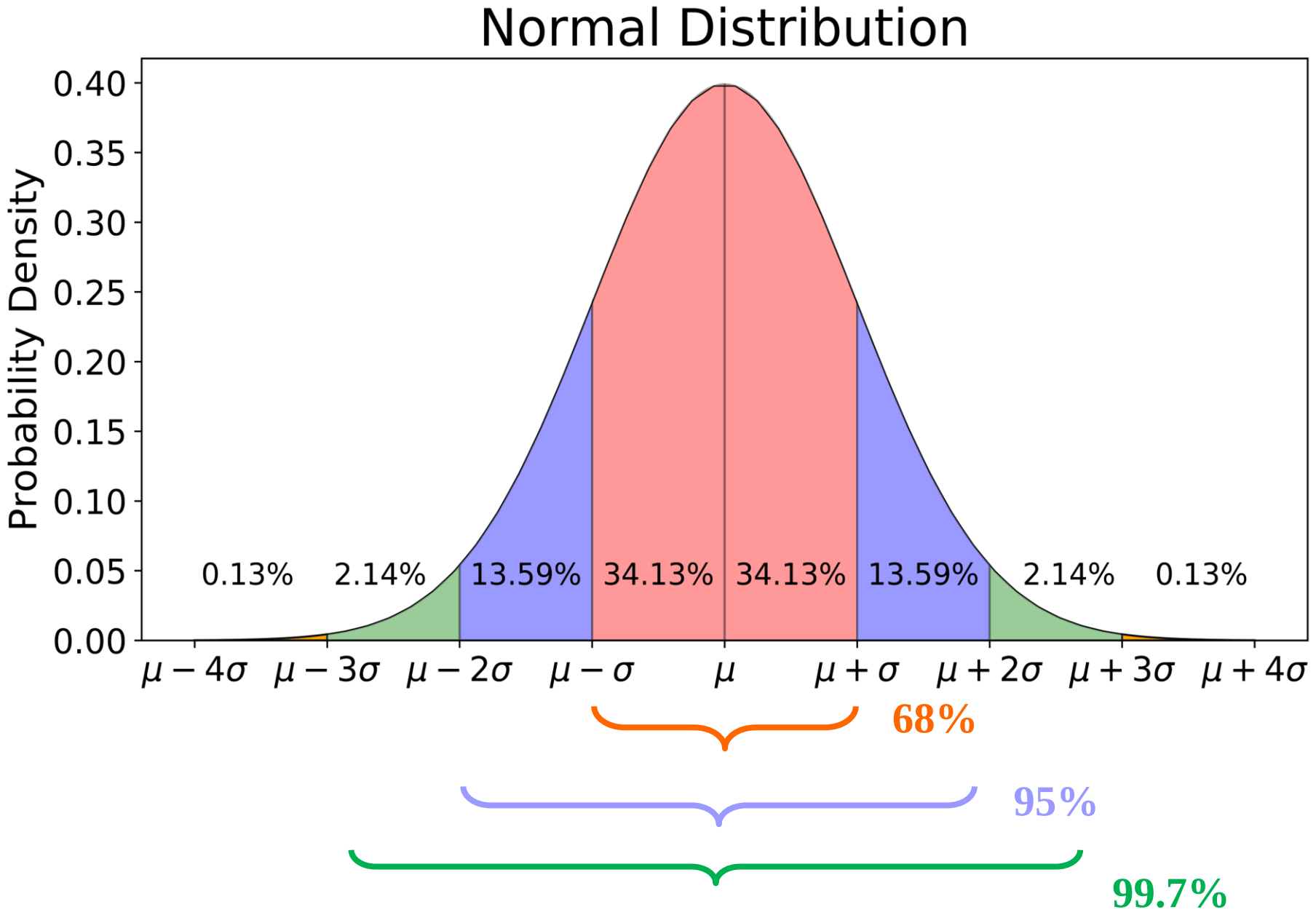


Features:

- Symmetric and bell shaped.
- Centered on mean μ
- Larger σ = flatter curve
- Smaller σ = sharper curve
- Standard normal curve

N(0, 1)

Dispersion as a function of standard deviation σ



Interpretation

Central limit theorem: The sum of a large number of independent random variables (under certain conditions) has a Gaussian distribution.

Remark: This is also true if the distribution of these random variables are very different from Gaussian.

This theorem explains why so many phenomena involving large populations (in nature and society) have bell shaped Gaussian histograms, and justifies the use of the Gaussian distribution as their model.

Generation of Random from $Uni\_c(0,1)$

Sometimes it is of interest the generate/study a r.v. as a function of other variables. To this end we can refer firstly introduce the next result:

Probability integral transform: Let X be a continuous r.v. with strictly increasing cdf $F_X(x)$. Then the r.v. $Y := F_X(X)$ has a standard normal uniform distribution, i.e. Y~Uni_c(0,1) and $F_Y(y) = y$

Proof: Let $y \in [0,1]$, assume for simplicity $\exists\, F_X^{-1}(y)$ (i.e. $F_X(x) = y$ is bijective thus strictly increasing) then:

$$
\begin{aligned}
F_Y(y) &= \Pr(\, Y \le y \,) \\
&= \Pr(\, F_X(X) \le y \,) \\
&= \Pr\left(\, X \le F_X^{-1}(y) \,\right) \quad \text{(this is the cdf of } X \text{)} \\
&= F_X(F_X^{-1}(y)) \\
&= y \qquad \Longrightarrow \qquad Y \sim \mathrm{Uni}_c(0,1)
\end{aligned}
$$

Generation of Random from $Uni\_c(0,1)$

Inverse Transform Sampling: Let $U \sim Uni\_c(0,1)$ and let X a target continuous r.v. with cdf $F_X(x)$, then $Y = F_X^{-1}(U)$ has the same distribution as X, i.e. $Y \sim X$.

It is just the inverse of the probability integral transform.

Proof. If the result is true, then $F_X(X) = U$. Thus, note that

$$\Longrightarrow \quad Y = F_X^{-1}(U) = F_X^{-1}(F_X(X)) \quad \Longrightarrow \quad Y = X$$

Example: Generate $X \sim Exp(\lambda)$ random numbers from $U \sim Uni\_c(0,1)$.

Let $X \sim Exp(\lambda)$ be the target r.v. and $U : \Pr(U \le u) = u$

$$F_X(X) = U \quad \Longrightarrow \quad 1 - e^{-\lambda Y} = U \quad \Longrightarrow \quad Y = F_X^{-1}(U) = \frac{ln(1-U)}{-\lambda}$$

$$F_Y(y) = \Pr\left(\overbrace{\frac{ln(1-U)}{-\lambda}}^{Y} \le y\right) = \Pr(1 - U \ge e^{-\lambda y})$$

In matlab:
u=rand(10^6,1); % U~Uni_c(0,1)
y = -lambda^-1*log(1-u);

$$= \Pr(U \le 1 - e^{-\lambda y}) = 1 - e^{-\lambda y} = F_X(y) \quad \Longrightarrow \quad Y \sim X$$

Summary

Part 1 - Introduction to probability

Part 2 – Random variables

- Discrete random variables
- Continuous random variables
- Mean, variance and other moments

90

Sample Mean and Sample Variance

Let us consider a sample consisting in n outcomes/realizations (or observations) of a r.v X, denoted by $X(i)$, i=1, …, n.

The sample mean (denoted by S_m) is often used to estimate $\mu_X = E[X]$

$$S_m = \frac{\sum_{i=1}^{n} X(i)}{n}$$

Similarly, the sample variance is used to estimate Var[X].

$$S_v = \frac{\sum_{i=1}^{n} [X(i) - S_m]^2}{n-1}$$

Mean (or expected value)

Discrete r.v. $X = (S, p)$ has mean

$$E[X] = \mu_X = \sum_{k \in S} k \cdot p(k)$$

Remark: In general it may happen that $E[X] \notin S$

Continuous r. v. $X = (\mathbb{R}, f\ )$ has mean

$$E[X] = \mu_X = \int_{-\infty}^{+\infty} x \cdot f(x) dx$$

In both cases the mean is the baricenter of the distribution

Interpretation/Motivating example (discrete case)

The height of person in centimeters is a r.v. $X = (S, p)$ where

- $S = \{50, \ldots, 250\}$: sample space in centimeters
- $p(k)$: probability that a person's height is k cm (for all $k \in S$).

Consider a sample of N people each with a height h_i.

Let n_k be the #people with a height of k cm.

Consider the so-called **sample mean:** $$S_m = \frac{1}{N}\sum_{i=1}^{N} h_i$$

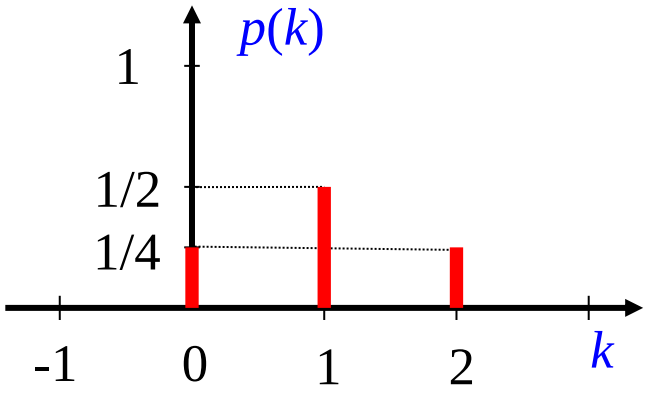
Then note that $$\frac{1}{N}\sum_{i=1}^{N} h_i = \frac{1}{N}\sum_{k\in S} k \cdot n_k = \sum_{k\in S} k \cdot \frac{n_k}{N}$$

As $N \rightarrow \infty$ the frequency n_k/N goes to $p(k)$ and we get

$$E[X] = \mu_X = \lim_{N\to\infty} \sum_{k\in S} k \cdot \frac{n_k}{N} = \sum_{k\in S} k \cdot p(k)$$

Example: number of heads flipping 2 coins $X = (S,p)$

s	$Pr(s)$	x	$p(x)$
TT	1/4	0	1/4
TH	1/4	1	1/2
HT	1/4		
HH	1/4	2	1/4



Mean

$$\mu_X = E[X] = \sum_{x \in S} k \cdot p(k) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

Remark: Other way to compute it? Yes: see moment generating function

Example: How much should a lottery ticket really cost to be fair?

Biglietti venduti: 17 949 331	Costo del biglietto: 3 €
Premi (in €)	
s_1	5 mln
s_2	2 mln
s_3	1 mln
s_4	800 000
s_5	700 000
s_6	600 000
s_7	500 000
s_8	400 000
s_9	300 000
s_{10}	200 000
da s_{11} a s_{40}	100 000
da s_{41} a s_{90}	50 000
Montepremi totale: $s = s_1 + s_2 + \cdots + s_{90} = 17$ mln	

Let $X(S, p_X)$ with $S \in \{s_1, s_2 \dots s_{90} \dots s_{1794431}\}$ be the r.v. associated to my winning (because I bought 1 ticket), my **expected winning** is

$$E[X] = \sum_{k=1}^{1794331} s_k \cdot q = q \sum_{k=1}^{90} s_k = q \cdot s \approx 0.95€$$

with $$q = \frac{1}{17949331} \approx 5 \times 10^{-8}$$

0.95€ is the fair price but instead…

Organizer's profit per ticket $\approx 3 - 0.95 \approx 2.05€$

Organizer's gross profit $\approx 17949331 \cdot 2.05\text{M€} \approx 36.8\text{M€}$

Organizer's net profit $\approx 36.8\text{M€} - 17\text{M€} = 19.8\text{M€}$

Lottery: A tax on people who are bad at math!

Moments (of *n*-th order)

Discrete r.v. $X = (S, p)$ has *n*-th order moment

$$E[X^n] = \sum_{x_i \in S} x_i^n \cdot p(x_i)$$

Continuous r. v. $X = (\mathbb{R}, f)$ has *n*-th order moment

$$E[X^n] = \int_{-\infty}^{+\infty} x^n \cdot f(x) dx$$

In both cases:

- The Moment of order 0 is always equal to 1 (normalized total mass)
- Moment of order 1 is the mean (center of mass)

Example: number of heads flipping 2 coins $X = (S,p)$

k	$p(k)$
0	1/4
1	1/2
2	1/4

Moments:

$$E[X^0] = \sum_{k \in S} k^0 \cdot p(k) = \sum_{k \in S} p(k) = 1$$

$$E[X^1] = \sum_{k \in S} k^1 \cdot p(k) = \sum_{k \in S} k \cdot p(k) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1 \equiv \mu_X$$

$$E[X^2] = \sum_{k \in S} k^2 \cdot p(k) = 0^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{1}{2} + 2^2 \cdot \frac{1}{4} = 1.5$$

$$E[X^3] = \sum_{k \in S} k^3 \cdot p(k) = 0^3 \cdot \frac{1}{4} + 1^3 \cdot \frac{1}{2} + 2^3 \cdot \frac{1}{4} = 2.5$$

…

Remark: Other way to compute it? Yes: see moment generating function

Central moments (of *n*-th order)

Discrete r.v. $X = (S, p)$ has *n*-th order central moment

$$E[(X - \mu_X)^n] = \sum_{k \in S} (k - \mu_X)^n \cdot p(k)$$

Continuous r. v. $X = (\mathbb{R}, f)$ has *n*-th order central moment

$$E[(X - \mu_X)^n] = \int_{-\infty}^{+\infty} (x - \mu_X)^n \cdot f(x) dx$$

In both cases:

- Central moment of order 0 is always 1
- Central moment of order 1 is always 0
- Central moment of order 2 called **Variance** (info on data's dispersion)
- Central moment of order 3 called **Skewness** (info on data's asymetricity)

Example: number of heads flipping 2 coins $X = (S,p)$

k	$p(k)$
0	1/4
1	1/2
2	1/4

Central moments (here $\mu_X = 1$):

$$E[(X-\mu_X)^0] = \sum_{k\in S}(k-\mu_X)^0 \cdot p(k) = \sum_{k\in S} p(k) = 1$$

$$E[(X-\mu_X)^1] = \sum_{k\in S}(k-\mu_X)^1 \cdot p(k) = \sum_{k\in S} k\cdot p(k) - \mu_X \sum_{k\in S} p(k) = \mu_X - \mu_X = 0$$

$$E[(X-\mu_X)^2] = \sum_{k\in S}(k-\mu_X)^2 \cdot p(k) = (0-1)^2\cdot\frac{1}{4} + (1-1)^2\cdot\frac{1}{2} + (2-1)^2\cdot\frac{1}{4} = 0.5$$

$$E[(X-\mu_X)^3] = \sum_{k\in S}(k-\mu_X)^3 \cdot p(k) = (0-1)^3\cdot\frac{1}{4} + (1-1)^3\cdot\frac{1}{2} + (2-1)^3\cdot\frac{1}{4} = 0$$

…

Remark: Other way to compute it? Yes: see moment generating function

Variance

The second central moment is called the variance.

Discrete r.v. $X = (S, p)$ has variance

$$Var[X] = \sigma^2 = E[(X - \mu_X)^2] = \sum_{x_i \in S} (x_i - \mu_X)^2 \cdot p(x_i)$$

Continuous r. v. $X = (\mathbb{R}, f)$ has variance

$$Var[X] = \sigma^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 \cdot f(x) dx$$

Standard deviation $\sigma = \sqrt{Var[X]}$

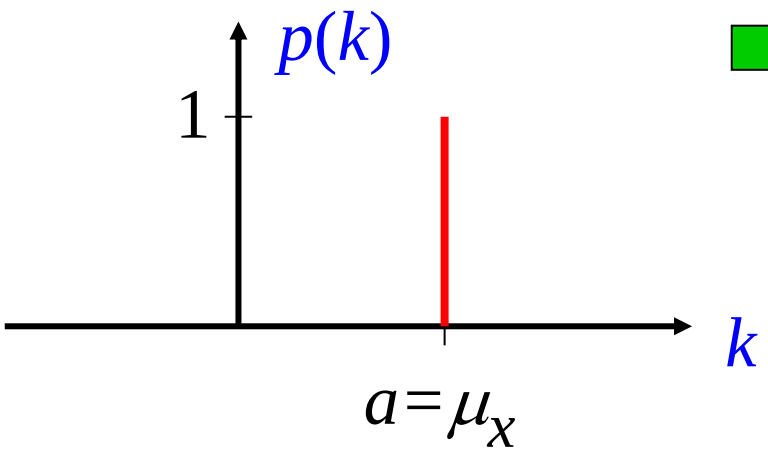
It gives a measure of the dispersion w.r.t the mean

The variance measures the statistical dispersion of a data, i.e., how far a set of (random) numbers are spread out from the its mean value.

It follows that, deterministic distributions (since have only one value with probability 1) have null variance.

Indeed:

k	$p(k)$
a	1



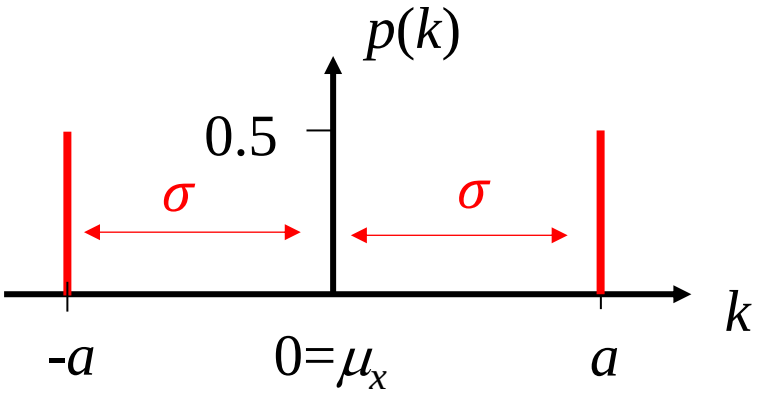
$$\mu_X = E[X] = \sum_{k \in S} k \cdot p(k) = a \cdot 1 = a$$

$$Var[X] = \sum_{k \in S} (k - \mu_X)^2 \cdot p(k) = 0 \cdot 1 = 0$$

Variance can be arbitrarily large

For instance, a distribution with only two values: $S = \{-a, a\}$:

k	$p(k)$
$-a$	0.5
a	0.5



$$\mu_X = E[X] = \sum_{k \in S} k \cdot p(k) = (-a) \cdot \frac{1}{2} + a \cdot \frac{1}{2} = 0$$

$$Var[X] = \sum_{k \in S} (k - \mu_X)^2 \cdot p(k) = (-a)^2 \cdot \frac{1}{2} + a^2 \cdot \frac{1}{2} = a^2$$

$$\sigma = \sqrt{Var[X]} = a$$

As a increases, so does the variance

Some useful results (algebra of random variables)

The mean of the sum of n (arbitrary) r.v. is the sum of the means:

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

Due to the linearity of both the summation operator and the mean operator

The variance of the sum of n independent r. v. is the sum of the variances:

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

The mean of the product of n independent r. v. is the product of the means:

$$E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i]$$

because of independence

$$p_{\prod_i X_i}\left(\prod_i x_i\right) = \prod_i p_{X_i}(x_i)$$

Fundamental result on variance

$$Var[X] = E[X^2] - \mu_X^2$$

Proof:

Mean of sum

$$Var[X] = E[(X - \mu_X)^2] = E[X^2 - 2X\mu_X + \mu_X^2]$$

$$= E[X^2] - 2\mu_X E[X] + E[\mu_X^2] = E[X^2] - 2\mu_X^2 + \mu_X^2$$

$$= E[X^2] - \mu_X^2$$

Sum of means

Shows the relationship between 2*nd* order central moment (variance) and 2*nd* order moment ($E[X^2]$)

Bernoulli distribution $X = (S, p) \sim \text{Ber}(\pi)$

Parameter: $\pi \in [0, 1]$

Sample space $S = \{0, 1\}$

Probability function : $p(k) = \begin{cases} 1 - \pi & \text{if } k = 0 \\ \pi & \text{if } k = 1 \end{cases}$

$$\mu_X = \pi$$

$$Var[X] = \pi \cdot (1 - \pi)$$

$$\mu_X = E[X] = \sum_{k \in S} k \cdot p(k) = 0 \cdot (1 - \pi) + 1 \cdot \pi = \pi$$

$$Var[X] = E[(X - \mu_X)^2] = \sum_{k \in S} (k - \mu_X)^2 \cdot p(k)$$

$$= (0 - \pi)^2 \cdot (1 - \pi) + (1 - \pi)^2 \cdot \pi = \pi \cdot (1 - \pi) \cdot [\pi + 1 - \pi]$$

$$= \pi \cdot (1 - \pi)$$

Geometric distribution $X = (S, p) \sim \text{Geo}(\pi)$

Parameter: $\pi \in [0, 1]$

Sample space: $S = \mathbb{N}_+ = \{1, 2, 3, \ldots\}$

Probability function : $p(k) = \pi \cdot (1 - \pi)^{k-1}$

$$\mu_X = \frac{1}{\pi}$$

$$Var[X] = \frac{1 - \pi}{\pi^2}$$

$$\mu_X = E[X] = \sum_{k=1}^{\infty} k \cdot p(k) = \sum_{k=1}^{\infty} k \cdot \pi \cdot (1 - \pi)^{k-1}$$

$$\sum_{k=1}^{\infty} k \cdot a^k = \frac{a}{(1 - a)^2}$$

$$= \pi \cdot \sum_{k=1}^{\infty} k \cdot (1 - \pi)^{k-1} = \frac{\pi}{1 - \pi} \cdot \sum_{k=1}^{\infty} k \cdot (1 - \pi)^k$$

$$= \frac{\pi}{1 - \pi} \cdot \frac{1 - \pi}{\pi^2} = \frac{1}{\pi}$$

Ex: if $\pi = 0.5 \rightarrow E[X] = 2$
i.e. I need 2 trial on average to get a success

$Var[X]$: prove it using moment generating function (see later)

Binomial distribution $X = (S, p) \sim \text{Bin}(n, \pi)$

Parameters: $n \in \mathbb{N}_+ = \{1, 2, 3, \ldots\}$; $\pi \in [0, 1]$

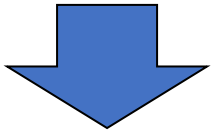
Sample space: $S = \{0, 1, 2, 3, \ldots, n\}$

Probability function : $p(k) = \binom{n}{k} (1 - \pi)^{n-k} \cdot \pi^k$

$$\mu_X = n \cdot \pi$$

$$Var[X] = n \cdot \pi \cdot (1 - \pi)$$

The binomial random variable **Bin(n, π)** is a sum of n independent identically distributed Bernoulli random variables **Ber(π)**



Mean and variance of **Bin(n, π)** are n times the mean and variance of **Ber(π)**

Poisson distribution $X = (S, p) \sim \text{Pois}(\lambda)$

Parameters: $\lambda \in \mathbb{R}_+$

Sample space: $S = \mathbb{N} = \{0, 1, 2, 3, \dots\}$

Probability function : $p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$

$$\mu_X = \lambda$$

$$Var[X] = \lambda$$

$$\mu_X = E[X] = \sum_{k=0}^{\infty} k \cdot p(k) = e^{-\lambda} \cdot \sum_{k=1}^{\infty} k \cdot \frac{\lambda^k}{k!}$$

$$= e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = e^{-\lambda} \cdot \lambda \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda} \cdot \lambda \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

$$= e^{-\lambda} \cdot \lambda \cdot e^{\lambda} = \lambda$$

Exponential Taylor expansion $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$

$Var[X]$: prove it using moment generating function (see later)

Discrete uniform distribution X = (S, p) ~ Uni(a,b)

Parameters: $a, b \in \mathbb{Z}, a < b$

Sample space: $S = \{a, a+1, \dots, b\}$

Probability function : $p(k) = \frac{1}{b-a+1}$

$$\mu_X = \frac{a+b}{2}$$

$$Var[X] = \frac{(b-a)^2}{12}$$

$$\mu_X = E[X] = \sum_{k=a}^{b} k \cdot p(k) = \frac{1}{b-a+1} \cdot \sum_{k=a}^{b} k$$

$$\sum_{k=0}^{n} k = \frac{n \cdot (n+1)}{2}$$

$$= \frac{1}{b-a+1} \cdot (a + (a+1) + \cdots + (a+b-a-1) + (a+b-a))$$

$$= \frac{1}{b-a+1} \cdot (\overbrace{a + a + \cdots + a}^{(b-a+1) \text{ times}} + \overbrace{0 + 1 + \cdots + (b-a)}^{})$$

$$= \frac{1}{b-a+1} \cdot \left(a(b-a+1) + \frac{(b-a)(b-a+1)}{2} \right) = a + \frac{b-a}{2} = \frac{a+b}{2}$$

Var[X] : prove it using moment generating function (see later)

Continuous uniform distribution $X = (S, f) \sim \text{Uni\_c}(a,b)$

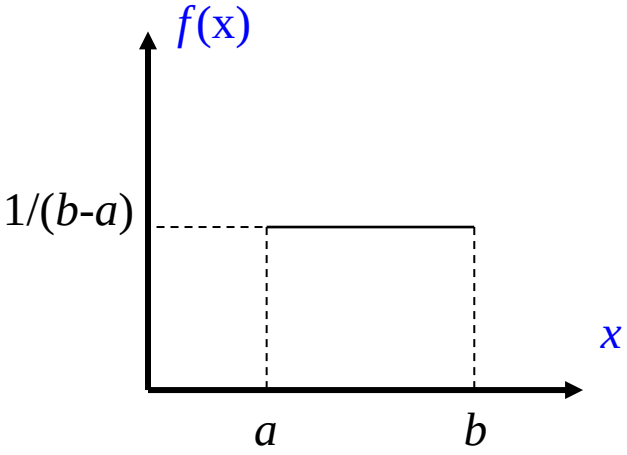
Parameters: $a, b \in \mathbb{Z}, a < b$

Sample space: $S = [a, b]$

Probability density function : $f(x) = \frac{1}{b-a}$

$$\mu_X = \frac{a+b}{2}$$

$$Var[X] = \frac{(b-a)^2}{12}$$



$$\mu_X = \int_a^b x \cdot f(x)dx = \frac{1}{b-a}\int_a^b x\,dx =$$

$$= \frac{1}{b-a} \cdot \left[\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$$

$$Var[X] = E[X^2] - \mu_X^2 = \left(\frac{1}{b-a}\int_a^b x^2\,dx\right) - \mu_X^2 = \frac{1}{b-a} \cdot \left[\frac{x^3}{3}\right]_a^b - \mu_X^2$$

$$= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} - \frac{(a+b)^2}{4}$$

$$= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{(b-a)^2}{12}$$

Exponential distribution $X = (S, f) \sim \text{Exp}(\lambda)$

Parameter: $\lambda \in \mathbb{R}_{>0}$ (rate)

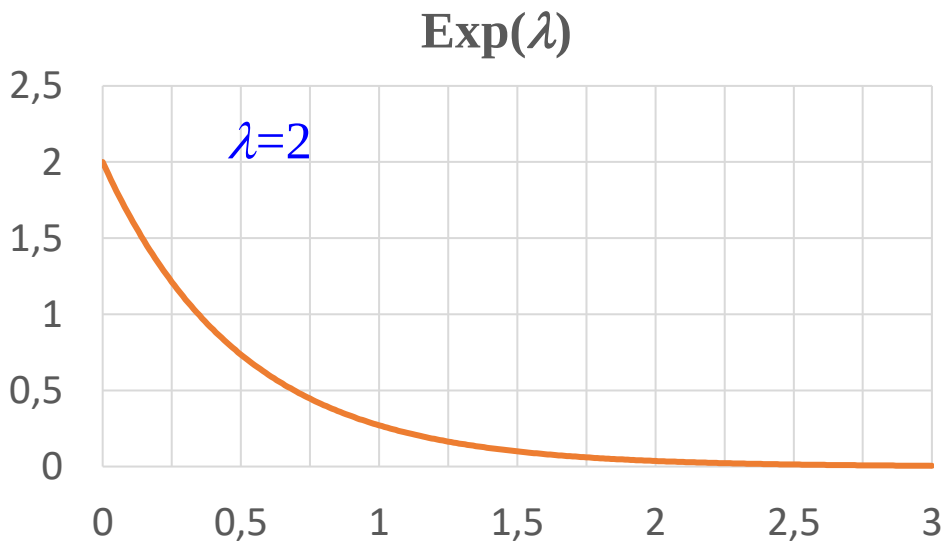
Sample space: $S = \mathbb{R}_{\geq 0}$

Probability density function: $f(x) = \lambda\, e^{-\lambda x}$

$$\mu_X = \frac{1}{\lambda}$$

$$Var[X] = \frac{1}{\lambda^2}$$

See in the following for a computation of mean and variance using moment generation function.



Normal (or Gaussian) distribution $X = (S, f) \sim N(\mu, \sigma^2)$

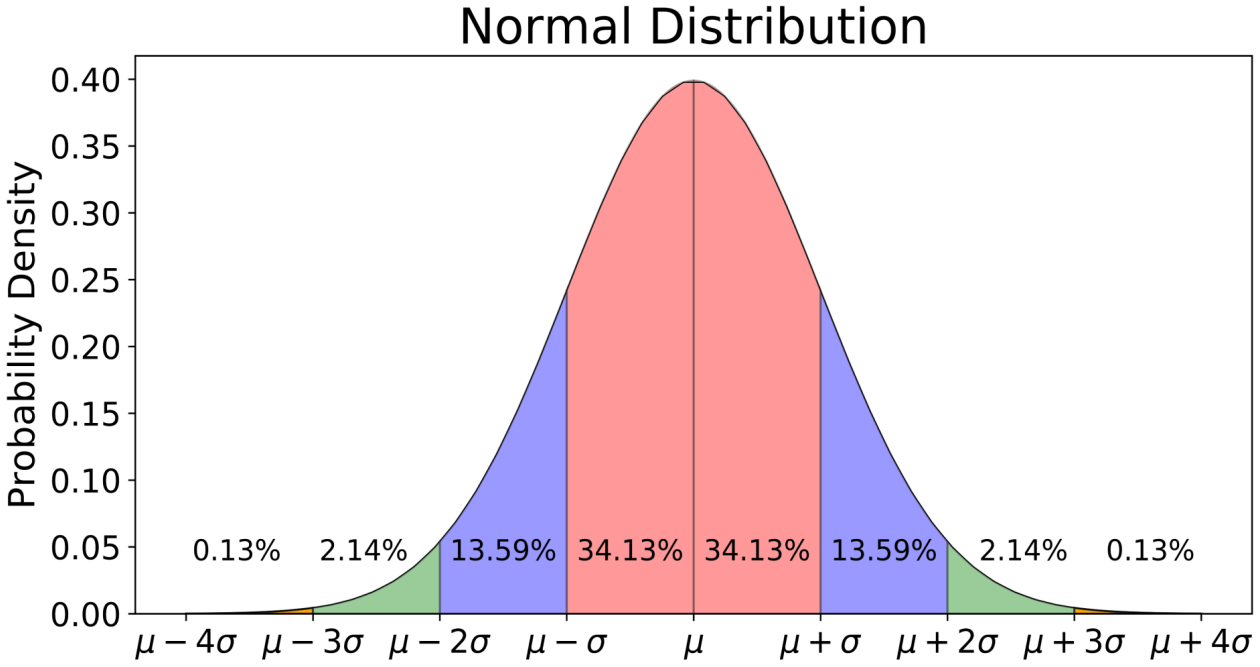
Parameters: $\mu \in \mathbb{R}$ (mean); $\sigma \in \mathbb{R}_{>0}$ (standard dev.)

Sample space: $S = \mathbb{R}$

Probability density function: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$

$$\mu_X = \mu$$

$$Var[X] = \sigma^2$$



Mean μ and variance σ^2 are the two parameters that characterize the distribution.

Moment Generating Function (discrete case)

Consider a discrete and nonnegative random variable $X = (S,p)$.

The moment generating function (MGF) of X is the Z-transform of its pmf $p(k)$

$$\Pi_X(z) = \mathcal{Z}[p(k)] \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} p(k) z^{-k}$$

All moments can be derived from it. Means and variance are:

$$\mu_X = E[X] = -\left.\frac{d\Pi_X(z)}{dz}\right|_{z=1} \qquad Var[X] = \left.\frac{d^2\Pi_X(z)}{dz^2}\right|_{z=1} - \mu_X - \mu_X^2$$

Proof for μ_X

$$\frac{d}{dz}\Pi_X(z) = \frac{d}{dz}\left(\sum_{k=0}^{\infty} p(k) z^{-k}\right) = \sum_{k=0}^{\infty} p(k)\frac{dz^{-k}}{dz} = \sum_{k=0}^{\infty} p(k)\cdot(-k)\cdot z^{-k-1}$$

$$\Rightarrow \mu_X = E[X] = -\left.\frac{d\Pi_X(z)}{dz}\right|_{z=1} = \sum_{k=0}^{\infty} k\cdot p(k) \equiv \mu_X$$

Example: number of heads flipping 2 coins $X = (S,p)$

k	$p(k)$
0	1/4
1	1/2
2	1/4

Moment generating function:

$$\Pi_X(z) = \mathcal{Z}[p(k)] \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} p(k) z^{-k} = \frac{1}{4} + \frac{1}{2} z^{-1} + \frac{1}{4} z^{-2}$$

Mean:
$$\mu_X = E[X] = -\frac{d\Pi_X(z)}{dz}\bigg|_{z=1} = -\left(\frac{-z^{-2}}{2} + \frac{-2z^{-3}}{4}\right)\bigg|_{z=1} = 1$$

Variance
$$Var[X] = \frac{d^2\Pi_X(z)}{dz^2}\bigg|_{z=1} - \mu_X - \mu_X^2 = \left(z^{-3} + \frac{6z^{-4}}{4}\right)\bigg|_{z=1} - 1 - 1$$

$$= \frac{10}{4} - 2 = \frac{1}{2}$$

Remark: Same values we obtained by direct calculation.

Example: Geometric random variable

$$\sum_{k=0}^{\infty} a^k = \frac{1}{1-a} \qquad |a| < 1$$

pdf: $p(k) = \pi \cdot (1-\pi)^{k-1} \quad \text{for } k \in \mathbb{N}_{>0}$

MGF: $\Pi_X(Z) = \mathcal{Z}[\pi \cdot (1-\pi)^{k-1}] = \sum_{k=1}^{\infty} \pi \cdot (1-\pi)^{k-1} z^{-k}$

$$= \frac{\pi}{1-\pi} \sum_{k=1}^{\infty} \left(\frac{1-\pi}{z}\right)^k = \frac{\pi}{1-\pi}\left(-1 + \sum_{k=0}^{\infty} \left(\frac{1-\pi}{z}\right)^k\right)$$

$$= \frac{\pi}{1-\pi}\left(-1 + \frac{1}{1-(1-\pi)/z}\right) = \frac{\pi}{1-\pi}\left(-1 + \frac{z}{z-(1-\pi)}\right) = \frac{\pi}{z-(1-\pi)}$$

mean

$$\mu_X = E[X] = -\frac{d\Pi_X(z)}{dz}\bigg|_{z=1} = \frac{\pi}{(z-(1-\pi))^2}\bigg|_{z=1} = \frac{1}{\pi}$$

variance

$$Var[X] = \frac{d^2\Pi_X(z)}{dz^2}\bigg|_{z=1} - \mu_X - \mu_X^2 = \frac{2\pi}{(z-(1-\pi))^3}\bigg|_{z=1} - \frac{1}{\pi} - \frac{1}{\pi^2}$$

$$= \frac{2}{\pi^2} - \frac{1}{\pi} - \frac{1}{\pi^2} = \frac{1-\pi}{\pi^2}$$

Moment Generating Function (continuous case)

Consider a continuous and nonnegative random variable $X = (S, f)$.

The moment generating function (MGF) of X is the Laplace-transform of its probability density function $f(x)$, i.e.,

$$\Pi_X(s) = \mathcal{L}[f(x)] \stackrel{\text{def}}{=} \int_0^\infty e^{-sx} f(x)\,dx$$

All moments can be derived from it. **Mean** and **variance** are:

$$\mu_X = E[X] = -\left.\frac{d\Pi_X(s)}{ds}\right|_{s=0} \qquad Var[X] = \left.\frac{d^2\Pi_X(s)}{ds^2}\right|_{s=0} - \mu_X^2$$

Proof for μ_X

$$\frac{d}{ds}\Pi_X(s) = \frac{d}{ds}\int_0^\infty e^{-sx} f(x)dx = \int_0^\infty \frac{d}{ds} e^{-sx} \cdot f(x)dx = \int_0^\infty (-x)e^{-sx} \cdot f(x)dx$$

$$\Rightarrow \mu_X = E[X] = -\left.\frac{d\Pi_X(s)}{ds}\right|_{s=0} = -\int_0^\infty (-x)e^{-sx} \cdot f(x)dx \,\Big|_{s=0} = \int_0^\infty x \cdot f(x)dx \equiv \mu_X$$

Example: Exponential random variable

pdf: $f(x) = \lambda\ e^{-\lambda x}$

MGF: $\Pi_X(s) = \mathcal{L}[\ \lambda e^{-\lambda x}] = \dfrac{\lambda}{s+\lambda}$

mean $\mu_X = E[X] = -\dfrac{d\Pi_X(s)}{ds}\Big|_{s=0} = \dfrac{\lambda}{(s+\lambda)^2}\Big|_{s=0} = \dfrac{1}{\lambda}$

variance $Var[X] = \dfrac{d^2\Pi_X(s)}{ds^2} - \mu_X^2 = \dfrac{2\,\lambda}{(s+\lambda)^3}\Big|_{s=0} - \dfrac{1}{\lambda^2} = \dfrac{1}{\lambda^2}$

Covariance and Correlation

We saw that variance of the sum of n independent r. v. is the sum of the variances:

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

Assume the random variables are positively dependent, namely, when one takes high values, the others are likely to obtain high value as well.

In such a case, the variance of their sum may be much higher than the sum of the individual variances.

Therefore, there is a need to define a quantitative measure for dependence between random variables. Such measures are

$$Cov[X_1, X_2] = E[(X_1 - E[X_1]) \cdot (X_2 - E[X_2])]$$

$$Corr[X_1, X_2] = \frac{Cov[X_1, X_2]}{\sigma_{X_1}\sigma_{X_2}}$$

Covariance

The covariance provides a measure of how much the two random variables vary together. It gives information about their dependence.

Def: The covariance of two random variables X_1 and X_2 is

$$Cov[X_1, X_2] = E[(X_1 - E[X_1]) \cdot (X_2 - E[X_2])]$$

Interpretation: if high values of X_1 imply high values of X_2, and low values of X_1 imply low values of X_2, the covariance is high.

It is easy show that

$$Cov[X_1, X_2] = E[X_1 \cdot X_2] - E[X_1] \cdot E[X_2]$$

Hence, if X_1 and X_2 are independent, then $Cov[X_1, X_2] = 0$.

(The converse does not always hold.)

It also holds that:

$$Var[X_1 + X_2] = Var[X_1] + Var[X_2] + 2 \cdot Cov[X_1, X_2]$$

Covariance, Auto-correlation and Correlation matrix

Let $\boldsymbol{X} = [X_1, X_2, \dots, X_N]$ a random column vector, stacking many random variables X_i

The covariance matrix $K_{\boldsymbol{XX}}$ is a square matrix collecting the covariances between each pair of elements of $\boldsymbol{X}$, namely:

$$K_{\boldsymbol{XX}} = \left[K_{X_i X_j}\right] \quad \text{where} \quad K_{X_i X_j} = Cov\left[X_i, X_j\right] = E\left[X_i \cdot X_j\right] - E[X_i] \cdot E\left[X_j\right]^T$$

It is symmetric and positive semi-definite and its main diagonal contains variances (i.e., the covariance of each element with itself).

The covariance matrix $K_{\boldsymbol{XX}}$ is related also to the correlation matrix $Corr(\boldsymbol{X})$

$$Corr(\boldsymbol{X}) = diag(K_{\boldsymbol{XX}})^{-1} \cdot K_{\boldsymbol{XX}} \cdot diag(K_{\boldsymbol{XX}})^{-1} \qquad \text{where} \quad Corr(X_i, X_i) = 1$$

The covariance matrix $K_{\boldsymbol{XX}}$ is related also to the auto-correlation matrix $R_{\boldsymbol{XX}}$

$$K_{\boldsymbol{XX}} = R_{\boldsymbol{XX}} - E[\boldsymbol{X}] \cdot E[\boldsymbol{X}]^T \qquad thus \qquad R_{\boldsymbol{XX}} = K_{\boldsymbol{XX}} + E[\boldsymbol{X}] \cdot E[\boldsymbol{X}]^T$$

Estimation of K_{XX}, and $Corr(X)$ from series

```
s=rng(0); % Specifies the seed for the MATLAB® random number generator
% Generate two time series of lenght n from two Gaussian random variables
n=1000;
mu1 = 0; sigma1 = 1;
x1 = normrnd(mu1, sigma1, [n,1]); % column vector of n x 1
mu2 = 0; sigma2 = 2;
x2 = x1+ normrnd(mu2, sigma2, [n,1]);

X=[x1,x2];
% Compute expectation and variance as a check test
exp_X=mean(X) % correspoing to >> sum(X)/n
var_X=var(X) % corresponding to >> sum( (X-exp_X).^2)/(n-1)

% Compute the covariance
K_XX = cov(X)
K_x1_x2=sum( (x1-exp_X(1)).*(x2-exp_X(2)))/(n-1)
% Compute the correlation
cor_X = corrcoef(X)

% Display the results of interest
disp(['Covariance between x1 and x2: ', num2str(K_XX(1, 2))]);
disp(['Correlation between x1 and x2: ', num2str(cor_X(1, 2))]);
```

```
exp_X =

   -0.0326    0.0412


var_X =

    0.9979    4.9497


K_XX =

    0.9979    0.9794
    0.9794    4.9497


K_x1_x2 =

    0.9794


cor_X =

    1.0000    0.4407
    0.4407    1.0000

Covariance between x1 and x2: 0.97942
Correlation between x1 and x2: 0.44069
```

Probability function estimations from data-series using «*ksdensity*»

```
>> help ksdensity
 ksdensity - Kernel smoothing function estimate for univariate and bivariate data
    This MATLAB function returns a probability density estimate, f, for the
    sample data in the vector or two-column matrix x.
```

```
s=rng(0); % Specifies the seed for the MATLAB® random number generator
% Generate two time series of lenght n from two Gaussian random variables
n=1000;
mu1 = 0; sigma1 = 1;
x1 = normrnd(mu1, sigma1, [n,1]); % column vector of n x 1
mu2 = 0; sigma2 = 2;
x2 = x1+ normrnd(mu2, sigma2, [n,1]);

% Estimate the PDF
x = [-10:0.1:10]';
% Returns a probability density estimate from a time series
pdf_x1 = ksdensity(x1, x, 'Kernel', 'normal', 'Bandwidth', 0.5);
pdf_x2 = ksdensity(x2, x, 'Kernel', 'normal', 'Bandwidth', 0.5);

% Plot the PDFs
figure(1); hold on; grid on
plot(x, pdf_x1, 'LineWidth', 2);
plot(x, pdf_x2, 'LineWidth', 2);
xlabel('x'); ylabel('PDF');
legend('x1', 'x2');
```