Basics of Probability Theory

Antonio Capone

Politecnico di Milano

Content

Probabilities

Basic definitions

Probability theory provides a mathematical description of the possible outcomes of experiments. Let $\mathcal{E}$ be an experiment which provides a finite number of outcomes $\alpha_1, \alpha_2, \ldots, \alpha_n$.
We have the following definitions:

- **Trial** is the execution of $\mathcal{E}$ which leads to a outcome, or sample, α and only one.
- **Space or stochastic universe** associated with the experiment $\mathcal{E}$ is the set $S = \alpha_1, \alpha_2, \ldots . \alpha_n$ of all possible outcomes of $\mathcal{E}$.
- **Event** is any set A of outcomes and what is a any subset of S.
- An **Elementary Event** (or Sample Event) is a set $E \subset S$ with a single outcome.

Basic definitions

(cont'd) definitions:

- A **Certain Event** is the event corresponding to S.
- The **Impossible Event** is the empty set $\emptyset$.
- Events can be combined with operations in use in Set theory, obtaining events such as **union** (or sum) events, **conjunction** (or product or intersection) events, **complement** events, and **difference** events.
- **We say that in a trial event A occurs if the outcome of the trial belongs to A.**

Basic properties

Some basic properties directly follow from definitions:

- The certain event always occurs;
- The impossible event never occurs;
- A union event occurs if at least one of the component events occurs;
- A joint event occurs if all the components events occur simultaneously;
- Disjoint events can not occur simultaneously, and for this reasons they are called mutually exclusive (as for example sample events and complementary events).

Probability definition

The probability $P(A)$ of an event $A \subset S$ is a **measure** *defined on S so as to satisfy the following axioms:*

Axiom I: *$P(A)$ is a nonnegative real number associated with the event.*

$$P(A) \geq 0 \tag{1}$$

Axiom II: *the probability of the certain event is one.*

$$P(S) = 1 \tag{2}$$

Axiom III: *if A e B are disjoint events*

$$P(A + B) = P(A) + P(B) \tag{3}$$

Probability definition

Based on the axioms, probabilities have the following properties:

$$P(A) = 1 - P(\bar{A}) \leq 1 \tag{4}$$

$$P(\phi) = 0 \tag{5}$$

If $B \subset A$, then

$$P(B) \leq P(A) \tag{6}$$

If $A_1, A_2, \ldots, A_n$ are disjoint events, and $A = A_1 + A_2 + \ldots + A_n$, then we have

$$P(A) = P(A_1) + P(A_2) + \ldots + P(A_n). \tag{7}$$

Probability space description

- A complete stochastic description of an experiment is obtained when a sufficient number of probabilities to events have been assigned
- From a mathematical point of view the assignments are arbitrary within the limits of the axioms but obviously for applications we use probability for measuring the likelihood that an event occurs
- A formal link between probabilities and occurrence of events in repeated trials is provided by the law of large numbers (see later).

We say that an experiment $\mathcal{E}$ (or probability space S) is completely described from the probabilistic point of view when probabilities are given for each elementary event E_i

$$p_i = P(E_i)$$

The probability of any event A can then be obtained as the sum of the probabilities of its elementary events.

Uniform spaces

- If all the n_S elementary events E_i are equally probable, the space S is called **uniform**
- For a uniform space S of n_S elements, the probability of an event A composed of r_A elementary events is:

$$P(A) = \frac{r_A}{n_S} \tag{8}$$

- Therefore for uniform spaces, probability calculations are performed by counting techniques (as in combinatorial calculus) in order to get r_A and n_S.

Urn model

- A useful way of describing uniform finite spaces is through the *urn model*

- It is an ideal experiment consisting in drawing k objects (elements) from an urn containing n objects (like e.g. numbered or colored balls)
- The model assumes that all possible outcomes consisting of all the groups that can be formed with k out of n objects are equally probable

Counting groups

- Counting the number of k out of n objects and the number of groups in a given event allows to completely characterize experiments in the urn model
- If groups differ in at least one element or in the order they appear in the group, and if objects are drawn together or one by one with no replacement

$$n_S = (n)_k = n(n-1)\dots(n-k+1) = \frac{n!}{(n-k)!}.$$

- While, if objects are drawn one by one with re-insertion/replacement of the drawn element in the urn

$$n_S = n^k$$

Counting groups

- If the order does not count we have to divide by the number $k!$ of possible permutations and we get:

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

in the case of drawn with no replacement, and

$$n_S = \frac{n^k}{k!},$$

in the case of with replacement.

Example (1)

In an urn there are ten objects representing the ten digits $0, 1, \ldots, 9$. Evaluate the probability that, upon drawing of 3 elements, the three digits form the event

- $A =$ number 567
- $B =$ number with three consecutive increasing digits

We have $n_S = (10)_3 = 10 \cdot 9 \cdot 8 = 720$

$$r_A = 1 \qquad P(A) = \frac{1}{720}$$

$$r_B = 8 \qquad P(B) = \frac{8}{720}$$

Example (2)

Like in previous example but assuming that we have three consecutive drawings with the replacement of the element previously drawn. Evaluate the probability of events:

- $A =$ number 567
- $B =$ number with three consecutive increasing digits
- $C =$ number with all equal digits.

$$n_S = 10^3$$

$$r_A = 1 \qquad P(A) = \frac{1}{1000}$$

$$r_B = 8 \qquad P(B) = \frac{8}{1000}$$

$$r_C = 10 \qquad P(C) = \frac{10}{1000}$$

Example (3)

Evaluate the minimum number of people k you need to pick so that the probability of having at least two people born on the same day is greater or equal 0.5.
The experiment is equivalent to drawing k numbers out of an urn that contains 365 objects, each representing a different day of the year.
Denoting with $D_k = \{$ *extraction k objects all different* $\}$ we have

$$P(D_k) = \frac{(365)_k}{(365)^k} \qquad k \leq 365$$

The probability that at least two people are born in the same day is:

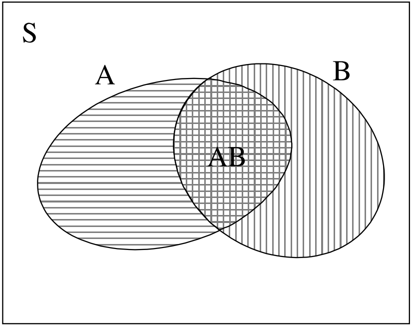
$$P(\overline{D}_k) = 1 - P(D_k) = 1 - \frac{365!}{(365-k)!\ 365^k} > 0,5.$$

The non linear equation can be solved numerically. The result is $k = 23$.

Union of non disjoint events

Given events A and $B \subseteq S$ we have:

$$P(A+B) = P(A) + P(B) - P(AB) \qquad (9)$$



Example (4)

In a throw of dice, evaluate the probability that number is either even or less than 3.
Denote the event "even number" as A and the event "less than three" as B we have

$$P(D) = P(A + B) = P(A) + P(B) - P(AB),$$

We evaluate $P(A), P(B)$, and $P(AB)$ with the counting process and we find

$$P(A) = \frac{3}{6}, \qquad P(B) = \frac{2}{6}, \qquad P(AB) = \frac{1}{6}.$$

Substituting we get

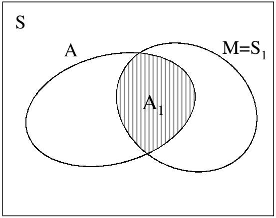
$$P(D) = \frac{1}{3} + \frac{1}{2} - \frac{1}{6} = \frac{4}{6}.$$

Conditional events

- It is often useful evaluating probabilities of events when we know that another event certainly occurs (conditional events)
- Formally, we want to evaluate the conditional probability that the outcome of a trial of $\mathcal{E}$, $\alpha \in A$ knowing that $\alpha \in M$.
- Obviously, knowing that the outcome is in set M gives some additional information on the possible occurrence of a given α, since all α that do not belong to M are excluded.
- For this reason the probability of the occurrence of A is no longer the original one $P(A)$, usually referred to as "a priori" probability, but a different one, usually referred to as "a posteriori" (after knowing $\alpha \in M$).

Conditional probability

- The probability description of conditional events can be given as a function of unconditional probabilities considering a new space $S_1 = M$ and a new event $A_1 = AM$



- Using axioms it is easy to show that the probability of A conditioned to M is:

$$P(A/M) = \frac{P(AM)}{P(M)}. \tag{10}$$

Example (5)

Evaluate the probability that the outcome of a throw of the dice is 2 knowing that the result is even.
We have

$$S = \{1, 2, 3, 4, 5, 6\} \qquad S_1 = \{2, 4, 6\}$$

and from (10)

$$P(2/\text{ even}) = P_1(2) = \frac{1}{3}$$

Total Probability

Theorem of Total Probability

Given $M_1, M_2, \ldots M_n$ disjoint events such that $M_1 + M_2 + \ldots + M_n = S$ (or more in general $M_1 + M_2 \ldots M_N \supset A$), we have

$$P(A) = \sum_{i=1}^{n} P(A/M_i)P(M_i). \tag{11}$$

In fact, since events AM_i are disjoint, and their union provides A, we can write

$$P(A) = \sum_{i=1}^{n} P(AM_i), \tag{12}$$

and using the relation

$$P(AM_i) = P(A/M_i)P(M_i), \tag{13}$$

we get (11). ♣

Example (6)

Note: in all cases where calculating the probability of an event conditioned to others is easier that direct calculation, total probability theorem is particularly useful

A box contains three types of objects, some of which are defective, in these proportions

type A - 2500 of which 10% defective
type B - 500 of which 40% defective
type C - 1000 of which 30% defective

If we draw an object at random, what is the probability $P(D)$ that, drawing an object, this is found to be defective?

Example (6)

The probability to draw an object of type A, B, C are respectively

$$P(A) = \frac{2500}{4000} = \frac{5}{8}; \quad P(B) = \frac{500}{4000} = \frac{1}{8}; \quad P(C) = \frac{1000}{4000} = \frac{2}{8}.$$

Then we have

$$P(D/A) = \frac{10}{100}; \quad P(D/B) = \frac{40}{100}; \quad P(D/C) = \frac{30}{100};$$

and, finally, from total probability equation:

$$P(D) = P(D/A)P(A) + P(D/B)P(B) + P(D/C)P(C) = \frac{3}{16}$$

Example (7)

A game is based on the following experiment. A box contains n tags, each one reporting a number arbitrarily determined. The player draws at first r tags and observes their maximum value m_r. Then, further drawings are performed until a value m is observed such as $m > M_r$.

Player wins if $m = M$, where M is the maximum value among those reported on the n tags. We want to evaluate the probability $P(V)$ to win.

Example (7)

Since the positions of the maximum are equally likely, the probability that M is in position k is

$$P(M \text{ in } k) = \frac{1}{n}$$

The probability to win, with M in k, is zero if $k \leq r$.
For $k > r$ player wins if the maximum m_{k-1} among the first $k-1$ tags is within the first r, and this happens with probability

$$P_r(V/M \text{ in } k) = \frac{r}{k-1}.$$

with $k > r$.

Example (7)

By the Total Probability Theorem (11) we have:

$$P_r(V) = \sum_{k=1}^{n} P_r(V/M \text{ in } k)P(M \text{ in } k) =$$

$$= \sum_{k=1}^{r} P_r(V/M \text{ in } k)P(M \text{ in } k) + \sum_{k=r+1}^{n} P_r(V/M \text{ in } k)P(M \text{ in } k) =$$

$$= \sum_{k=r+1}^{n} \frac{r}{k-1}\,\frac{1}{n} = \frac{r}{n}\sum_{k=r}^{n-1} \frac{1}{k}.$$

Bayes' Formulas

If we use total probability two times in a direct and a reverse way we get:

$$P(M/A) = \frac{P(AM)}{P(A)} = P(A/M)\frac{P(M)}{P(A)}. \tag{14}$$

Using the total probability theorem (11) for the denominator, the above expression can be re-written as:

$$P(M_k/A) = \frac{P(A/M_k)P(M_k)}{\sum_{j=1}^{n} P(A/M_j)P(M_j)}. \tag{15}$$

Bayes' Formulas

- Bayes' formulas are particularly useful for evaluating conditional probabilities before and after observing the occurrence of an event.
- In particular, the second formula is referred to as "Bayes' rule for the 'a posteriori' probability", that is after observing the occurrence of the event A.
- In this case, $P(M_i)$ are called "a priori probabilities" and $P(M_i/A)$ "a posteriori probabilities".

Example (8)

An object drawn at random from the box in Example (6) is found to be defective. Evaluate the probabilities that it is of type A, B and C respectively.

Using Bayes' formula and the preceding results we have

$$P(A/D) = \frac{P(D/A)P(A)}{P(D)} = \frac{10}{30}; \qquad P(B/D) = \frac{P(D/B)P(B)}{P(D)} = \frac{8}{30}.$$

Similarly we have

$$P(C/D) = \frac{12}{30}.$$

Note that while "a priori" the most likely type of object is A, after observing that the object is defective the most likely type of object is C.

Example (9)

Assume that you are presented with three dices, two of them fair and the other a counterfeit that always gives 6. If you randomly pick one of the three dices, the probability that it's the counterfeit is $1/3$.

This is the a priori probability of the hypothesis that the dice is counterfeit. Now after throwing the dice, you get 6 for two consecutive times. Seeing this new evidence, you want to calculate the revised a posteriori probability that it is the counterfeit.

Example (9)

The 'a priori' probability of counterfeit dice is

$$P(D_c) = \frac{1}{3},$$

while that of a fair dice is

$$P(D_f) = \frac{2}{3}.$$

We have:

$$P(66/D_f) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

$$P(66/D_c) = 1$$

and then using Bayes' formula:

$$P(D_c/66) = \frac{P(66/D_c)P(D_c)}{P(66/D_c)P(D_c) + P(66/D_f)P(D_f)} = \frac{18}{19}$$

Statistical independence

Two events A and $B \subset S$ are said statistically independent if and only if

$$P(AB) = P(A)P(B) \tag{16}$$

The meaning of statistical independence is immediate if we observe that if A and B are independent:

$$P(A/B) = P(A), \quad P(B/A) = P(B).$$

This means that the probability of A is not influenced by the occurrence of B and vice versa.

Example (10)

In a throw of the dice, check whether the following events
$A =$ even number
$B =$ number one, or two or three
are statistically independent.

We have

$$P(A) = 1/2; \qquad P(B) = 1/2; \qquad P(AB) = 1/6;$$

that is

$$P(AB) \neq P(A)P(B)$$

Hence, events A e B are not statistically independent.

Example (11)

In a throw of the dice, check whether the following events $A =$ an even number appears
$B =$ number one, or two, or three, or four appears
are statistically independent.

We have

$$P(A) = 1/2; \qquad P(B) = 2/3; \qquad P(AB) = 2/6$$

and

$$P(AB) = P(A)P(B)$$

Hence, events A e B are indeed statistically independent.

Random Variables

Spaces with countable outcomes

To deal with spaces S with infinite outcomes, we must add another axiom that extends the summation of the probability measure over infinite terms:

Axiom IIIa: *If* $A_1, A_2, \ldots, A_n \ldots$ *are disjoint events and* $A = A_1 + A_2 + \ldots + A_n + \ldots$, *then*

$$P(A) = P(A_1) + P(A_2) + \ldots + P(A_n) + \ldots \tag{17}$$

An example of this type of space is the number of coin flips to get a head. This number is not limited, as the head could never appear in n trial, whichever n is.
These spaces are said **countable** (number of elements of the same cardinality of natural numbers), and can be managed with the methods of finite spaces.

Spaces with uncountable outcomes

- We can consider also **uncountable** spaces, such as is the case when the outcomes is, for example, a point in an interval, or in any general geometrical space
- The extension to these spaces is however not straightforward and require new instruments
- The approach used is that of "transforming" the space of outcomes into another one more convenient for assigning probabilities.
- In particular, we map outcomes and events (subsets) of S into the space of real numbers (using integer numbers as a subset to include countable spaces as a special case)

Definition of Random Variable

- Let us consider a real function $X(\alpha)$ defined on the space S of the outcomes that binds the set S and the set of real numbers R in order to match every $\alpha \in S$ with one and only one value $X(\alpha) in R$
- With this function, each event $A \subset S$ corresponds a set $I \subset R$ such that for every $\alpha \in A$ we have $X(\alpha) \in I$.
- In this way the description of an experiment in terms of results α, A and probability events for $P_S(A)$ in S, can be replaced with description in terms of real numbers x, sets I and probabilities $P_R(I)$ in R.

A function $X(\alpha)$ which satisfies the above conditions is called **random variable**.

Typically, the notation is simplified omitting the relation with α and capital letters, such as X, Y, Z, are used to indicate random variables.

Cumulative Distribution Function (CDF)

Let X be a Random Variable (RV) and x a real number.
The probability of the event $\{X \leq x\}$ is a function of the real variable x:

$$F_X(x) = P(X \leq x) \tag{18}$$

and is called **Cumulative Probability Distribution Function** (CDF) of X.

$F_X(x)$ completely describes RV X.
In fact we have, for any x_1, x_2 and x:

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1) \tag{19}$$

$$P(X = x) = F(x) - F(x^-) \tag{20}$$

We denote $F(x^+) = \lim_{\varepsilon \to 0}$ and $F(x^-) = \lim_{\varepsilon \to 0} F(x - \varepsilon)$

Cumulative Distribution Function (CDF)

The CDF has the following properties:

1. it has the following limits

$$F(-\infty) = 0 \qquad F(+\infty) = 1 \tag{21}$$

2. it is a monotonic non decreasing function of x:

$$F(x_1) \leq F(x_2) \ \text{ per } \ x_1 \leq x_2 \tag{22}$$

3. it is right continuous :

$$F(x^+) = F(x) \tag{23}$$

Continuous Random Variables

Probability Density Function (pdf)

- A RV X is **continuous** if its CDF $F_X(x)$ is a continuous function in R, together with its first derivative, except at most a countable set of points where the derivative does not exist.
- Since for a continuous RV $F_X(x)$, we have

$$P(X = x) = 0.$$

- It is therefore useful introducing the **probability density function** (pdf) of RV X, $f_X(x)$ defined as the derivative of the corresponding CDF:

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{24}$$

  The definition is then completed by assigning arbitrary positive values where the derivative does not exist.

Probability Density Function (pdf)

From the definition and properties of $F(x)$ we have

$$f(x) \geq 0 \tag{25}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{26}$$

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(x)dx \tag{27}$$

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx \tag{28}$$

Directly from the definition we have:

$$f(x) = \lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x}. \tag{29}$$

This shows that the pdf can be interpreted as the normalized probability that the RV belongs to a small interval around x and, dimensionally, is a density, hence the name.

Example (12) - Uniform RV

We want to find the CDF and pdf of RV X, defined as the coordinate of a point randomly selected in interval $[a, b]$ of x axis.

We have

$$F_X(x) = \begin{cases} \dfrac{x-a}{b-a} & (a \le x \le b) \\ \\ 0 & (x < a) \\ 1 & (x > b) \end{cases} \tag{30}$$

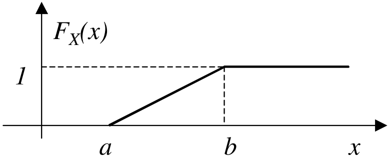
and from definition (29):

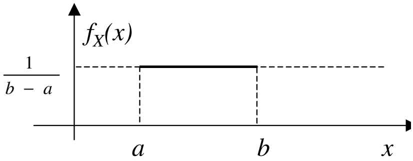
$$f(x) = \lim_{\Delta x \to 0} \frac{\Delta x}{b-a} \frac{1}{\Delta x}$$

Example (12)

we get

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & (a \leq x \leq b) \\ \\ 0 & \text{elsewhere} \end{cases} \tag{31}$$

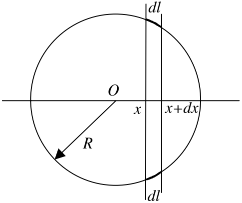




A RV that satisfies (30) and (31) is called "uniformly distributed" and the pdf is said "uniform".

Example (13)

A point P is drawn uniformly on a circumference of radius R and center in the origin of axes. Find the pdf of RV X, defined as the coordinate of orthogonal projection of P on the horizontal axis.



To find the pdf let us use the definition (29). With reference to the figure, $P(x < X \leq x + \Delta x)$ is the probability that P lies in one of two small arcs shown in the figure, each having a length

$$d\ell = \sqrt{dx^2 + dy^2} = dx\sqrt{1 + (\frac{dy}{dx})^2}$$

Example (13)

Being $y = \sqrt{R^2 - x^2}$, we get:

$$dy = -\frac{xdx}{\sqrt{R^2 - x^2}}$$

by replacing it in the expression above we get

$$d\ell = dx\sqrt{1 + \frac{x^2(dx)^2}{R^2 - x^2}\frac{1}{(dx)^2}} = \frac{dx}{\sqrt{1 - (\frac{x}{R})^2}}.$$

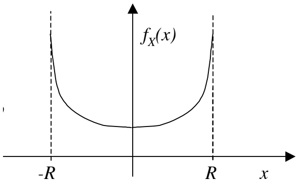
Example (13)

Then we have:

$$f_X(x) = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \frac{2\Delta l}{2\pi R} = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \frac{1}{2\pi R} \frac{2\Delta x}{\sqrt{1-(x/R)^2}} =$$

$$= \frac{1}{\pi R} \frac{1}{\sqrt{1-(x/R)^2}}$$

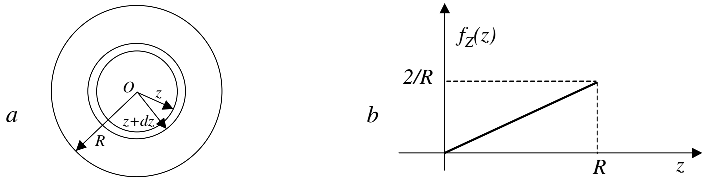
for $(\mid x \mid \leq R)$ and zero elsewhere.



Example (14)

A point P is drawn uniformly in a circle of radius R. Derive the pdf of RV Z, defined as the distance of P from the center O of the circle.

$P(z < Z \leq z + \Delta z)$ is the probability that P is taken in the annulus shown in the figure, whose area is $2\pi z \Delta z$.



By definition (29) we get

$$f_Z(z) = \lim_{\Delta z \to 0} \frac{2\pi z \Delta z}{\pi R^2} \frac{1}{\Delta z} = \frac{2z}{R^2} \quad (0 \leq z \leq R)$$

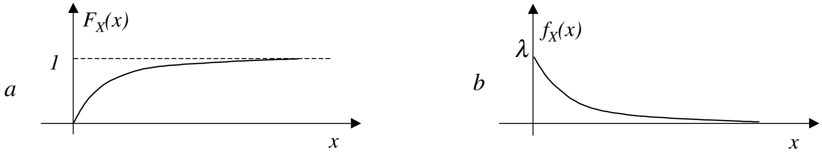
Example (15) - Negative Exponential RV

The "Negative Exponential" pdf is defined as:

$$F(x) = 1 - e^{-\lambda x} \quad (x \geq 0) \tag{32}$$

we have:

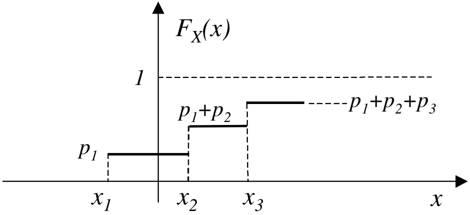
$$f(x) = \lambda e^{-\lambda x} \quad (x \geq 0) \tag{33}$$



Discrete Random Variables

Probability Distribution

A discrete RV X is characterized by a CDF $F_X(x)$ of a staircase type, with discontinuities in a countable set of points $x_i(i = 0, \pm 1, \pm 2 \dots)$, where it presents steps of value p_i:



In this case we get

$$P(X = x) = \begin{cases} p_i & x = x_i \\ 0 & x \neq x_i \end{cases} \tag{34}$$

This is called *Probability Distribution* of X.

Probability Distribution

For the Probability Distribution, we have

$$p_i \geq 0 \tag{35}$$

$$\sum_{i=-\infty}^{\infty} p_i = 1 \tag{36}$$

$$F(x) = \sum_{i=-\infty}^{M} p_i \tag{37}$$

where M is the maximum i for which $x_i \leq x$.
If the values x_i are integers, then RV X is said an **integer RV**.

Examples of discrete variables

- Distribution of a constant c:

$$P(X=x)=\begin{cases}1 & \text{for} \quad x=c \\ 0 & \text{elsewhere}\end{cases} \tag{38}$$

- The Bernoulli (binary) distribution:

$$P(X=x)=\begin{cases}p & \text{for} \quad x=1 \\ 1-p=q & \text{for} \quad x=0 \\ 0 & \text{elsewhere}\end{cases} \tag{39}$$

- The uniform distribution

$$P(X=x)=\begin{cases}\frac{1}{n} & \text{for} \quad x=x_i \quad (i=1,\ldots,n) \\ 0 & \text{elsewhere}\end{cases} \tag{40}$$

already encountered in examples with dices, draws from urns, etc.

Binomial distribution

- We can consider experiments obtained from repeating a single experiment multiple independent times
- Repeated independent trials, each of which with only two possible outcomes, say *success* (S) and *failure* (F) are called *Bernoulli* trials.
- Denoted $P(S) = p$ and $P(F) = q = 1 - p$, the probability $P(S_n = k)$ that in n Bernoulli trials k successes and $n - k$ failures occur is given by the following distribution

$$P(S_n = k) = \binom{n}{k} p^k q^{n-k} \qquad (0 \leq k \leq n) \tag{41}$$

- This distribution is called *Binomial of order n*

Example (16)

A quality control process tests some components out of a factory and components are found defective with a probability $p = 10^{-2}$. *Evaluate the probability that out of 10 components checked there are*

$A =$ only one defective

$B =$ two defective

$C =$ at least one defective

The 10 tests can be modeled as Bernoulli trials with success probability $p = \frac{1}{100}$. Then we have

$$P(A) = \binom{10}{1}\left(\frac{1}{100}\right)^1\left(\frac{99}{100}\right)^9 = 0,0913\ldots$$

$$P(B) = \binom{10}{2}\left(\frac{1}{100}\right)^2\left(\frac{99}{100}\right)^8 = 0,00415\ldots$$

$$P(C) = \sum_{k=1}^{10}\binom{10}{k}\left(\frac{1}{100}\right)^k\left(\frac{99}{100}\right)^{10-k} = 1-\binom{10}{0}\left(\frac{1}{100}\right)^0\left(\frac{99}{100}\right)^{10} = 0,0956\ldots$$

Moments of a pdf

For the pdf we can define some parameters that resume some properties of the function. These are called *moments*, and the most used are:

1. $k-$th order moments $(k = 1, 2, \ldots)$

$$m_k = \int_{-\infty}^{+\infty} x^k \; f(x)dx \tag{42}$$

2. $k-$th order central moments

$$\mu_k = \int_{-\infty}^{+\infty} (x - m_1)^k f(x)dx \tag{43}$$

Note that, depending on the specific pdf, some moments may not exist.

Moments of a pdf

Parameters of the same meaning can be given also for discrete variables in the form:

$$m_k = \sum_{i=-\infty}^{\infty} x_i^k \; p_i \tag{44}$$

$$\mu_k \sum_{i=-\infty}^{\infty} (x_i - m_1)^k \; p_i \tag{45}$$

Moments of a pdf

- The first order moment:

$$m_1 = \int_{-\infty}^{+\infty} x \, f(x) dx \quad \text{(also denoted } m_X\text{)}$$

  This can be considered as the coordinate of the center of mass interpreting the pdf as a mass distribution along the x with density $f(x)$.

- The index μ_2:

$$\mu_2 = \int_{-\infty}^{+\infty} (x - m_1)^2 f(x) dx \quad \text{(also denoted } \sigma_X^2\text{)}$$

  provides an index of the dispersion of the distribution around $x = m_X$

- we have

$$\mu_2 = m_2 - m_1^2$$

Law of large numbers

- Let X be a RV whose pdf has first order moment m_1
- Denote with $X_1, X_2, \ldots, x_n$ the outcomes of the RV in n independent repetitions of the experiment
- and $\overline{X}_n$ their arithmetic mean:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- we have the following **Low of large numbers**:

$$P(\lim_{n\to\infty} \overline{X}_n = m_1) = 1 \quad (46)$$

Law of large numbers

- Law of large numbers states that the average performed on a number n of outcomes of n independent trials, tends with probability 1 to m_1 when n tends to infinity.
- For this reason, m_1 is also called the *mean value* or *expected value* of RV X and it is also denoted by $E[X]$.
- This law is of great importance since it provides a relationship between a pure mathematical parameter, m_1, to another one $\overline{X}_n$ directly derived from an experiment.

Interpretation of probability

- Let's formulate the law of large numbers for probability p_A of event A
- Define the binary RV X such that it is $X = 1$ if A occurs and $X = 0$ otherwise
- If we perform n trials we have

$$\sum_{i=1}^{n} X_i = n_A$$

  being n_A the number of times A occurs
- we also observe that

$$m_1(X) = p_A$$

  and

$$\overline{X}_n = \frac{n_A}{n}.$$

- Therefore, the law of large numbers can be written as

$$P(\lim_{n\to\infty} \frac{n_A}{n} = p_A) = 1 \quad (47)$$

Properties of E[X]

Some properties of $E[X]$ are (for proofs see lecture notes):

1. If $f(x)$ is symmetric around a value of a and m_1 exists, then $m_1 = a$
2. If m_1 exists, it can be expressed as
$$m_1 = \int_0^{\infty} (1 - F(x))dx - \int_{-\infty}^{0} F(x)dx \tag{48}$$
3. If $F_X(x) = 0$ for $x < 0$, for $\alpha > 0$ the following inequality holds
$$P(X \geq \alpha) \leq \frac{E[X]}{\alpha} \tag{49}$$
Setting $v = \dfrac{\alpha}{E[X]}$ we get a different expression
$$P(X \geq vE[X]) \leq \frac{1}{v} \tag{50}$$
that shows how to establish a constraint upon the part of pdf that lies above the mean value ($v > 1$), based on the sole knowledge of the mean value.

Tchebichev inequality

- Central moment μ_2, is also called *variance* of RV X and denoted by σ_X^2, whereas σ_X is called *standard deviation*
- The variance represents a measure of the dispersion of $f(x)$ around its average value
- This is shown by the **Tchebichev Inequality**

$$P(\mid X - m_1 \mid \geq \ v\sigma) < \frac{1}{v^2} \tag{51}$$

- By setting $v\sigma = \varepsilon$ we get alternatively

$$P(m_1 - \varepsilon < X < m_1 + \varepsilon) \ \geq \ 1 - \frac{\sigma^2}{\varepsilon^2} \tag{52}$$

$$P(|X - m_1| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \tag{53}$$

Example (17)

Let us apply Tchebichev inequality to bound the probability that the frequency of HEADS in flipping a fair coin n times exceeds $0.5 \pm \varepsilon$.

The frequency of HEADS in n trials is H/n where H is the RV number of HEADS in n trials. This has a Binomial distribution with average $n/2$ and $\sigma^2(H) = n/4$. Therefore,

$$m_1(H/n) = \frac{1}{2}$$

$$\sigma^2(H/n) = \frac{1}{4n}$$

Tchebichev inequality says

$$P(|H/n - m_1| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

and substituting

$$P(|H/n - 0.5| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2}$$

Example (17)

we have

$$\begin{aligned}
\varepsilon = 0.1, \qquad n &= 10, \qquad P \leq 2.5(???) \\
\varepsilon = 0.1, \qquad n &= 100, \qquad P \leq 0.25 \\
\varepsilon = 0.1, \qquad n &= 1000, \qquad P \leq 0.025 \\
\varepsilon = 0.1, \qquad n &= 10000, \qquad P \leq 0.0025
\end{aligned}$$

We also see that

$$\lim_{n \to \infty} P(|H/n - 0.5| \geq \varepsilon) = 0, \qquad \forall \varepsilon > 0$$

that provides a kind of demonstration of the law of large numbers.

Conditional distributions and densities

Let M be an event of space S where RV X is defined. We define CDF of X conditional to M (provided that $P(M) \neq 0$) the function:

$$F_X(x/M) = P(X \ \leq \ x/M) \tag{54}$$

and similarly for the density a

$$f_X(x/M) = \frac{dF_X(x/M)}{dx} = \lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x/M)}{\Delta x} \tag{55}$$

It is easy to check that the above defined functions have all the properties of the CDF and pdf.

Conditional distributions and densities

Interesting cases are those where also event M is described in terms of RV X.
We define probability of an event A conditional to the value x assumed by a RV X, assuming that $f_X(x) \neq 0$, as the limit

$$P(A/X = x) = \lim_{\Delta x \to 0} P(A/x < X \leq x + \Delta x) \tag{56}$$

From Bayes formula (14) we get:

$$P(A/X = x) = \lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x/A)P(A)}{P(x < X \leq x + \Delta x)}$$

and multiplying by Δx above and below, and taking the limit, we have finally

$$P(A/X = x) = \frac{f_X(x/A)P(A)}{f_X(x)} \tag{57}$$

Total probability law for the continuous case

From (57) we have by integrating left and right sides:

$$\int_{-\infty}^{+\infty} f_X(x/A)P(A)dx = \int_{-\infty}^{+\infty} P(A/X = x)f_X(x)dx$$

and, by observing that $\int_{-\infty}^{+\infty} f_X(x/A)dx = 1$, we have

$$P(A) = \int_{-\infty}^{+\infty} P(A/X = x)f_X(x)dx \qquad (58)$$

This is the Total probability law for the continuous case.

Bayes' formula for the continuous case

Furthermore from (57), and using (58), we obtain:

$$f_X(x/A) = \frac{P(A/X = x) f_X(x)}{P(A)} = \frac{P(A/X = x) f_X(x)}{\int_{-\infty}^{+\infty} P(A/X = x) f_X(x) dx} \quad (59)$$

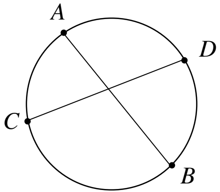
which represents the Bayes theorem extended to the continuous. case.

Example (18)

Four points A, B, C and D are chosen uniformly and independently on a circumference. Find the probability of event $I = \{intersection\ of\ chords\ \ AB\ and\ \ CD\}$.

Denoted by L the length of the circumference and by x the RV length of arc $\widehat{AB}$(oriented), and assumed $X = x$, we have

$$P(I/X = x) = P(D \in \widehat{AB})P(C \in \widehat{BA}) + P(D \in \widehat{BA})P(C \in \widehat{AB}) =$$
$$= 2\frac{x(L-x)}{L^2}$$



Example (18)

From total probability law, and being

$$f_X(x) = \frac{1}{L}, \quad (0 < x < L),$$

we get

$$P(I) = \int_0^L P(I/X = x) f_X(x) dx = \int_0^L 2\frac{x(L-x)}{L^3} \, dx = \frac{1}{3}$$

The result can be found also observing that, once A is taken, the sequences derived from the permutations of the other 3 points are equally likely, and among these only two lead to a chord intersection.

Multiple Random Variables

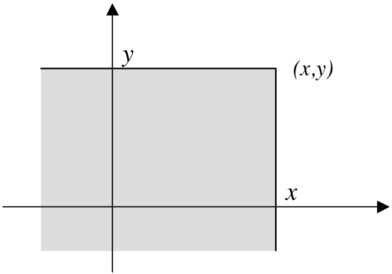
- We can extend definitions of a scalar RV (defined on real numbers R) to the case of multiple RV's defined on multidimensional spaces
- We focus on the case of two RVs, being the extension to more than two RVs straightforward
- Consider two RV $X(\alpha)$ and $Y(\alpha)$ defined in the same result space S
- We have a correspondence between each event $A \subset S$ and a set D_{xy} of the Cartesian plane, such that for every $\alpha \in A$ the point with coordinates $X(\alpha)$ and $Y(\alpha)$ belongs to D_{xy}.
- Thus, a joint event in S is represented by a domain D_{xy} in the Cartesian plane.

Joint CDF

The probability of the joint events $\{X \leq x, Y \leq y\} = \{X \leq x\}\{Y \leq y\}$ is a function of the pair of real variables x and y:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) \tag{60}$$

Such a function is called *joint CDF* of RVs X and Y.



Joint CDF

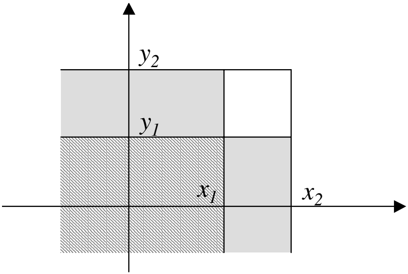
From the definition we can easily verify the following relations:

$$F(x,\infty) = F_X(x); \qquad F(\infty, y) = F_Y(y) \tag{61}$$

$$F(\infty,\infty) = 1 \tag{62}$$

$$F(x,-\infty) = 0; \qquad F(-\infty, y) = 0 \tag{63}$$

$$P(x_1 < X \le x_2, y_1 < Y \le y_2) = F(x_2,y_2) - F(x_1,y_2) - F(x_2,y_1) + F(x_1,y_1) \tag{64}$$



Joint pdf

Assuming now that $F_{XY}(x,y)$ is derivable, the joint pdf of RVs X and Y is

$$f_{XY}(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} \tag{65}$$

The properties hold

$$f(x,y) \geq 0 \tag{66}$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dx dy = 1 \tag{67}$$

Furthermore, from the definition of the joint derivative we have

$$f(x,y) = \lim_{\Delta x, \Delta y \to 0} \frac{P(x < X < x + \Delta x, y < Y < y + \Delta y)}{\Delta x \Delta y} \tag{68}$$

Joint pdf

The event including all results where $X(\alpha)$ and $Y(\alpha)$ belong to a domain D can be written as a union or intersection of elementary events of the type

$$\{x < X \leq x + \Delta x,\ y < Y \leq y + \Delta y\}$$

and, therefore, we have

$$P((X,Y) \in D) = \int\int_D f(x,y)dxdy \tag{69}$$

where the integral is extended over the domain D.
It also follows that

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy; \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx \tag{70}$$

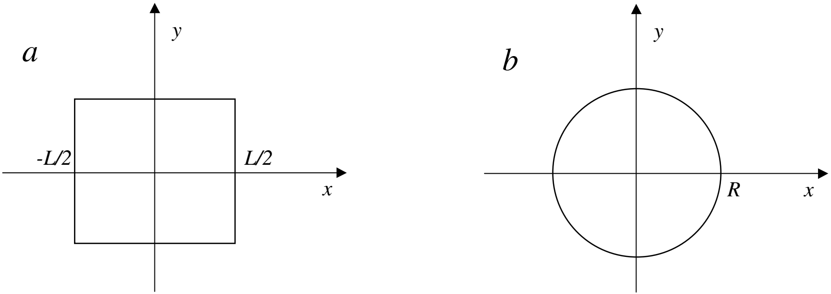
When dealing with multiple RVs, the pdf of each RV is called *marginal*.

Example (19)

Find the joint and marginal pdf of RV's X and Y Cartesian coordinates of a point Q chosen uniformly in a

a) *square of side L and centered at the origin*

b) *circle of radius R and center at the origin*



Example (19)

To find the joint density we use definition (68).
In this expression the probability at the numerator the probability Q lies into the rectangle of coordinates x, $x + \Delta x$, y, $y + \Delta y$, but since Q is picked uniformly, this probability has value $\frac{\Delta x \Delta y}{S}$, S being the area of the domain, regardless of the location of the small rectangle. Therefore, we obtain

$$f(x,y) = \begin{cases} \frac{1}{S} & \text{for}(x,y) \in S \\ 0 & \text{elsewhere} \end{cases} \quad (71)$$

Such a pdf is still called Uniform in S and the value of the constant $1/S$ depends only from the area of the domain and not by its shape.

Example (19)

About the marginal pdf we have

a)

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \int_{-\frac{L}{2}}^{\frac{L}{2}} \frac{1}{L^2}dy = \frac{1}{L}; \qquad \left(-\frac{L}{2} < x < \frac{L}{2}\right)$$

and similarly

$$f_Y(y) = \frac{1}{L}; \qquad \left(-\frac{L}{2} < y < \frac{L}{2}\right)$$

In this case, the marginal pdf are uniform.

b)

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \frac{1}{\pi R^2}dy = \frac{2}{\pi R^2}\sqrt{R^2-x^2}; \ (|\ x\ |< R)$$

Here, the marginal pdf's are no longer uniform. In fact, the shape of the domain of point (X,Y) influences the result.

Conditional pdf

- pdf of RV Y conditioned by the value assumed by another RV X

$$f_Y(y/X=x) = \frac{f_{XY}(x,y)}{f_X(x)} \tag{72}$$

- Total Probability Theorem

$$f_Y(y) = \int_{-\infty}^{+\infty} f_Y(y/X=x) f_X(x) dx \tag{73}$$

- Bayes' Theorem

$$f_Y(y/X=x) = \frac{f_X(x/Y=y) f_Y(y)}{f_X(x)} \tag{74}$$

Conditional pdf

- Conditional mean

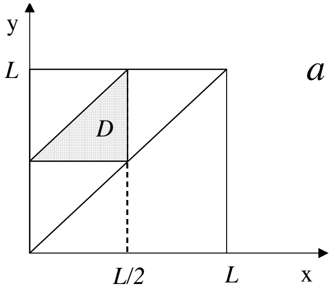
$$E[Y/X = x] = \int_{-\infty}^{+\infty} y f_Y(y/X = x) \qquad (75)$$

- Total Probability Theorem with respect to the mean

$$E[Y] = \int_{-\infty}^{+\infty} E[Y/x] f_X(x) dx \qquad (76)$$

Example (20)

A point of coordinate X is uniformly selected within interval $[0; L]$ of x axis; Another point of coordinate Y is uniformly selected within interval $[X; L]$. Find the joint pdf of X, Y, the marginal pdf of Y and the probability P that the three segments of length X, Y, and $Y - X$ can form a triangle.



Example (20)

We have

$$f_X(x) = \frac{1}{L}; \qquad (0 < x < L)$$

$$f_Y(y/X = x) = \frac{1}{L - x}; \qquad (x < y < L)$$

Using (72) we get

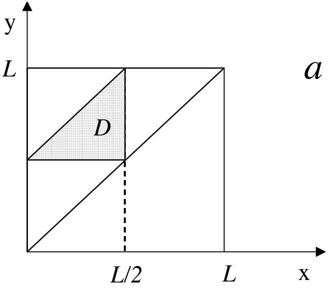
$$f_{XY}(x, y) = \frac{1}{L(L - x)}; \qquad (0 < x < y < L)$$

and from (73), by observing that the expression under integration is zero for $x > y$, we have

$$f_Y(y) = \int_0^y \frac{1}{L(L - x)} dx = \frac{1}{L} ln \frac{L}{L - y}; \qquad (0 < y < L)$$

Example (20)

The domain D, where X and Y are such as to allow the construction of the triangle, is shown in the figure



and we thus have

$$p = \int_0^{\frac{L}{2}} dx \int_{\frac{L}{2}}^{x+\frac{L}{2}} \frac{1}{L(L-x)} dy = ln2 - \frac{1}{2} = 0,1931\ldots$$

Statistically independent RVs

Two RV X and Y are said to be statistically independent if events $\{X \leq x\}$ e $\{Y \leq y\}$ are statistically independent for each x and y.
It follows then that two random RV are independent if one of the following relations holds

$$
\begin{aligned}
F_{XY}(x,y) &= F_X(x)F_Y(y) \\
f_{XY}(x,y) &= f_X(x)f_Y(y) \\
f_X(x/Y=y) &= f_X(x) \\
f_Y(y/X=x) &= f_Y(y)
\end{aligned}
$$

Example (21)

Given two RV's X and Y independent and exponentially distributed with the same average $1/\lambda$, find:

a) *the probability of the event $\{Y > \alpha X\}$ with α real positive;*

b) *the pdf $f_Y(y/Y > \alpha X)$.*

a) We could use the (69), being D the domain in which $y > \alpha x$, and given the independence we have

$$f_{XY}(x,y) = f_X(x) f_Y(y) = \lambda^2 e^{-\lambda(x+y)} \quad (x, y > 0)$$

More immediately we can use the Total Probability Theorem

$$\begin{aligned} P(Y > \alpha X) &= \int_0^\infty P(Y > \alpha X / X = x) f_X(x) dx = \int_0^\infty e^{-\lambda \alpha x} \; \lambda e^{-\lambda x} dx \\ &= \frac{1}{\alpha + 1} \int_0^\infty \lambda(\alpha+1) e^{-\lambda(\alpha+1)x} dx = \frac{1}{(\alpha+1)} \end{aligned}$$

Example (21)

b) From the definition of conditional pdf, and from the result of point a) we get:

$$
\begin{aligned}
f_Y(y/Y > \alpha X) &= \lim_{\Delta y \to 0} \frac{1}{\Delta y} \frac{P(y < Y \le y + \Delta y, Y > \alpha X)}{P(Y > \alpha X)} = \\
&= \lim_{\Delta y \to 0} \frac{1}{\Delta y} \frac{P(y < Y \le y + \Delta y, X < y/\alpha)}{P(Y > \alpha X)} = \\
&= \frac{\int_0^{y/\alpha} f_{XY}(x, y) dx}{P(Y > \alpha X)} = \frac{\lambda e^{-\lambda y} \int_0^{y/\alpha} \lambda e^{-\lambda x} dx}{1/(\alpha + 1)} = \\
&= (\alpha + 1)\lambda e^{-\lambda y}(1 - e^{-\lambda y/\alpha}) \quad (y > 0)
\end{aligned}
$$

Joint Moments

Given two RV's X and Y the joint moments of order h and k are defined as

$$m_{hk} = \int\int x^h y^k f_{xy}(x,y)dxdy$$

and the central moments of order h and k

$$\mu_{hk} = \int\int (x - m_x)^h (y - m_y)^k f_{xy}(x,y)dxdy.$$

The mixed second-order central moment μ_{11}, said also **Covariance** of RVs X and Y, is of particular interest. It is linked to m_{11} by the following relation

$$\mu_{11} = m_{11} - m_{10}m_{01} \tag{77}$$

The sum of two continuous RV's

Given the two continuous RV's X e Y, whose joint pdf is known, we want to find the pdf of their sum

$$Z = X + Y \tag{78}$$

To this purpose, we note that

$$f_Z(z/X = x) = f_Y(z - x/X = x) \tag{79}$$

From the total probability theorem we have

$$f_Z(z) = \int f_Z(z/X = x) f_X(x) dx = \int f_Y(z - x/X = x) f_X(x) dx \tag{80}$$

which provides the final formula

$$f_Z(z) = \int f_{XY}(x, z - x) dx \tag{81}$$

Symmetrically we have

$$f_Z(z) = \int f_{XY}(z - y, y) dy \tag{82}$$

The sum of two continuous RV's

If X and Y are statistically independent the two above become

$$f_Z(z) = \int f_X(x) f_Y(z - x) dx$$

$$f_Z(z) = \int f_X(z - y) f_Y(y) dy$$

The operations above are known as the convolution of pdf's.
In fact, the convolution of functions $f(x)$ and $g(y)$ (need not to be pdf's) is defined as

$$f(z) * g(z) = \int f(x) g(z - x) dx = \int f(z - x) g(x) dx$$

Example (22)

Find the pdf of RV $Z = X + Y$ where X and Y are independent RVs with the same pdf, namely

a) $f(x) = \frac{1}{a} \quad (0 < x < a)$
b) $f(x) = \lambda e^{-\lambda x} \quad (x > 0)$

a) The integrating function in (90) is different from zero when both the following conditions apply:

$$\begin{cases} 0 < x < a \\ 0 < z - x < a \end{cases}$$

or

$$\begin{cases} 0 < x < a \\ z - a < x < z \end{cases}$$

Example (22)

Such conditions depend on z and, therefore, we must distinguish the following cases:

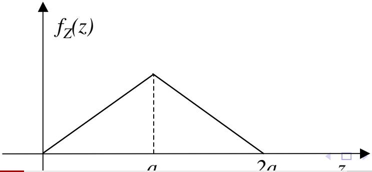
- for $z \leq 0$ $\quad f_Z(z) = 0$
- for $0 \leq z < a$ $\quad$ condition $0 < x < z$ holds, and therefore we have

$$f_Z(z) = \frac{1}{a^2} \int_0^z dz = \frac{z}{a^2};$$

- for $a < z \leq 2a$ condition $z - a < x < a$ holds, and therefore we have

$$f_Z(z) = \frac{1}{a^2} \int_{z-a}^a dx = \frac{2-z}{a^2};$$

- for $z > 2a$ $\quad f_Z(z) = 0;$



Example (22)

b) The integrating function in (90) is different from zero when $\begin{cases} x > 0 \\ z - x > 0 \end{cases}$ that is $\begin{cases} x > 0 \\ x < z \end{cases}$ and, therefore, we have

$$f_Z(z) = \int_0^z \lambda e^{-\lambda x} \lambda e^{-\lambda(z-x)} dx = \lambda^2 z e^{-\lambda z} \quad (z > 0)$$