# Probability and Random Signals I: Class Notes for the Course ECSE 305

Benoit Champagne
Department of Electrical & Computer Engineering
McGill University, Montreal, Quebec, Canada

Fall 2003

# Chapter 1

# Introduction

## 1.1 Randomness versus determinism

*Determinism in science and engineering:*

- Deterministic view in science: provided sufficient information is available about the initial state and operating conditions of a natural process or a man-made system, its future behavior can be predicted exactly.
- This operational viewpoint has been the prevailing one in most of your college and university education (mechanics, circuit theory, etc.)
- A typical example is provided by classical mechanics:
  - Consider the motion of a particle under the influence of various forces in three-dimensional space.
  - If we know the initial position and velocity vectors of the particle, its mass and the total force field, Newton's laws can be used to calculate (i.e. predict) the future trajectory of the particle.

*The concept of randomness:*

- The above view is highly idealistic: In most "real-life" scientific and engineering problems, as well as many other situations of interest (e.g. games of chance), we cannot do exact predictions about the phenomena or systems under consideration.
- Two basic reasons for this may be identified:
  - we do not have sufficient knowledge of the initial state of the system or the operating conditions (e.g. motion of electrons in a microprocessor circuit).
  - due to fundamental physical limitations, it is impossible to make exact predictions (e.g. uncertainty principle in quantum physics)
- We refer to such phenomena or systems as random, in the sense that there is uncertainty about their future behavior: a particular result or situation may or may not occur.
- The observation of specific quantities derived from such a random system or phenomenon is often referred to as a random experiment.

*Examples:*

- Consider the following game of chance:
  - We roll an ordinary six-sided die once and observe the number showing up, also called outcome.
  - Possible outcomes are represented by the set of numbers $S = \{1, 2, 3, 4, 5, 6\}$.
  - We cannot predict what number will show up as a result of this experiment.
  - Neither can we predict that a related event $A$, such as obtaining an even number (represented by $A = \{2, 4, 6\}$), will occur.
- Consider a more sophisticated example from communications engineering:
  - Using an appropriate modulation scheme, we transmit an analog speech signal $s(t)$ over a radio channel.
  - Due to channel and receiver noise, and other possible disturbances during the transmission, the received signal $r(t)$ is generally different from the transmitted one, i.e. $s(t)$.
  - In general, it is not possible for the radio engineer to predict the exact shape of the error signal $n(t) = r(t) - s(t)$.

## 1.2 The object of probability

*Regularity in randomness:*

- OK, we cannot predict with certainty the particular outcome in a single realization of a random experiment, but...
- In many practical situations of interest (games of chance, digital communications, etc..), it has been observed that when a random experiment is repeated a large number of times, the sequence of results so obtained shows a high degree of *regularity*.
- Let us be more specific: Suppose we repeat a random experiment (e.g. rolling a die) $n$ times. Let $\eta(A, n)$ be the number of times that a certain event $A$ occur (e.g. the result is even). It has been observed that

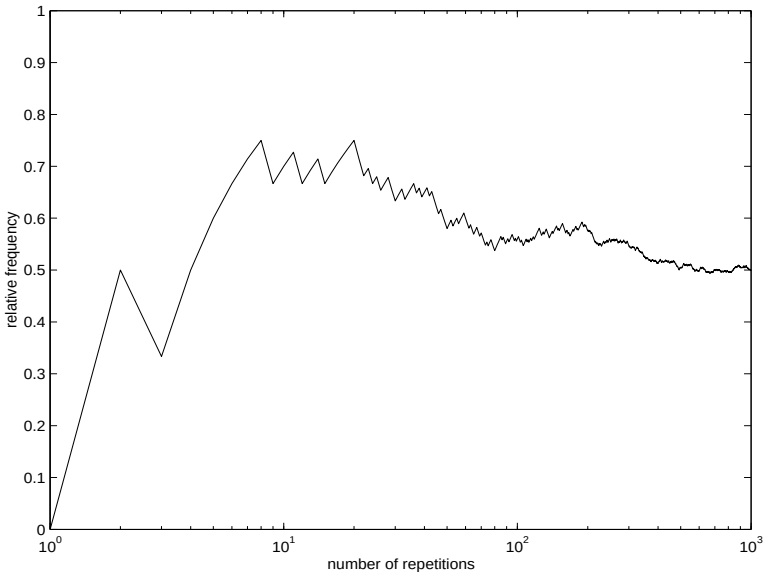
$$\frac{\eta(A, n)}{n} \to \text{constant} \quad \text{as} \quad n \to \infty \tag{1.1}$$

- The ratio $\eta(A, n)/n$ is called the relative frequency.
- The constant provides a quantitative measure of the likelihood of $A$.

*Example:*

- Consider a simple experiment consisting in flipping a coin. Let $A$ denote the event that a head shows up.
- Now suppose this experiment is repeated $n$ times. The quantity $\eta(A, n)$ simply represents the number of times that head shows up, out of $n$ similar flips of the coin.
- Assuming that the coin is fair (unloaded), we expect the ratio $\eta(A, n)/n$ to approach $1/2$ as $n$ gets larger and larger.
- Results of a computer simulation experiment:
  - Sequence of observed outcomes: THTHHH...
  - Relative frequency versus $n$:




  - Example:

$$\frac{\eta(A, 3)}{3} = \frac{1}{3} = 0.33$$

*Goal of probability theory:*

- To provide quantitative measures of the likelihood of various events. These measures will be called probabilities.
- To provide standard mathematical models for the efficient description and computation of such probabilities.
- To provide the tools and techniques necessary for computing the probability of more complex events, and related quantities, from the probabilities of simpler events (deductive theory).
- To provide fundamental insight, mathematical formalism, and general guidance about certain more philosophical aspects and questions of the theory. For example, why do relative frequencies converge?

*Notes on applications:*

- Probability theory finds applications in almost every branches of natural and social sciences as well as engineering: mathematical statistics, physical sciences, computer sciences, essentially all fields of engineering (electrical, mechanical, industrial, etc.), economy and finance, behavioral sciences, epidemiology, etc.
- In electrical and computer engineering: Probability theory is very useful in the study of systems or problems involving the manipulation of large quantities of data having a random nature: rate of failures in microprocessor production, performance of digital radio receiver, throughput of communication network, data compression algorithms, etc.
- Whenever we want to analyze or design such engineering systems, probability theory can provide extremely valuable information.
- In fact, in many situations of interest, probability is the only reliable and practical tool available for the study of this type of systems.

## 1.3 Approaches to probability

Probability draws its origins in the games of chance and specially in the development of approaches and strategies for maximizing the odd of winning in such games. Over the last 500 years, various definitions of probability, and eventually probability theories, have evolved. Some of the most well known definitions are the following:

- classical approach
- relative frequency
- axiomatic approach

*Classical approach (Laplace 1812):*

- Consider a random experiment in which the set $S$ of possible outcomes is finite, containing $N$ elements.
- Suppose that an event of interest to us, say $A$ a subset of $S$, contains $N_A$ elements.
- In the classical approach, the probability of $A$ is defined as

$$P(A) = \frac{N_A}{N} \tag{1.2}$$

- Example: $A$ = even number showing up when rolling a die once
  $S = \{1, 2, 3, 4, 5, 6\} \Rightarrow N = 6$
  $A = \{2, 4, 6\} \Rightarrow N_A = 3$
  $P(A) = 3/6 = 1/2.$
- Problems with classical approach:
  - too restrictive ($S$ must be finite)
  - definition of elements may lead to ambiguity

*Relative frequency (von Mises 1919):*

- Suppose that we can repeat the random experiment an $\infty$ of times.
- Define the probability of event $A$ as

$$P(A) = \lim_{n\to\infty} \frac{\eta(A,n)}{n} \tag{1.3}$$

- Problems:
  - don't know if the limit exists?
  - even if limit exists, cannot repeat experiment an $\infty$ of times?
  - what is the error introduced if a large, but finite number of expereements is used in computing $P(A)$?

*Axiomatic approach (Kolmogorov, 1933):*

- We only require that the function $P(A)$ satisfy a minimal number of axioms, from which more complex probabilities (and related quantities) may be computed in a systematic manner.
- Simplified version of the axioms:

  A1. For any event $A$, $P(A) \geq 0$

  A2. Let $S$ denote the set of all possible outcomes, then $P(S) = 1$

  A3. If events $A$ and $B$ cannot occur simultaneously, then $P(A \text{ or } B) = P(A) + P(B)$
- The theory is developed in a rigorous and systematic way around this irreducible set of axioms.
- As long as the axioms are satisfied, the definition and interpretation of the function $P(A)$ in a particular application are left to the user.
- This is by far the most commonly used theory nowadays: It provides fundamental justification for the classical and relative frequency approaches.
- The axiomatic approach is the one used in this course.

# Chapter 2

# Background material

*Chapter overview:*

- Review of set theory
- Combinatorial methods

## 2.1 Set theory

### 2.1.1 Basic terminology

*Definition of a set:*

- A set is a collection of objects (concrete or abstract), called elements, that usually share some common attributes, but are not otherwise restricted in any fashion.
- The curly brackets { and } are used as delimiters when specifying the content of a set. This may be achieved by either listing all the elements of the set explicitly, as is

$$\{1, 2, 3, 4, 5, 6\} \tag{2.1}$$

  or by stating the common properties satisfied by its elements, as in

$$\{a : a \text{ is a positive integer } \leq 6\} \tag{2.2}$$

  In the latter case, the notation "$a$ :" should read "all $a$ such that".
- To indicate that an object $a$ is an element of a set $A$, we write $a \in A$; we also say that $a$ is a member of, or belongs to, $A$. If $a$ is not an element of $A$, we write $a \notin A$.
- Two sets $A$ and $B$ are identical (or equal) if and only if (iff) they have the same elements, in which case we write $A = B$. If $A$ and $B$ are not identical, we write $A \neq B$.
- Example: Let $A = \{1, 2, ..., 6\}$ and $B = \{2, 4, 6\}$. Then $A \neq B$ because $1 \in A$ while $1 \notin B$.

*Subset:*

- If every element of a set $A$ is also an element of a set $B$, we say that $A$ is contained in $B$, or that $A$ is a subset of $B$, and write $A \subseteq B$.
- If $A$ is a subset of $B$ but there exists $b$ such that $b \in B$ and $b \notin A$, we sometimes say that $A$ is a proper subset of $B$ and write $A \subset B$.
- The negations of the set relations $\subseteq$ and $\subset$ are denoted by $\not\subseteq$ and $\not\subset$, respectively.
- Example: let $A = \{1, 2, ..., 6\}$, $B = \{2, 4, 6\}$ and $C = \{0, 1\}$, then $B \subseteq A$, $B \subset A$, $C \not\subseteq A$, etc.

*Sample space and empty set:*

- In practical applications of set theory, all sets of interest in a given situation are usually subsets of a larger set called sample space, or universal set, and denoted by the letter $S$.
- It is also common practice to introduce a degenerate set containing no elements; the latter is called the empty set, or null set, and is denoted by the symbol $\emptyset$.

*Theorem 2.1:* Let $A$, $B$ and $C$ denote arbitrary subsets of a sample space $S$. The following relations hold:

(a) $A \subseteq A$

(b) $A \subseteq B$ and $B \subseteq C$ implies $A \subseteq C$

(c) $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$

(d) $\emptyset \subseteq A \subseteq S$

*Proof:* These basic properties follow directly from the preceding definitions; their proof is left as an exercise to the reader. $\square$

*Commonly used sets of numbers:*

- Basic sets of numbers:
  - Positive integers, or natural numbers: $\mathbb{N} = \{1, 2, 3, ...\}$
  - Integers: $\mathbb{Z} = \{0, \pm 1, \pm 2, ...\}$
  - Rational numbers: $\mathbb{Q} = \{\frac{a}{b} : a, b \in \mathbb{Z} \text{ and } b \neq 0\}$
  - Real numbers: $\mathbb{R}$
  - Complex numbers: $\mathbb{C} = \{a + jb : a, b \in \mathbb{R}\}$, where $j = \sqrt{-1}$

  Note that $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$.

- Let $a$ and $b$ be two arbitrary real numbers. The following subsets of $\mathbb{R}$ are called intervals from $a$ to $b$:
  - Open interval: $(a, b) = \{x \in \mathbb{R} : a < x < b\}$
  - Closed interval: $[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$
  - Left semi-open interval: $(a, b] = \{x \in \mathbb{R} : a < x \leq b\}$
  - Right semi-open interval: $[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$

  Note that these intervals are empty, i.e. identical to the empty set $\emptyset$, when $a > b$.

*Finite versus infinite sets:*

- A set is called finite if it is empty or contains a finite number of elements; otherwise it is called infinite.
- A set is called countable if it is finite (countably finite) or if it is infinite but can be put into a one-to-one correspondence with the set of positive integers $\mathbb{N}$ (countably infinite). In the latter case, the elements of the set can be indexed sequentially.
- A set that is not countable is said to be uncountable or uncountably infinite.
- Examples:
  - The set $S = \{1, 2, 3, 4, 5, 6\}$ is countably finite.
  - Examples of countably infinite sets include $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$.
  - Examples of uncountably infinite sets include $\mathbb{R}$ and the open intervals $(a, b)$ for any $a < b$ in $\mathbb{R}$.

*Product sets:*

- Let $A$ and $B$ be two arbitrary sets, not necessarily associated to the same sample space. The product set of $A$ and $B$, denoted $A \times B$, is the set of all ordered pairs $(a, b)$ such that $a \in A$ and $b \in B$. That is,

$$A \times B = \{(a, b) : a \in A \text{ and } b \in B\} \tag{2.3}$$

[1]

- The product set $A \times A$ is also denoted in a more compact form as $A^2$.
- The generalization of this concept to products of more than two sets is immediate.
- As an example, the notation $\mathbb{R}^n$, for $n$ a positive integer, denotes the set of all $n$-tuples $(a_1, a_2, ..., a_n)$ where $a_i \in \mathbb{R}$ for $i = 1, ..., n$.

*Example 2.1:*

► (a) Consider a single toss of a coin. The set of all possible observable results, or outcomes, can be described as

$$S_1 = \{H, T\}$$

where $H$ denotes heads and $T$ denotes tails.

(b) Consider two consecutive tosses of a coin. The set of all possible outcomes is

$$S_2 = \{HH, HT, TH, TT\}$$

where, for example, the ordered sequence $HT$ corresponds to $H$ on the first toss and $T$ on the second toss. Observe that

$$S_2 = \{H, T\} \times \{H, T\} = S_1^2$$

The event that at least one head is observed can be represented by the subset

$$A = \{HH, HT, TH\} \subset S_2$$

◄

[1] test

### 2.1.2 Set operations

*Definitions:* Let $A$ and $B$ be arbitrary subsets of a sample space $S$. We define the following operations:

- The *union* of the sets $A$ and $B$, denoted $A \cup B$, is the set of all elements that belong to at least one of the sets $A$ or $B$:

$$A \cup B = \{x \in S : x \in A \text{ or } x \in B\} \tag{2.4}$$

- The *intersection* of the sets $A$ and $B$, denoted $A \cap B$, is the set of all elements that belong to both $A$ and $B$:

$$A \cap B = AB = \{x \in S : x \in A \text{ and } x \in B\} \tag{2.5}$$

- The *complement* of the set $A$, denoted $A^c$, is the set of all elements of $S$ that do not belong to $A$:

$$A^c = \{x \in S : x \notin A\} \tag{2.6}$$

- The *difference* of the sets $A$ and $B$, denoted $A - B$, is the set of all elements of $A$ that do not belong to $B$:

$$A - B = \{x \in S : x \in A \text{ and } x \notin B\} \tag{2.7}$$

*Remarks:*

- In the above definition of the union, the "or" is a logical one, meaning that $x$ may be in $A$ or $B$ or both.
- In the probability literature, the symbol $\cap$ for the intersection is sometimes omitted, so that the notations $A \cap B$ and $AB$ are equivalent.
- If $A \cap B = \emptyset$, we say that $A$ and $B$ are mutually exclusive (or disjoint).
- Finally, note that $A^c = S - A$ and $A - B = A \cap B^c$.

*Theorem 2.2:* Let $A$, $B$ and $C$ be arbitrary subsets of a sample space $S$. The following identities hold:

(a) Basic identities:

$$A \cup A = A \quad \text{and} \quad A \cap A = A \tag{2.8}$$

$$A \cup S = S \quad \text{and} \quad A \cap S = A \tag{2.9}$$

$$A \cup \emptyset = A \quad \text{and} \quad A \cap \emptyset = \emptyset \tag{2.10}$$

(b) Commutative laws:

$$A \cup B = B \cup A \quad \text{and} \quad A \cap B = B \cap A \tag{2.11}$$

(c) Associative laws:

$$(A \cup B) \cup C = A \cup (B \cup C) \quad \text{and} \quad (A \cap B) \cap C = A \cap (B \cap C) \tag{2.12}$$

(d) Distributive laws:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad \text{and} \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \tag{2.13}$$

(e) Complementarity laws:

$$A \cup A^c = S \quad \text{and} \quad A \cap A^c = \emptyset \tag{2.14}$$

$$(A^c)^c = A, \quad S^c = \emptyset \quad \text{and} \quad \emptyset^c = S \tag{2.15}$$

(f) DeMorgan's laws:

$$(A \cup B)^c = A^c \cap B^c \quad \text{and} \quad (A \cap B)^c = A^c \cup B^c \tag{2.16}$$

*Proof:* The proof of such properties is tedious but otherwise straightforward. For any of the above equalities, we have to show set inclusion in both direction, that is: any arbitrary element of the set on the left-hand side is also an element of the set on the right-hand side, and vice versa. This requires the use and manipulation of logical assertions and operators. Familiarity with the latter concepts is assumed.

As an example, consider the DeMorgan's identity $(A \cap B)^c = A^c \cup B^c$. We have

$$\begin{aligned} x \in (A \cap B)^c &\iff x \notin A \cap B \\ &\iff x \notin A \text{ or } x \notin B \\ &\iff x \in A^c \text{ or } x \in B^c \\ &\iff x \in A^c \cup B^c \end{aligned}$$

The other identities may be proved in a similar way; this is left as an exercise for the reader. □

*Example 2.2:*

► A die is rolled once. The set of possible outcomes is

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

Define the subsets

$$A = \{x \in S : x \leq 3\} = \{1, 2, 3\}$$
$$B = \{x \in S : x \text{ even } \} = \{2, 4, 6\}$$

We have:

$$\begin{aligned} A \cap B &= \{2\} \\ A \cup B &= \{1, 2, 3, 4, 6\} \\ A^c &= \{4, 5, 6\} \\ B^c &= \{1, 3, 5\} \end{aligned}$$

Let us verify DeMorgan's Laws, i.e. Theorem 2.2 (f). From the above, we have

$$\begin{aligned} (A \cup B)^c &= \{5\} \\ A^c \cap B^c &= \{5\} \end{aligned}$$

which shows that the first identity in (2.16) is satisfied. In the same way,

$$\begin{aligned} (A \cap B)^c &= \{1, 3, 4, 5, 6\} \\ A^c \cup B^c &= \{1, 3, 4, 5, 6\} \end{aligned}$$

which shows the validity of the second identity in (2.16). ◄

*Venn diagrams:*

- Venn diagrams provide a useful mechanism for visualizing various set-theoretic operations.
- Basic idea:
  - represent sets as planar areas delimited by closed contours;
  - these contours are included in a larger rectangular area representing the sample space $S$ itself;
  - an operation between various sets is shown as a shaded area.
- This is illustrated in Figure 2.1 for the following operations: $A \cup B$, $A \cap B$, $A^c$ and $A - B$.

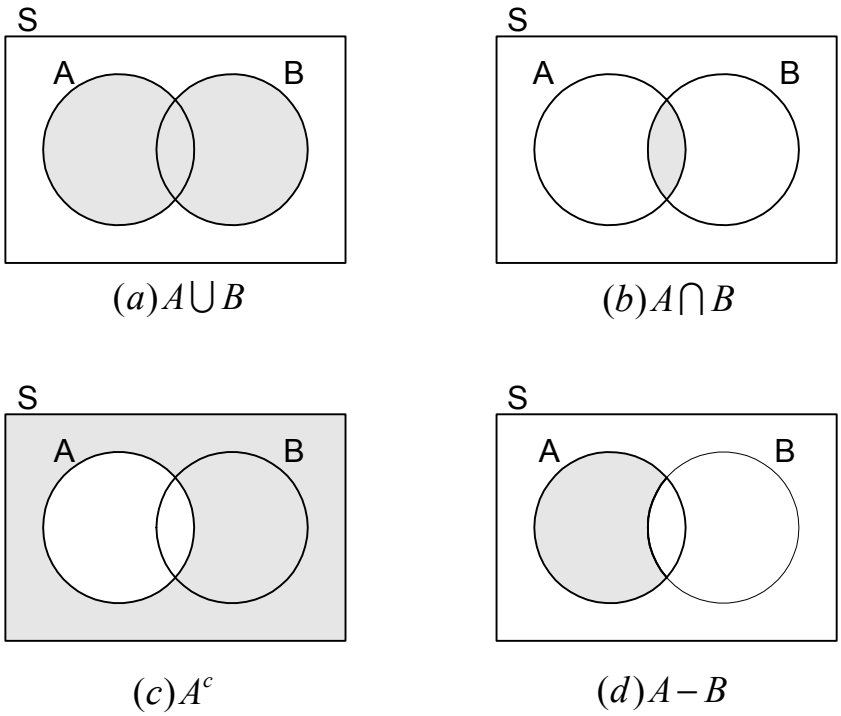



Figure 2.1: Use of Venn diagrams to illustrate set operations.

- Venn diagrams are often used as an intuitive device for gaining insight into complex set relations and operations, although their use in the formal proof of set properties is not quite appropriate.
- As an example, the following theorem may be easily justified on the basis of Venn diagrams.

*Theorem 2.3:* Let $A$ and $B$ be arbitrary subsets of a sample space $S$. Anyone of the following conditions is equivalent to the inclusion $A \subseteq B$:

(a) $A \cap B = A$

(b) $A \cup B = B$

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

(d) $A^c \cup B = S$

(e) $B^c \subseteq A^c$

*Justification based on Venn diagrams:* A Venn diagram's interpretation of Theorem 2.3 (a) and (b) is illustrated in Figure 2.2.

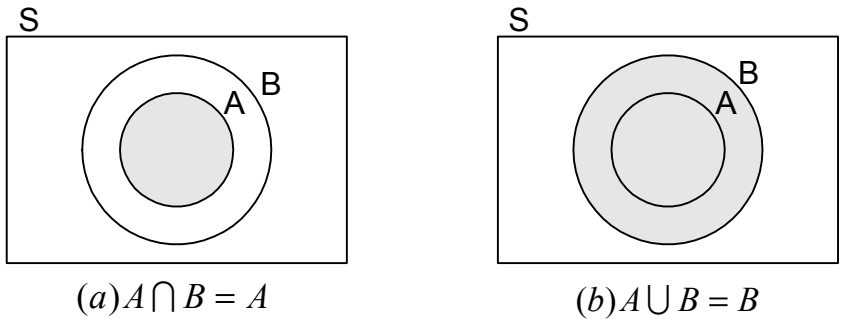



Figure 2.2: Interpretation of Theorem 2.3 based on Venn diagrams.

*Some generalizations:*

- Consider a sequence of indexed subsets of $S$, say $A_i$ where the index $i \in I$, with $I$ being a subset (finite or infinite) of the natural numbers $\mathbb{N}$.
- The union and intersection of the sets $A_i$, $i \in I$, are defined as

$$\bigcup_{i \in I} A_i = \{x \in S : x \in A_i \text{ for some } i \in I\} \tag{2.17}$$

$$\bigcap_{i \in I} A_i = \{x \in S : x \in A_i \text{ for all } i \in I\} \tag{2.18}$$

  When $I = \mathbb{N}$, these may be denoted as $\bigcup_{i=1}^{\infty}$ and $\bigcap_{i=1}^{\infty}$, respectively.
- De Morgan's laws admit immediate generalization to this case:

$$(\bigcup_{i \in I} A_i)^c = \bigcap_{i \in I} A_i^c \quad \text{and} \quad (\bigcap_{i \in I} A_i)^c = \bigcup_{i \in I} A_i^c \tag{2.19}$$

- We say that the sequence $A_i, i \in \mathbb{N}$, is increasing if $A_1 \subseteq A_2 \subseteq A_3 \subseteq ...$ In this case, we define

$$\lim_{i \to \infty} A_i = \bigcup_{i=1}^{\infty} A_i \tag{2.20}$$

- We say that the sequence $A_i, i \in \mathbb{N}$, is decreasing if $A_1 \supseteq A_2 \supseteq A_3 \supseteq ...$ In this case, we define

$$\lim_{i \to \infty} A_i = \bigcap_{i=1}^{\infty} A_i \tag{2.21}$$

*Example 2.3:*

► Consider the real plane, $S = \mathbb{R}^2$. Define $A_i$ as the subsets of all points in on or inside a circle of radius $i$ centered at the origin, where $i$ is a positive integer. That is

$$A_i = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq i^2\}, \quad i \in \mathbb{N}$$

Observe that

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$$

so that the sequence $A_i$ is increasing. This is illustrated in Figure 2.3.

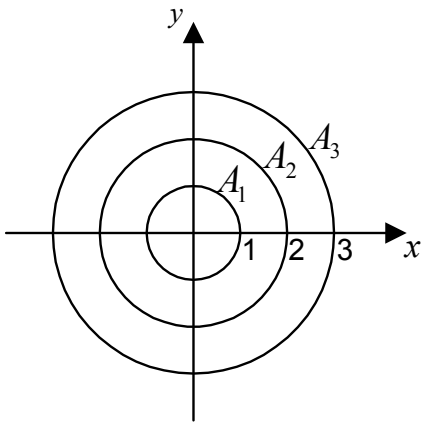



Figure 2.3: An increasing sequence of subsets.

Also note that (try to prove it)

$$\bigcup_{i=1}^{\infty} A_i = \{(x, y) \in \mathbb{R}^2 : (x, y) \in A_i \text{ for some } i\} = \mathbb{R}^2$$

Therefore,

$$\lim_{i \to \infty} A_i = \mathbb{R}^2$$

◄

### 2.1.3 Sets of sets

The elements of a set may themselves be sets. Such sets of sets are often called *classes* or *families* of sets. Sets having for elements subsets of a sample space $S$ play a central role in probability. Below, we develop these concepts.

*Power set:*

- The set of all the subsets of a set $S$ is called the *power set* of $S$ and is denoted by $\mathcal{P}_S$, or simply $\mathcal{P}$.
- Since $\emptyset \subseteq S$ and $S \subseteq S$, we have by definition of $\mathcal{P}_S$ that $\emptyset \in \mathcal{P}_S$ and $S \in \mathcal{P}_S$.
- For example, let $S = \{0, 1\}$. Then $\mathcal{P}_S = \{\emptyset, \{0\}, \{1\}, S\}$

*Remarks:*

- When the sample space $S$ is uncountably infinite (e.g. $S = \mathbb{R}$), the power set $\mathcal{P}_S$ will typically contain undesirable subsets that pose serious mathematical difficulties.
- In such situations, it is usually desirable to work with a much smaller subset of $\mathcal{P}_S$, that do not include the undesirable subsets.
- This leads to the notion of set algebras.

*Set algebra:* Let $\mathcal{F}$ be a set of subsets of $S$, that is $\mathcal{F} \subseteq \mathcal{P}_S$. We say that $\mathcal{F}$ is an algebra iff

(a) $S \in \mathcal{F}$

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

(c) $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$

*Remarks:*

- From (a) and (b), it follows that $\emptyset \in \mathcal{F}$ since $\emptyset = S^c$.
- According to the above definition, the algebra $\mathcal{F}$ is closed under the operations of complementation and union.
- Using DeMorgan's laws, you should be able to show that $\mathcal{F}$ is also closed under the operation of intersection, that is: $A, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}$.

*Example 2.4:*

▶ Let $S = \{0, 1\}$. The corresponding power set is

$$\mathcal{P}_S = \{\emptyset, \{0\}, \{1\}, S\}.$$

It is easy to check that $\mathcal{P}_S$ is an algebra...

(a) $S \in \mathcal{P}_S$

(b) For any $A \in \mathcal{P}_S$, we have $A^c \in \mathcal{P}$. For example:

$$\emptyset^c = S \in \mathcal{P}_S, \quad \{0\}^c = \{1\} \in \mathcal{P}_S, \quad \text{etc.}$$

(c) For any $A, B \in \mathcal{P}_S$, we have $A \cup B \in \mathcal{P}_S$. For example:

$$\emptyset \cup \emptyset = \emptyset \in \mathcal{P}_S, \quad \emptyset \cup \{0\} = \{0\} \in \mathcal{P}_S, \quad \text{etc.}$$

Note that there are 10 such identities to check.

◀

*Sigma algebra:*

- In probability theory, a more specific type of algebra, called sigma-algebra, or simply $\sigma$-algebra, is employed.
- The latter is defined as above, but condition (c) is replaced by:

  (c') If the sets $A_1, A_2, A_3, ...$ belong to $\mathcal{F}$, so does their union $\cup_{i=1}^{\infty} A_i$.
- The use of an infinite sequence in (c') ensures that $\mathcal{F}$ is closed under any countable combination of complement, union and intersection operations.
- Note that any finite algebra (i.e. with $\mathcal{F}$ containing a finite number of elements) is also a $\sigma$-algebra.

## 2.2 Combinatorial analysis

*What is combinatorial analysis?*

- Part of mathematics dealing with the study and development of systematic methods for counting.
- Find applications in many areas of sciences and engineering: probability and statistics, information theory, data compression, genetics, etc.
- The calculation of probabilities often amounts to counting the number of elements in various sets. Combinatorial techniques will be of great help in the solution of these problems.

### 2.2.1 Basic counting techniques

*r-tuples:*

- Let $r$ be a positive integer.
- A $r$-tuple is an ordered list (or vector) of elements, of the form $(x_1, x_2, ..., x_r)$, or simply $x_1x_2...x_r$ (when there is no ambiguity).
- Two $r$-tuples are equal ($=$) if and only if each of the corresponding elements are identical.

*Theorem 2.4 (Generalized counting principle):* Let $A$ be a set of $r$-tuples, $x_1x_2...x_r$, such that there are, firstly, $n_1$ different ways in which to chose $x_1$, secondly, $n_2$ different ways in which to chose $x_2$, ... and finally, $n_r$ different ways in which to chose $x_r$. Then $A$ contains

$$N(A) = n_1n_2...n_r \tag{2.22}$$

different $r$-tuples.

*Remarks:*

- Theorem 2.4, which is to some extent obvious, can be proven by mathematical induction (left as an exercise).
- The theorem specifies only the number of possible choices that are available at each step: the specific choices in the $r$th step may depend on previous choices, but not their number $n_r$.

*Example 2.5:*

► In Quebec, license plate numbers are made up of 3 letters followed by 3 digits, that is $l_1 l_2 l_3 d_1 d_2 d_3$ where $l_i$ is any one of 26 possible letters from a to z, and $d_i$ is any one of the possible digits from 0 to 9. Thus there are, in principle,

$$26 \times 26 \times 26 \times 10 \times 10 \times 10 = 26^3 \times 10^3 = 17,576,000 \tag{2.23}$$

different license plate numbers. ◄

*Corollary:* Suppose the sets $A_1, A_2, ..., A_r$ contain $n_1, n_2, ..., n_r$ elements, respectively. Then the product set

$$A_1 \times A_2 \times ... \times A_r = \{(a_1, a_2, ..., a_r) : a_i \in A_i\} \tag{2.24}$$

contains $n_1 n_2 ... n_r$ elements.

*Remarks:*

- This result is an immediate consequence of Theorem 1.
- In particular, if $A$ contains $n$ elements, then $A^r$ contains $n^r$ elements.

*Theorem 2.5:* A set $S$ containing $n$ elements has $2^n$ different subsets, or equivalently, its power set $\mathcal{P}_S$ contains $2^n$ elements.

*Proof:* Let $S = \{s_1, s_2, ..., s_n\}$. The essence of the proof is to realize that every subset $A$ of $S$ may be represented uniquely by a binary sequence of length $N$, say $b_1 b_2 ... b_n$, where for $i = 1, ..., n$, we have $b_i = 1$ if $s_i \in A$ and $b_i = 0$ if $s_i \notin A$. The number of subset of $S$ is therefore equal to the number of binary sequences $b_1 b_2 ... b_n \in \{0, 1\}^n$, which is equal to $2^n$. □

*Example 2.6:*

▶ Consider a set $S$ with two elements, say $S = \{a, b\}$. The basic idea used in the above proof is illustrated in the table below:

| Subset | Binary representation |
|---|---|
| $\emptyset$ | 00 |
| $\{a\}$ | 10 |
| $\{b\}$ | 01 |
| $S = \{a, b\}$ | 11 |

In this case, we have $2^2 = 4$ subsets in $\mathcal{P}_S$. ◀

*Tree diagrams:*

- Useful when counting principle does not apply directly.
- For example, when the number of ways of selecting a second element depends on the choice made for the first element, and so on.
- Tree diagram provides systematic identification of all possibilities.

*Example 2.7:*

▶ In a certain binary coding scheme, individual pieces of information (e.g. letters, digits, etc.) are represented by specific sequences of 0s and 1s, called codewords.

List all possible codewords that terminate upon the occurrence of symbol 0 or a maximum of 3 bits?

◀

### 2.2.2 Permutations

*Definition:* An ordered arrangement of $r$ elements taken without replacement from a set $A$ containing $n$ elements ($0 < r \leq n$) is called an $r$-element permutation of $A$. The number of such permutations is denoted $P(n, r)$.

*Example 2.8:*

▶ Consider the set $A = \{a, b, c\}$. All the possible 2-element permutations of $A$ are:

$$ab, ac, ba, bc, ca, cb$$

The number of these permutations is $P(3, 2) = 6$. ◀

*Remarks:*

- Repetitions are not allowed in a permutation. In the above example, once $a$ has been selected as the first element, the remaining choices for the second element are $b$ or $c$.
- A permutation is an ordered arrangement of $r$ elements, i.e. an $r$-tuple. Thus the order does matter: $ab \neq ba$

*Theorem 2.6:* The number of $r$-element permutations of a set $A$ containing $n$ elements is given by the product

$$P(n,r) = n\,(n-1)...(n-r+1) \tag{2.25}$$

*Proof:* Observe the following:

- there are $n$ ways in which to chose the 1st element, leaving us with $n-1$ remaining elements;
- there are $n-1$ ways in which to chose the 2nd element leaving us with $n-2$ remaining elements;...
- and finally, there are $n-r+1$ ways in which to chose the $r$th element.

Therefore, according to Theorem 2.4, there are $n(n-1)...(n-r+1)$ ways of forming all the possible permutations. □

*Factorial notation:*

- For any positive integer $n$, we define

$$n! = n(n-1)(n-2)...1 \tag{2.26}$$

  It is also convenient to define $0! = 1$.

- Alternatively, factorials may be defined (and computed) recursively as $n! = n\,(n-1)!$, with initial condition $0! = 1$.

- Factorials grow surprisingly fast: $10! = 3628800$, $20! \approx 2.4329 \times 10^{18}$, etc.

- For large values of $n$, may use Stirling's approximation:

$$n! \approx \sqrt{2\pi} n^{n+1/2} e^{-n} \tag{2.27}$$

- Useful to express $P(n,r)$ in terms of factorials:

$$\boxed{P(n,r) = \frac{n!}{(n-r)!}} \tag{2.28}$$

*Example 2.9:*

► How many different words can we form: (a) with the 4 letters P H I L; (b) with the 6 letters P H I L I P?

(a) Since the 4 letters P H I L are different, the number of different words is equal to the number of 4-element permutations of these letters, that is

$$P(4,4) = 4! = 24$$

(b) First assume that the 2 P's and 2 I's are distinct, as in P H I L I' P'. The number of possible permutations of these 6 "different" letters is

$$P(6,6) = 6! = 720$$

Clearly, some of these permutations result in the same word. For instance:

$$\text{P H I L I' P'} = \text{P' H I L I' P}$$

Specifically, we note that there are

$$2! \text{ permutations of the letters P P'}$$
$$2! \text{ permutations of the letters I I'}$$

Taking this into account, the number of different words that can be formed with the letters P H I L I P is

$$\frac{6!}{2!2!} = \frac{720}{4} = 180$$

◄

### 2.2.3 Combinations

*Definition:* An *unordered* arrangement of $r$ objects taken without replacement from a set $A$ containing $n$ elements ($0 < r \leq n$) is called an $r$-element combination of $A$. The number of such combinations is denoted $C(n, r)$.

*Example 2.10:*

▶ Consider the set $A = \{a, b, c\}$. All the possible combinations of the elements of $A$ taken 2 at a time are:

$$ab, ac, bc \tag{2.29}$$

Thus the number of such combinations is $C(3, 2) = 3$. ◀

*Remarks:*

- As in the case of permutations, repetitions are not allowed.
- Contrary to permutations, order does not matter: $ab$ and $ba$ are counted as one combination.
- Except for the absence of curly brackets (and commas), an $r$-element combination of $A$ is the same as an $r$-element subset of $A$.

*Theorem 2.8:* The number of $r$-element combinations of a set $A$ containing $n$ elements, is given by

$$C(n,r) = \frac{n!}{(n-r)!\,r!} \tag{2.30}$$

*Proof:* Simply observe that every $r$-element permutation of $A$ can be obtained by first selecting an $r$-element combination and then permuting the $r$ selected elements. Therefore, according to the basic counting principle:

$$C(n,r) \times r! = P(n,r) = \frac{n!}{(n-r)!} \tag{2.31}$$

from which the desired result follows. □

*Corollary:* A set $S$ containing $n$ elements has $C(n,r) = n!/(n-r)!\,r!$ different subsets of size $r$.

*Proof:* Simply recall that an $r$-element combination and an $r$-element subset of a set $S$ are conceptually equivalent.

*Definition:* For any integers $r$ and $n$, with $0 \leq r \leq n$, we define:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = C(n,r) \tag{2.32}$$

The expression $\binom{n}{r}$ (read "$n$ choose $r$") is also called binomial coefficient.

*Theorem 2.9:* The binomial coefficients satisfy the following relations:

$$\binom{n}{0} = \binom{n}{n} = 1 \tag{2.33}$$

$$\binom{n}{r} = \binom{n}{n-r} \tag{2.34}$$

$$\binom{n+1}{r} = \binom{n}{r} + \binom{n}{r-1} \tag{2.35}$$

*Proof:* Left as exercise.

*Note:* Can you give an intuitive interpretation of (2.35) and (2.36)?

*Example 2.11:*

► In a 6/49 lottery, players pick 6 different integers between 1 and 49, without repetition, the order of the selection being irrelevant. The lottery commission then selects 6 winning numbers in the same manner. A player wins the first prize if his/her selection matches the 6 winning numbers. The player wins the second prize if exactly 5 of his/her chosen numbers match the winning selection. How many different winning combinations are there?

- 1st prize: Here, the player selection must be identical to that of the lottery commission. There is only one possible way of doing that.
- 2nd prize: Here, there are $\binom{6}{5} = 6$ ways of selecting 5 numbers out of the 6 winning numbers. There are also $\binom{43}{1} = 43$ ways of choosing one number out of the 49-6=43 non-winning numbers. Thus, the number of different combinations leading to a 2nd prize is

$$6 \times 43 = 257$$

◄

### 2.2.4 Sampling problems

*Motivation:*

- Many counting problems can be interpreted as sampling problems, in which objects are selected from a population.
- Below, we define four types of sampling problems and for each one, we provide a general counting formula.
- In all these cases, a selection of $r$ objects from a population is made. The latter is represented by a set $A$ initially containing $n$ distinct objects.

*Sampling with replacement and with ordering:*

- After selecting an object from $A$ and noting its identity in an ordered list, the object is put back into $A$.
- This corresponds to the basic counting situation (Theorem 2.5). Thus, the number of distinct ordered lists is

$$N_1(n, r) = n^r \tag{2.36}$$

*Sampling without replacement and with ordering:*

- After selecting an object from $A$ and noting its identity in an ordered list, the object is discarded $A$.
- The number of distinct lists is equal to the number of $r$-element permutations from set $A$. Therefore (Theorem 2.6), we have

$$N_2(n,r) = P(n,r) = \frac{n!}{(n-r)!} \tag{2.37}$$

*Sampling without replacement and without ordering:*

- After selecting an object from $A$ and noting its identity in a non-ordered list, the object is discarded.
- The number of distinct lists is equal to the number of $r$-element combinations from set $A$. Therefore (Theorem 2.8), we have

$$N_3(n,r) = \binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{2.38}$$

*Sampling with replacement and without ordering:*

- After selecting an object from $A$ and noting its identity in a non-ordered list, the object is put back in $A$.
- In order to count the number of possibilities, we need to specify the way in which the observations are recorded.
- The standard approach consists in listing for each object how many times it is selected.
- For example, suppose $n = 6$ and $r = 5$. A possible observation is then $(3, 0, 0, 1, 0, 1)$, which can also be represented as

$$xxx| \ | \ |x| \ |x$$

- The number of distinct possible observations (or lists) is equal to the number of distinguishable permutations of $n + r - 1$ objects of two different types, of which $r$ are alike (the $x$'s) and $n - 1$ are alike (the |'s).
- Therefore (Theorem 2.7), we have

$$N_4(n, r) = \frac{(n + r - 1)!}{r!(n - 1)!} \tag{2.39}$$

### 2.2.5 Miscellaneous results

*Theorem 2.10 (binomial expansion):* For any integer $n \geq 0$,

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^{n-i} y^i \tag{2.40}$$

*Proof:* By induction on $n$ (left as an exercise).

*Example 2.12:*

- Suppose set $S$ contains $n$ elements. We can use (2.41) to show that $\mathcal{P}_S$, the Power set of $S$, contains $2^n$ elements (see also Theorem 2.5):
  - Power set = set of all subsets of $S$
  - A subset of $S$ may contains $r$ elements, with $0 \leq r \leq n$
  - Number of $r$-element subsets: $C(n,r) = \binom{n}{r}$
  - Total number of subsets:

$$\binom{n}{0} + \binom{n}{1} + ... + \binom{n}{n-1} + \binom{n}{n} = (1+1)^n = 2^n \tag{2.41}$$

◀

*Theorem 2.11 (multinomial expansion):* For any integer $n \geq 0$,

$$(x_1 + x_2 + ... + x_k)^n = \sum_{n_1+n_2+...+n_k=n} \frac{n!}{n_1!\, n_2! \,...\, n_k!}\, x_1^{n_1}\, x_2^{n_2} \,...\, x_k^{n_k} \tag{2.42}$$

## Problems

1. Suppose that $A \subseteq B$, show that $A \cap (B - A) = \emptyset$.

2. It is a tradition for *business men* in North America to shake hands prior to a meeting. In a meeting involving $n$ so-called business men, how many handshakes will be exchanged?

3. How many different decimal numbers of 4 digits can we form that contain at least one 7?

4. 10 persons are waiting at an airport counter, of whom 5 are from Canada and five are from France. How many different line-ups can be formed so that no two persons from the same country are next to each other?

5. Use mathematical induction to proof Theorem 2.10.

# Chapter 3

# Axiomatic approach to probability

*Chapter Overview:*

- Axioms of probability and terminolgy
- Basic probability theorems
- Special cases of probability space:
  - Discrete (finite and countably infinite)
  - Continuous (uncountably infinite)

## 3.1 Axioms of probability

*Random experiment:*

- An experiment, either natural or man-made, in which one among several identified results are possible, is called a *random experiment*.
- The possible results of the experiments are called *outcomes*.
- A particular realization of the experiment, leading to a particular outcome, is called a *trial*.

*Probability space:*

- In the axiomatic approach to probability, a random experiment is modeled as a *probability space*, the latter being a triplet $(S, \mathcal{F}, P)$, where
  - $S$ is the sample space,
  - $\mathcal{F}$ is the set of events (events algebra),
  - $P(.)$ is the probability function.
- These concepts are described individually below.

*Sample space:*

- The sample space $S$ is the set of all possible results, or outcomes, of the random experiment.
- In practical applications, $S$ is defined by the very nature of the problem under consideration. $S$ may be finite, countably infinite or uncountably infinite.
- The elements of $S$, i.e. the experimental outcomes, will usually be denoted by lower case letters (e.g.: $s, a, x$, etc...)

*Example 3.1:*

▶ Consider a random experiment that consists in flipping a coin twice. A suitable sample space may be defined as

$$S = \{HH, HT, TH, TT\}$$

where, for example, outcome $HT$ corresponds to heads on the first toss and tails on the second. Here, $S$ is finite with only 4 outcomes. ◀

*Events:*

- In probability theory, an event $A$ is defined as a subset of $S$, i.e. $A \subseteq S$.
- Referring to a particular trial of the random experiment, we say that $A$ *occurs* if the experimental outcome $s \in A$.
- Special events $S$ and $\emptyset$:
  - Since for any outcome $s$, we have $s \in S$ by definition, $S$ always occurs and is thus called the certain event.
  - Since for any outcome $s$, we have $s \notin \emptyset$, $\emptyset$ never occurs and is thus called the impossible event.

*Example 3.1 (continued):*

▶ Consider the event $A = \{\text{getting heads on the first flip}\}$. This can equivalently be represented by the following subset of $S$:

$$A = \{HH, HT\} \subset S$$

Let $s$ denote the outcome of a particular trial:

$$\begin{aligned} \text{if } s = HH \text{ or } HT &\Rightarrow A \text{ occurs} \\ \text{if } s = TH \text{ or } TT &\Rightarrow A \text{ does not occur} \end{aligned}$$

◀

*Events algebra:*

- Let $\mathcal{F}$ denote the set of all events under consideration in a given random experiment. Note that $\mathcal{F}$ is a set of subsets of $S$
- Clearly:
  - $\mathcal{F}$ must be large enough to contain all interesting events,
  - but not so large as to contain impractical events that lead to mathematical difficulties. (This may be the case when $S$ is uncountably infinite, e.g. $S = \mathbb{R}^n$.)
- In the axiomatic approach to probability, it is required that $\mathcal{F}$ be a $\sigma$-algebra:
  (a) $S \in \mathcal{F}$
  (b) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
  (c) $A_1, A_2, ... \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$
- Whenever $S$ is finite, the simplest and most appropriate choice for $\mathcal{F}$ is generally the power set $\mathcal{P}_S$.
- The proper choice for $\mathcal{F}$ when $S$ in infinite will be discussed later.

*Example 3.1 (continued):*

▶ Consider flipping a coin twice and let $S = \{HH, HT, TH, TT\}$ be the corresponding sample space. An appropriate choice for $\mathcal{F}$ here is $\mathcal{P}_S$, i.e. the set of all subsets of $S$:

$$\begin{aligned}\mathcal{P}_S \;=\; & \{\emptyset, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \{HH, TH\}, \\ & \{HH, TT\}, \{HT, TH\}, \{HT, TT\}, \{TH, TT\}, \{HH, HT, TH\}, \\ & \{HH, HT, TT\}, \{HH, TH, TT\}, \{HT, TH, TT\}, S\}\end{aligned}$$

Note that $\mathcal{F} = \mathcal{P}_S$ contains $16 = 2^4$ different subsets, i.e. events, that may or may not occur during a particular realization of the random experiment. For example, the event $\{HH, HT, TH\} \in \mathcal{F}$ corresponds to obtaining at least one heads when you flip the coin twice.

If you think about it, each event corresponds to a specific statement about the experimental outcome and here, there are only 16 possible different statements of this type that can be made. ◀

*The probability function:*

- $P$ is a function that maps events $A$ in $\mathcal{F}$ into real numbers in $\mathbb{R}$, that is:

$$P : A \in \mathcal{F} \rightarrow P(A) \in \mathbb{R} \tag{3.1}$$

  The number $P(A)$ is called the probability of the event $A$.

- The function $P(.)$ must satisfy the following **axioms**:

  Axiom 1: The function $P$ is non-negative:

$$P(A) \geq 0 \tag{3.2}$$

  Axiom 2: The function $P$ is normalized so that

$$P(S) = 1 \tag{3.3}$$

  Axiom 3: Let $A_1, A_2, A_3, ...$ be a sequence of mutually exclusive events, that is, $A_i \cap A_j = \emptyset$ for $i \neq j$. Then

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \tag{3.4}$$

*Remarks:*

- From an operational viewpoint, the number $P(A)$ may be interpreted as a measure of the likelihood of event $A$ in a particular realization of the random experement.
- If $P(A) = P(B)$, we say that events $A$ and $B$ are equally likely (this does NOT imply that $A = B$).
- As a special case of Axiom 3, it follows that for any events $A$ and $B$,

$$A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B) \tag{3.5}$$

- In the special case of a finite sample space $S$, it can be shown that (3.5) is in fact equivalent to Axiom 3. Thus, when $S$ is finite, we may replace Axiom 3 (infinite additivity) by the simpler condition (3.5).

*Example 3.1 (continued):*

▶ Let the function $P$ be defined as follows, for any $A \in \mathcal{F}$:

$$P(A) \triangleq \frac{N(A)}{4}$$

where $N(A)$ denotes the number of elements in subset $A$. For example, consider event $A = \{\text{at least on tails}\}$; we have

$$\begin{aligned} A = \{TH, HT, TT\} \quad &\Rightarrow \quad N(A) = 3 \\ &\Rightarrow \quad P(A) = \frac{3}{4} \end{aligned}$$

It can be verified easily that function $P$ satisfies all the axioms of probability:

- *Axiom 1:* For any event $A$, $N(A) \geq 0$ and therefore, $P(A) = N(A)/4 \geq 0$.
- *Axiom 2:* Since $N(S) = 4$, we immediately obtain $P(S) = N(S)/4 = 1$.
- *Axiom 3:* Observe that if $A \cap B = \emptyset$, then $N(A \cup B) = N(A) + N(B)$ and therefore

$$\begin{aligned} P(A \cup B) &= \frac{N(A \cup B)}{4} \\ &= \frac{N(A)}{4} + \frac{N(A)}{4} = P(A) + P(B) \end{aligned}$$

◀

## 3.2 Basic theorems

*Introduction:* Several basic properties follow from the axiomatic definition of the probability function $P(A)$. These are listed below as theorems along with their proof.

*Theorem 3.1:* For any event $A \in \mathcal{F}$:

$$P(A^c) = 1 - P(A) \tag{3.6}$$

*Proof:* Observe that $A \cap A^c = \emptyset$ and $A \cup A^c = S$. Thus, using Axiom 3, we have: $P(A) + P(A^c) = P(A \cup A^c) = P(S) = 1$, or equivalently, $P(A^c) = 1 - P(A)$. $\square$

*Corollary:* For any event $A \in \mathcal{F}$:

$$0 \leq P(A) \leq 1 \tag{3.7}$$

*Proof:* Left as exercise. $\square$

*Theorem 3.2:*

$$P(\emptyset) = 0. \tag{3.8}$$

*Proof:* Observe that $\emptyset = S^c$. Thus, invoking Theorem 3.1 and Axiom 2, we have: $P(\emptyset) = P(S^c) = 1 - P(S) = 0$ . $\square$

*Theorem 3.3:* If $A \subseteq B$, then

$$(a) \qquad P(B - A) = P(B) - P(A) \tag{3.9}$$

$$(b) \qquad P(A) \leq P(B) \tag{3.10}$$

*Proof:* Since $A \subseteq B$, set $B$ may be expressed as the union $B = A \cup (B - A)$ where $A$ and $B - A$ are mutually exclusive, that is $A \cap (B - A) = \emptyset$. The Venn diagram below illustrates this situation:

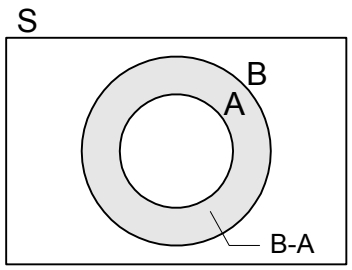



Figure 3.1: Venn diagram for Theorem 3.3.

Using axiom 3, we have

$$P(B) = P(A \cup (B - A)) = P(A) + P(B - A) \tag{3.11}$$

which proves part (a). To prove part (b), simply note (see Axiom 1) that $P(B - A) \geq 0$. □

*Theorem 3.4:* For arbitrary events $A$ and $B$, we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3.12}$$

*Proof:* Observe that for any events $A$ and $B$, we can always write

$$A \cup B = A \cup (B - (A \cap B)) \tag{3.13}$$

where $A$ and $B-(A\cap B)$ are mutually exclusive. This is illustrated by means of a Venn diagram below:

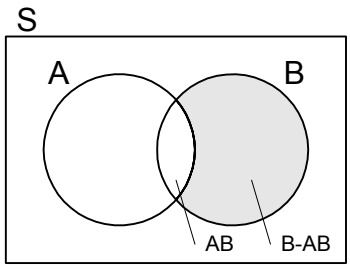



Figure 3.2: Venn diagram for Theorem 3.4. (Note: $AB \equiv A \cap B$.)

Invoking Axiom 3, we first obtain

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

Since $A \cap B \subseteq B$, Theorem 3.3 yields

$$P(B - (A \cap B)) = P(B) - P(A \cap B)$$

Eq. (3.12) follows by combining the above two identities. $\square$

*Remarks:*

- Theorem 3.4 may be generalized to a union of more than two events.
- In the case of three events, say $A$, $B$ and $C$, the following relation can be derived

$$P(A \cup B \cup C) = P(A)+P(B)+P(C)-P(AB)-P(AC)-P(BC)+P(ABC). \tag{3.14}$$

- The above formula can be proved by repeated application of Theorem 3.4. This is left as an exercise.
- See the textbook for a more general formula applicable to a union of $n$ events, where $n$ is an arbitrary positive integer.

*Theorem 3.5:* For any events $A$ and $B$:

$$P(A) = P(A \cap B) + P(A \cap B^c). \tag{3.15}$$

*Proof:* The theorem follows from Axiom 3 by noting that $A \cap B$ and $A \cap B^c$ are mutually exclusive and that their union is equal to $A$ (see Fig. 3.3). □

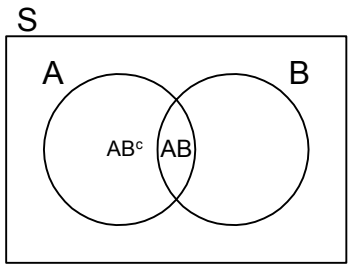



Figure 3.3: Venn diagram for Theorem 3.5.

*Example 3.2:*

► In a certain city, three daily newspapers are available, labelled here as A, B and C for simplicity. The probability that a randomly selected person reads newspaper A is $P(A) = .25$. Similarly, for newspapers $B$ and $C$, we have $P(B) = .20$ and $P(C) = .13$. The probability that a person reads both A and B is $P(AB) = P(A \cap B) = .1$. In the same way, $P(AC) = .08$, $P(BC) = .05$ and $P(ABC) = .04$.

(a) What is the probability that a randomly selected person does not read any of these three newspapers?

(b) What is the probability that this person reads only B, i.e. reads B but not A nor C?

◄

*Theorem 3.6:* For any increasing or decreasing sequence of events $A_1, A_2, A_3, ...$ we have

$$\lim_{i\to\infty} P(A_i) = P(\lim_{i\to\infty} A_i) \tag{3.16}$$

*Remarks:*

- Recall that a sequence $A_i, i \in \mathbb{N}$, is increasing if $A_1 \subseteq A_2 \subseteq A_3 \subseteq ...$, in which case we define $\lim_{i\to\infty} A_i = \bigcup_{i=1}^{\infty} A_i$.
- Similarly, a sequence $A_i, i \in \mathbb{N}$, is decreasing if $A_1 \supseteq A_2 \supseteq A_3 \supseteq ...$, in which case we define $\lim_{i\to\infty} A_i = \bigcap_{i=1}^{\infty} A_i$.
- Theorem 3.6 is essentially a statement about the continuity of the probability function $P$.
- Specifically, it says that under proper conditions on the sequence $A_i$ (i.e. increasing or decreasing), the limit operation in (3.16) can be passed inside the argument of $P(.)$.

*Proof (optional reading):* First consider the case of an increasing sequence, i.e. $A_1 \subseteq A_2 \subseteq A_3 \subseteq ...$ Define a new sequence of events as follows: $B_1 = A_1$ and $B_i = A_i - A_{i-1}$ for any integer $i \geq 2$. Note that the events $B_i$ so defined are mutually exclusive, i.e. $B_i \bigcap B_j = \emptyset$ if $i \neq j$. Furthermore, the following relations hold

$$\begin{aligned} \bigcup_{j=1}^{i} B_j &= A_i \\ \bigcup_{j=1}^{\infty} B_j &= \bigcup_{j=1}^{\infty} A_j \end{aligned}$$

Making use of above results together with Axiom 3, we first obtain:

$$P(\lim_{i\to\infty} A_i) = P(\bigcup_{j=1}^{\infty} A_j) = P(\bigcup_{j=1}^{\infty} B_j) = \sum_{j=1}^{\infty} P(B_j) \tag{3.17}$$

Finally, the infinite summation can be expressed in terms of limits as follows:

$$\sum_{j=1}^{\infty} P(B_j) = \lim_{i\to\infty} \sum_{j=1}^{i} P(B_j) = \lim_{i\to\infty} P(\bigcup_{j=1}^{i} B_j) = \lim_{i\to\infty} P(A_i) \tag{3.18}$$

A proof of (3.16) for decreasing sequences can be derived in a somewhat similar way. □

## 3.3 Discrete probability space

*Introduction:*

- In many applications of probability (games of chance, simple engineering problems, etc.), the sample space $S$ is either finite or countably infinite. The word discrete is used to describe anyone of these two situations.
- Specifically, we say that a probability space $(S, \mathcal{F}, P)$ is *discrete* whenever the sample space $S$ is finite or countably infinite.
- In this section, we discuss discrete spaces along with related special cases of interest.

### 3.3.1 Finite probability space

*Sample space:*

- The sample space $S$ is a finite set comprised of $N$ distinct elements:

$$S = \{s_1, s_2, ..., s_N\} \tag{3.19}$$

  where $N$ is a positive integer and $s_i$ denotes the $i$th possible outcome.

*Events algebra:*

- In the finite case, it is most convenient to take for events algebra the power set of the sample space $S$:

$$\begin{aligned} \mathcal{F} &= \mathcal{P}_S \\ &= \text{set of all subsets of } S \\ &= \{\emptyset, \{s_1\}, \{s_2\}, ..., \{s_N\}, \{s_1, s_2\}, \{s_1, s_3\}, ..., S\} \end{aligned} \tag{3.20}$$

- That is, the events algebra consists of all possible subsets of $S$. Indeed, in the finite case, it is usually not advantageous nor necessary to exclude certain subsets of $S$ from $\mathcal{F}$.

- Recall that $\mathcal{P}_S$, the power set of $S$, contains $2^N$ distinct elements (i.e. subsets). Thus, there are $2^N$ possible events or different statements that can be made about the experimental outcome.

*Probability function:*

- In the finite case, a standard way to define the probability function $P(.)$ is via the introduction of a *probability mass* $p_i$.
- To each $s_i \in S$, $i = 1, ..., N$, we associate a real number $p_i$, such that:

$$(a) \qquad p_i \geq 0, \quad i = 1, ..., N \tag{3.21}$$

$$(b) \qquad \sum_{i=1}^{N} p_i = 1 \tag{3.22}$$

- The probability of any event $A \in \mathcal{F}$ is then defined as

$$P(A) = \sum_{s_i \in A} p_i \tag{3.23}$$

  For example, if $A = \{s_1, s_4, s_6\}$, then $P(A) = p_1 + p_4 + p_6$.

- In particular, for the *elementary events* $\{s_i\}$, we have

$$P(\{s_i\}) = p_i, \quad i = 1, ..., N \tag{3.24}$$

*Axioms of probability:* It may be verified that the probability function $P(.)$ so defined satisfies the probability Axioms:

- Axiom 1: From the condition $p_i \geq 0$ in (3.21), it follows that

$$P(A) = \sum_{s_i \in A} p_i \geq 0$$

- Axiom 2: From condition (3.22), it follows that

$$P(S) = \sum_{i=1}^{N} p_i = 1$$

- Axiom 3: Suppose $A$ and $B$ have no common element (i.e. $A \cap B = \emptyset$), then we have

$$\begin{aligned} P(A \cup B) &= \sum_{s_i \in A \cup B} p_i \\ &= \sum_{s_i \in A} p_i + \sum_{s_i \in B} p_i = P(A) + P(B) \end{aligned}$$

*Example 3.3:*

► ◄

### 3.3.2 Equiprobable space

*Definition:*

- This is a special case of the finite probability space.
- We say that a probability space is equiprobable (also equilikely) if it is finite and the probability mass $p_i$ are all equal.

*The probability mass:*

- Let $N$ be the number of possible outcomes in the sample space $S$.
- Suppose that the numbers $p_i$ are all equal. Then, from condition (3.22), i.e. $\sum_{i=1}^{N} p_i = 1$, it follows that

$$P(\{s_i\}) = p_i = \frac{1}{N} \quad \text{for all } i = 1, ..., N \tag{3.25}$$

*Probability function:*

- Consider an arbitrary event $A \in \mathcal{F}$, containing $N(A)$ distinct elements. From (3.23) and (3.25), it follows that

$$\boxed{P(A) = \frac{N(A)}{N}} \tag{3.26}$$

*Remarks:*

- We say that the possible outcomes $s_i \in S$ are equally likely.
- Equation (3.26) corresponds to the classical definition of probability, as discussed in Chapter 1.
- In problem statements, the following standard terminology is used to indicate an equiprobable space:
  - random selection among $N$ possibilities;
  - a fair experiment
  - equiprobable or equilikely outcomes

*Example 3.4:*

▶ What is the probability of at least one 6 when rolling four fair dice? ◀

*Example 3.5: Standard birthday problem*

▶ What is the probability that at least two sutdents in a class of size $n$ have the same birtday? ◀

### 3.3.3 Countably infinite probability space

*Sample space:*

- The sample space $S$ is a countably infinite set represented as

$$S = \{s_1, s_2, s_3, ...\} \tag{3.27}$$

  where $s_i$, $i \in \mathbb{N}$, denotes the $i$th possible outcome.

- Example of countably infinite sets include $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$.

*Events algebra:*

- As in the finite case, it is usually most convenient to take as events algebra the power set of $S$:

$$\mathcal{F} = \mathcal{P}_S = \{A : A \subseteq S\} \tag{3.28}$$

- Observe that since $S$ is infinite, so is $\mathcal{F} = \mathcal{P}_S$ and thus the number of events under consideration is infinite.

- Some of these events are finite, such as the elementary events $\{s_i\}$ for $i \in \mathbb{N}$, while other are infinite, such as $S$ or, for example, $A = \{s_i : i \text{ is even } \} = \{s_2, s_4, s_6, ...\}$.

*Probability function:*

- Much the same way as in the finite case, the probability function $P(.)$ is defined via a probability mass $p_i$.
- To every $s_i \in S$, where $i$ now takes value in the set $\mathbb{N}$, we associate a real number $p_i$ such that:

$$(a) \qquad p_i \geq 0, \quad \text{for all } i \in \mathbb{N} \tag{3.29}$$

$$(b) \qquad \sum_{i=1}^{\infty} p_i = 1 \tag{3.30}$$

- The probability of any event $A \in \mathcal{F}$ is defined as

$$P(A) = \sum_{s_i \in A} p_i \tag{3.31}$$

  In particular, for any $i \in \mathbb{N}$, we have $P(\{s_i\}) = p_i$.

- It may be verified that the probability function $P(.)$ so defined satisfies all the probability Axioms.

*Remark:*

- The concept of an equiprobable space does not make sense here: If $p_i$ was constant, condition (3.30) could not be satisfied.

*Example 3.6:*

▶ Consider flipping a fair coin until heads is observed for the first time. What is the probability that the number of required flips is even?

*Solution:* ◀

## 3.4 Continuous probability space

*Introduction:*

- In many engineering applications of probability (e.g. design of a radio receiver, speech recognition system, image analysis, etc.) the sample space is uncountably infinite or, equivalently, continuous.
- We say that a probability space $(S, \mathcal{F}, P)$ is continuous whenever the sample space $S$ is uncountably infinite. The proper, formal mathematical treatment of this case is beyond the scope of this course.
- Here, we adopt an engineering approach, relying more on intuition than mathematical formalism. You will have to accept certain results and concepts without complete justification.
- Still, we try to explain some of the technical difficulties associated to continuous spaces and we describe some of the mathematical apparatus available to handle this situation.

### 3.4.1 One-dimensional (1D) continuous space

*Sample space:*

- $S$ is either the set of real numbers $\mathbb{R}$, or an interval thereof:

$$S = \mathbb{R} \quad \text{or} \quad S = (a, b) \subseteq \mathbb{R} \tag{3.32}$$

  where $a < b$ are real numbers.

- These are not the only possibilities but they cover most cases of interest.
- Note: the elements of $S$ cannot be counted.

*Example:*

- Waiting time of a person at a bus station.
- Analog voltage measurement on $\pm 5$ volts scale: $S = [-5, +5] \subset \mathbb{R}$
- The power dissipated in a resistor: $S = [0, \infty)$

*Events algebra:*

- In the continuous case, it is NOT convenient to take the power set of $S$ as events algebra, so: $\mathcal{F} \neq \mathcal{P}_S$:
- $\mathcal{P}_S$ includes some strange and complex subsets of $\mathbb{R}$ that are counter-intuitive, of no interest in engineering applications and pose serious mathematical difficulties.
- In practice, only those events that belong to the so-called Borel field of $S$, denoted $\mathcal{B}_S$, are included in the events algebra, that is

$$\mathcal{F} = \mathcal{B}_S \subset \mathcal{P}_S \tag{3.33}$$

- While $\mathcal{B}_S$ is smaller than $\mathcal{P}_S$, it contains all subsets of practical signif-icanve in applications of probability. This includes intervals of the real axis and various combinations thereof.
- See next page for additional explanations.

*Borel field (optional reading):*

- For simplicity, assume $S = \mathbb{R}$.
- Intervals from $\mathbb{R}$ may be combined via union, intersection and complementation to generate more complex subsets of $\mathbb{R}$.
- The Borel field of $\mathbb{R}$, denoted $\mathcal{B}_{\mathbb{R}}$ may be defined as the smallest $\sigma$-algebra that contains as elements all intervals of $\mathbb{R}$.
- For example, the following subsets of $\mathbb{R}$ all belong to $\mathcal{B}_{\mathbb{R}}$:
  - The intervals $(a, b)$, $[a, b)$, etc., with $a, b \in \mathbb{R}$.
  - Any subset of $\mathbb{R}$ obtained from such intervals via a countable number of union, intersection and/or complementation operations.
- Because the Borel field $\mathcal{B}_{\mathbb{R}}$ is made up of subsets of $\mathbb{R}$, it is a subset of the power set $\mathcal{P}_{\mathbb{R}}$. However, $\mathcal{B}_{\mathbb{R}}$ does not contain every subset of $\mathbb{R}$:

$$\mathcal{B}_{\mathbb{R}} \subset \mathcal{P}_{\mathbb{R}} \tag{3.34}$$

- The Borel field $\mathcal{B}_{\mathbb{R}}$ essentially contains those subsets of R which are meaningful from an application perspective. Other less interesting and problematic subsets are left out.
- Since $\mathcal{B}_{\mathbb{R}}$ is a $\sigma$-algebra, it can be used as an events algebra in a probability model.

*Probability function:*

- A standard way to define the probability function $P(.)$ is via a probability density $\rho(x)$.
- To each $x \in S \subseteq \mathbb{R}$, we associate a real number $\rho(x)$, such that:

$$(a) \qquad \rho(x) \geq 0, \quad \text{for all } x \in S \tag{3.35}$$

$$(b) \qquad \int_S \rho(x)dx = 1 \tag{3.36}$$

- The probability of any event $A \in \mathcal{F} = \mathcal{B}_S$ is then defined as

$$P(A) = \int_A \rho(x)dx \tag{3.37}$$

- It may be verified that the probability function $P(.)$ so defined satisfies the probability axioms A1, A2 and A3:
  - Axiom 1: From (3.37) and (3.35), it follows that

$$P(A) = \int_A \rho(x)\, dx \geq 0$$

  - Axiom 2: From (3.37) and (3.36), we have

$$P(S) = \int_S \rho(x)\, dx = 1$$

  - Axiom 3: Suppose $A \cap B = \emptyset$. Invoking basic properties of integration, we have

$$\begin{aligned} P(A \cup B) &= \int_{A\cup B} \rho(x)\, dx \\ &= \int_A \rho(x)dx + \int_B \rho(x)dx = P(A) + P(B) \end{aligned}$$

*Uniform probability space:*

- We say that a continuous 1D probability space is uniform if the sample space has finite length and the probability density $\rho(x)$ is constant. This is the simplest case of a 1D continuous probability space.

- The sample space $S$ is typically a bounded interval, as in $S = (a, b)$ or $S = [a, b]$ , where $a < b$ are bounded real numbers (i.e. $|a|, |b| < \infty$). It does not matter whether the interval $S$ is open, closed, or semi-open.

- Assuming that the function $\rho(x)$ is constant, it immediately follows from condition (3.36) that

$$\rho(x) = \frac{1}{b-a} \quad \text{for all } x \in (a, b) \tag{3.38}$$

- The probability function is easily obtained by inserting (3.38) into (3.37). Specifically, for any event $A \in \mathcal{F}$, we find:

$$P(A) = \frac{1}{b-a} \int_A dx = \frac{\text{length of } A}{b-a} \tag{3.39}$$

- The following special cases are of interest:

  - If $A$ is an interval of the type $A = (\alpha, \beta)$ contained in $S$, i.e. $a \leq \alpha \leq \beta \leq b$, then

$$P(A) = \frac{\beta - \alpha}{b-a} \tag{3.40}$$

  - For any $x \in S$, we have

$$P(\{x\}) = 0 \tag{3.41}$$

*Example 3.7:*

► Random selection of a point from the interval $[-1, 1]$... ◄

### 3.4.2 Continuous probability space in higher dimensions

In this section, we consider the generalization of the one-dimensional continuous probability space introduced in Section 3.4.1 to $n$ dimensions, where $n$ is a positive integer.

*Sample space:*

- The *sample space* is typically $\mathbb{R}^n$ or a subset thereof, i.e.: $S \subseteq \mathbb{R}^n$
- Examples include the plane $\mathbb{R}^2$, the three-dimensional space $\mathbb{R}^3$ or specific regions thereof (e.g. a delimited surface in $\mathbb{R}^2$ or volume in $\mathbb{R}^3$).

*Events algebra:*

- The standard choice is $\mathcal{F} = \mathcal{B}_S$, which contains all the subsets of practical interest in engineering applications.
- For example, if $S = \mathbb{R}^2$, the Borel field $\mathcal{B}_S$ will contain any geometrical region of practical interest within the real plane, such as:
  - points, lines, curves, and geometrically delimited areas.
  - other regions obtained from union, intersection and complentation of above regions.

*Probability function:*

- $P(.)$ may be defined via a probability density function $\rho(\mathbf{x})$, where $\mathbf{x} \in S \subseteq \mathbb{R}^n$ is now a vector (when $n \geq 2$).
- To each element $\mathbf{x}$ in $S$, we associate a real number $\rho(\mathbf{x})$, such that:

$$(a) \qquad \rho(\mathbf{x}) \geq 0, \quad \text{for all } \mathbf{x} \in S \tag{3.42}$$

$$(b) \qquad \int ... \int_S \rho(x)\, d\mathbf{x} = 1 \tag{3.43}$$

- The probability of any event $A \in \mathcal{F} = \mathcal{B}_S$ is then defined as

$$P(A) = \int ... \int_A \rho(\mathbf{x})\, d\mathbf{x} \tag{3.44}$$

- It may be verified that the probability function $P(.)$ so defined satisfies the probability Axioms.
- For now, we shall only consider a special case of (3.42)-(3.43) known as the uniform probability space.

*Uniform probability space:*

- Let $S \subseteq \mathbb{R}^n$. For any event $A \in \mathcal{B}_S$, we define

$$M(A) = \int ... \int_A d\mathbf{x} \tag{3.45}$$

  The number $M(A)$ $(0 \leq M(A) \leq \infty)$ is called the measure of $A$.

- A probability space is uniform (equilikely) if its sample space $S \subseteq \mathbb{R}^n$ has a finite measure, i.e. $M(S) < \infty$, and $\rho(\mathbf{x})$ is constant for all $\mathbf{x} \in S$.

- Suppose $p(\mathbf{x})$ is constant. Then, from (3.43), it follows that

$$\rho(\mathbf{x}) = \frac{1}{M(S)}, \quad \text{for all } \mathbf{x} \in S \tag{3.46}$$

- Consider an arbitrary event $A \in \mathcal{F}$ with measure $M(A)$. Using (3.44), (3.46) and (3.45), we obtain:

$$P(A) = \int ... \int_A \rho(\mathbf{x}) d\mathbf{x} = \frac{M(A)}{M(S)} \tag{3.47}$$

*Remarks:*

- For $n = 1, 2, 3$, the concept of measure admits an immediate physical interpretation:

  $A \subseteq \mathbb{R} \Rightarrow M(A) =$ length of $A$

  $A \subseteq \mathbb{R}^2 \Rightarrow M(A) =$ area of $A$

  $A \subseteq \mathbb{R}^3 \Rightarrow M(A) =$ volume of $S$

- In problem statements, look for:

  - random selection from
  - fair experiment
  - uniformly distributed outcomes

*Example 3.8:*

▶ Consider the random selection of two real numbers $x$ and $y$ from the interval $[0, 1]$. What is the probability that $x > 2y$? ◀

# Chapter 4

# Conditional Probability and Independence

- In the context of a random experiment, knowing that a certain event $B$ has occured may completely change the likelihood we associate to another event $A$.
- For example, suppose we roll two fair dice:
  - The sample space is $S = \{(x, y) : x, y \in \{1, 2, ..., 6\}\}$.
  - Let $A$ denote the event that the sum $x+y = 11$, i.e., $A = \{(5, 6), (6, 5)\}$, and let $B$ denote the event that $x = 1$, i.e. $B = \{(1, 1), (1, 2), ..., (1, 6)\}$.
  - Assuming that the dice are fair, the probability of $A$ is $P(A) = 2/36$.
  - Now, suppose we know that $B$ occurred, i.e. the first die shows 1.
  - Under this "condition", event $A$ is impossible, and its likelihood or probability becomes 0.

- Conditional probabilities provide quantitative measures of likelihood (probability) under the assumption that certain events have occurred, or equivalently, that certain *a priori* knowledge is available.
- In certain situations, knowing that $B$ has occurred does not change the likelihood of $A$; this idea is formalized via the mathematical concept of independence.
- The concepts of conditional probability and independence play a major role in the design and analysis of modern information processing systems, such as digital radio receivers, speech recognition systems, file compression algorithms, etc.

## 4.1 Conditional probability

*Relative frequency interpretation:*

- Consider a random experiment. Let $A$ and $B$ denote two events of interest with $P(B) > 0$.
- Suppose this experiment is repeated a large number of times, say $n$. According to the relative frequency interpretation of probability, we have

$$P(A) \approx \frac{\eta(A)}{n}, \quad P(B) \approx \frac{\eta(B)}{n}, \quad P(A \cap B) \approx \frac{\eta(A \cap B)}{n} \tag{4.1}$$

  where $\eta(A)$, $\eta(B)$ and $\eta(A \cap B)$ denote the number of occurrences of events $A$, $B$ and $A \cap B$ within the $n$ repetitions.
- Provided $\eta(B)$ is large, the probability of $A$, knowing or given that $B$ has occurred, might be evaluated as the ratio

$$P(A \text{ given } B) = \frac{\eta(A \cap B)}{\eta(B)}, \tag{4.2}$$

  also known as a *conditional relative frequency.*
- Using this approach, we have

$$P(A \text{ given } B) = \frac{\eta(A \cap B)}{\eta(B)} = \frac{\eta(A \cap B)/n}{\eta(B)/n} \approx \frac{P(A \cap B)}{P(B)} \tag{4.3}$$

- This and other considerations lead to the following definition.

*Definition:* Consider a random experiment $(S, \mathcal{F}, P)$. Let $B \in \mathcal{F}$ and assume that $P(B) > 0$. For every $A \in \mathcal{F}$, the conditional probability of $A$ given $B$, denoted $P(A|B)$, is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4.4}$$

*Remarks:*

- This definition extends the above concept of conditional relative frequency to the axiomatic probability framework.
- Note that $P(A|B)$ is defined only for the case $P(B) > 0$.

*Theorem 4.1:* Let $B \in \mathcal{F}$ with $P(B) > 0$. The function $P(\,.\,|B) : A \in \mathcal{F} \rightarrow P(A|B) \in \mathbb{R}$, as defined in (4.4), satisfies the axioms of probability, that is:

- Axiom 1:

$$P(A|B) \geq 0 \tag{4.5}$$

- Axiom 2:

$$P(S|B) = 1 \tag{4.6}$$

- Axiom 3: If $A_1, A_2, ...$ is a sequence of mutually exclusive events, then

$$P(\bigcup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} P(A_i|B) \tag{4.7}$$

*Proof:* Left as exercise.

*Further remarks:*

- For a given event $B$ with $P(B) > 0$, the mapping $A \rightarrow P(A|B)$ defines a valid probability function.
- Consequently, all the basic theorems of Section 3.2 apply to $P(A|B)$ as well, with trivial modifications in notation. For example, we have

  $$P(A|B) = 1 - P(A^c|B)$$
  $$P(A \cup C|B) = P(A|B) + P(C|B) - P(A \cap C|B)$$
  etc.

*Example 4.1:*

► A random experiment consists in flipping 3 fair coins. What are the chances of obtaining at least two tails, if we know that the first coin shows heads?

*Solution:* An adequate sample space is

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

Note that $S$ contains $N(S) = 8$ outcomes. Let $A$ denote the event "obtaining at least two tails", and $B$ denote the event "first coin shows head". Using set notation, we have

$$A = \{HTT, THT, TTH, TTT\}$$
$$B = \{HHH, HHT, HTH, HTT\}$$
$$A \cap B = \{HTT\}$$

Since the coins are assumed to be fair, we can use an equiprobable space as model. Therefore, we obtain

$$P(A) = N(A)/N(S) = 4/8 = 1/2$$
$$P(B) = N(B)/N(S) = 4/8 = 1/2$$
$$P(A \cap B) = N(A \cap B)/N(S) = 1/8$$

The conditional probability is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/8}{1/2} = \frac{1}{4}$$

Note that here, knowledge of $B$ significantly decreases the probability of $A$. ◄

*Reduction of sample space:*

- Generally, for an equiprobable space, we have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{N(A \cap B)/N(S)}{N(B)/N(S)} = \frac{N(A \cap B)}{N(B)} \tag{4.8}$$

- This leads to the so-called *reduced sample space* interpretation of the conditional probability $P(A|B)$:
  - Sample space $\rightarrow B$
  - Event $A \subseteq S \rightarrow A \cap B \subseteq B$
  - Probability $P(A|B) = \frac{N(A \cap B)}{N(B)}$

- The fact that neither $S$, nor $N(S)$ are explicitly needed in the computation of $P(A|B)$ may lead to important simplification when solving certain problems.

- The same ideas extend to uniform probability space in the continuous case:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{M(A \cap B)/M(S)}{M(B)/M(S)} = \frac{M(A \cap B)}{M(B)} \tag{4.9}$$

  where $M(A)$ denotes the measure of subset $A$ (length, area, etc.).

*Example 4.1 (revisited)*

▶ Knowing that $B$ has occurred is equivalent to working with the *reduced sample* space

$$B = \{HHH, HHT, HTH, HTT\} \tag{4.10}$$

Also, if we know that $B$ has occurred, then $s \in A$ is equivalent to $s \in A \cap B$, where

$$A \cap B = \{HTT\} \tag{4.11}$$

Thus, according to the reduced sample space interpretation for equiprobable space, we have

$$P(A|B) = \frac{N(A \cap B)}{N(B)} = \frac{1}{4}$$

◀

## 4.2 Conditional probability laws

### 4.2.1 Law of multiplication

*Introduction:*

- Consider the relation defining the conditional probability of $A$ given $B$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4.12}$$

  where it is assumed that $P(B) > 0$.

- From this relation, it follows that

$$P(A \cap B) = P(A|B)P(B) \tag{4.13}$$

- The probability that both $A$ and $B$ occur is equal to the conditional probability of $A$ given $B$, times the probability of $B$.

- This relation may be used advantageously to compute $P(A \cap B)$ when both $P(B)$ and $P(A|B)$ are available.

*Example 4.2: Car renting problem*

► A car rental agency has a fleet of 1000 Ford vehicles: 400 Escorts, 400 Taurus and 200 Explorers. These are equipped with either Firestone or Goodyear tires in the following proportions:

| | Firestone | Goodyear |
|---|---|---|
| Escort | 35 % | 65 % |
| Taurus | 55 % | 45 % |
| Explorer | 40 % | 60 % |

A customer selects a car at random: what is the probability that he/she ends up with an Escort equiped with Firestone tires?

*Solution:* Define the events:

$$\begin{aligned} A &= \{\text{Firestones tires}\} \\ B &= \{\text{Escort}\} \end{aligned}$$

We seek $P(A \cap B)$. From the problem statement, the following information is *directly* available:

$$P(B) = \frac{400}{1000} = 0.4$$
$$P(A|B) = 35\% = 0.35$$

Using relation (4.13), we obtain:

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ &= 0.4 \times 0.35 \\ &= 0.14 \end{aligned}$$

◄

*Remarks:*

- The multiplicative rule $P(A \cap B) = P(A|B)P(B)$ may be generalized to an intersection of $n$ events, where $n$ is an arbitrary integer $\geq 2$.
- To simplify notations, it is convenient to drop the $\cap$ sign for intersection, i.e. $AB = A \cap B$.

*Theorem 4.2:* Let $A_1, A_2, ..., A_n$ be such that $P(A_1A_2...A_{n-1}) > 0$. Then

$$P(A_1A_2...A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2...A_{n-1}) \quad (4.14)$$

*Proof:* First note that $P(A_1A_2...A_{n-1}) > 0$ implies $P(A_1) > 0$, $P(A_1A_2) > 0$, $\ldots$, $P(A_1A_2, ..., A_{n-1}) > 0$. Thus, all the conditional probabilities on the right-hand side (RHS) of (4.14) are well-defined and we have

$$\text{RHS} = P(A_1)\,\frac{P(A_1A_2)}{P(A_1)}\,\frac{P(A_1A_2A_3)}{P(A_1A_2)}\,\cdots\,\frac{P(A_1A_2...A_n)}{P(A_1A_2...A_{n-1})}$$

which is identical to the left-hand side (LHS) after simplification. □

*Remarks:*

- Theorem 4.2 is called the *law of multiplication*; it is also known as the chain rule of probability.
- The theorem is useful when it is desired to compute $P(A_1A_2...A_n)$ and the conditional probabilities in (4.14) may be easily evaluated.
- This often occurs for instance when dealing with temporal or logical sequences of events, as exemplified below.

*Example 4.3:*

▶ An urn contains 10 white balls and 5 black balls. We draw three balls from the urn without replacement. We assume that at each draw, each ball remaining in the urn is equally likely to be chosen. What is the probability that the three balls selected are all white?

*Solution:* Define the events

$$W_i \;=\; \{\text{selecting white ball at the } i^{th} \text{ draw}\}$$

We seek

$$P(W_1W_2W_3) = P(W_1)P(W_2|W_1)P(W_3|W_1W_2)$$

From the problem statement, we find:

$$P(W_1) = \frac{10}{15} \quad \text{and} \quad P(W_2|W_1) = \frac{9}{14}$$

since after the first draw, given a white ball was selected, only 14 balls remain out of which 9 are white. Similarly,

$$P(W_3|W_1W_2) = \frac{8}{13}$$

Therefore

$$P(W_1W_2W_3) = \frac{10}{15} \cdot \frac{9}{14} \cdot \frac{8}{13} = 0.264.$$

◀

### 4.2.2 Law of total probability

*Introduction:*

- Using Theorem 3.5 and the law of multiplication in (4.13), we can write:

$$\begin{aligned} P(A) &= P(AB) + P(AB^c) \\ &= P(A|B)P(B) + P(A|B^c)P(B^c) \end{aligned} \tag{4.15}$$

  where it is assumed that $P(B) > 0$ and $P(B^c) > 0$.

- This result is useful when we desire to compute $P(A)$ and the conditional probabilities $P(A|B)$ and $P(A|B^c)$ may be obtained easily.

*Example 4.4:*

▶ An urn contains 10 white balls and 5 black balls. We draw two balls from the urn at random, without replacement. What is the probability that the second ball is white?

*Solution:* Proceeding as in Example 4.3, define the events

$$\begin{aligned} W_i &= \{\text{selecting white ball at the } i^{th} \text{ draw}\} \\ B_i &= \{\text{selecting black ball at the } i^{th} \text{ draw}\} \end{aligned}$$

We seek $P(W_2)$. Using (4.15) with $A \equiv W_2$, $B \equiv W_1$ and $B^c \equiv B_1$, we obtain

$$\begin{aligned} P(W_2) &= P(W_2|W_1)P(W_1) + P(W_2|B_1)P(B_1) \\ &= \frac{9}{14} \cdot \frac{10}{15} + \frac{10}{14} \cdot \frac{5}{15} \\ &= \frac{10}{15} \cdot \left( \frac{9}{14} + \frac{5}{14} \right) = \frac{2}{3} \end{aligned}$$

One might find it surprising that the answer to this problem is 2/3, which is precisely the initial proportion of white balls in the urn, i.e. before the first draw. However, on second thought, in the absence of *a priori* knowledge about the result of the first draw, there is no apparent reason for the probability to be different from 2/3. ◀

*Partition:*

- A decomposition of a sample space $S$ into a union of 2 or more, disjoint, non-empty subsets is called a partition of $S$.
- Specifically, we say that the sets $B_1, B_2, ..., B_n$ form a partition of $S$ iff

  (1) $B_i \neq \emptyset$ for all $i \in \{1, ..., n\}$

  (2) $B_i B_j = \emptyset$ for all $i \neq j$

  (3) $\cup_{i=1}^{n} B_i = S$
- For example, the sets $B_1 = \{a, b\}$, $B_2 = \{c\}$ and $B_3 = \{d, e\}$ form a partition of $S = \{a, b, c, d, e\}$.

*Remarks:*

- Note that in (4.15), the sets $B$ and $B^c$ form a partition of $S$. ($B$ and $B^c$ are assumed non-empty, $B \cap B^c = \emptyset$ and $B \cup B^c = S$).
- It turns out that (4.15) can be generalized to an arbitrary partition of $S$ into $n$ disjoint subsets, where $n$ is a positive integer.

*Theorem 4.3:* Let $B_1, B_2, ..., B_n$ be a partition of $S$ and assume that $P(B_i) > 0$ for $i = 1, ..., n$. Then

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \tag{4.16}$$

*Proof:* Since $S = B_1 \cup B_2 \cup ... \cup B_n$, we have

$$\begin{aligned} A &= AS \\ &= A(B_1 \cup B_2 \cup ... \cup B_n) \\ &= (AB_1) \cup (AB_2) \cup ... \cup (AB_n) \end{aligned} \tag{4.17}$$

From $B_iB_j = \emptyset$ for $i \neq j$, it follows that $(AB_i) \cap (AB_j)$ for $i \neq j$. Using probability Axiom 3 and the law of multiplication, we finally have:

$$P(A) = \sum_{i=1}^{n} P(AB_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \quad \square \tag{4.18}$$

*Remarks:*

- Theorem 4.3 is called the *law of total probability.*
- We say *total* because the summation in (4.18) is over all the possible different ways of getting $A$.

*Example 4.5: Car renting problem revisited*

► A car rental agency has a fleet of 1000 Ford vehicules: 400 Escort, 400 Taurus and 200 Explorer. These are equiped with either Firestone or Goodyear tires in the following proportions:

| | Firestone | Goodyear |
|---|---|---|
| Escort | 35 % | 65 % |
| Taurus | 55 % | 45 % |
| Explorer | 40 % | 60 % |

A customer selects a car at random: what is the probability that he/she ends up with a car equiped with Firestone tires?

*Solution:* We seek $P(A)$ where

$$A = \{\text{Firestones tires}\}.$$

This information is not directly available from the problem statement. To overcome this difficulty, let us introduce

$$\begin{aligned} B_1 &= \{\text{Escort}\} \\ B_2 &= \{\text{Taurus}\} \\ B_3 &= \{\text{Explorer}\} \end{aligned}$$

We note that $B_1, B_2, B_3$ form a partition of the sample space. Thus, we may use the law of total probabilities to express $P(A)$ in terms of known quantities as follows:

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\ &= 0.35 \times \frac{400}{1000} + 0.55 \times \frac{400}{1000} + 0.4 \times \frac{200}{1000} \\ &= 0.44 \end{aligned}$$

◄

### 4.2.3 Bayes' formula

*Introduction:*

- Suppose that we know $P(B)$, $P(B^c)$, $P(A|B)$ and $P(A|B^c)$. How can we compute $P(B|A)$?
- Basic approach:

  (1) Use definition of conditional probability:

$$P(B|A) = \frac{P(AB)}{P(A)} \tag{4.19}$$

  (2) Use law of multiplication to expand numerator $P(AB)$:

$$P(AB) = P(A|B)P(B) \tag{4.20}$$

  (3) Use law of total probability to expand denominator:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) \tag{4.21}$$

- This approach may be summarized by the formula:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \tag{4.22}$$

*Example 4.6:*

► An urn contains 10 white balls and 5 black balls. We draw two balls from the urn at random, without replacement. Given the second ball is white, what is the probability that the first one was also white?

*Solution:* Define events $W_i$ and $B_i$ as in Example 4.4. We seek $P(W_1|W_2)$. Making use of (4.22), we obtain

$$\begin{aligned} P(W_1|W_2) &= \frac{P(W_2|W_1)P(W_1)}{P(W_2|W_1)P(W_1)+P(W_2|B_1)P(B_1)} \\ &= \frac{\frac{9}{14}\cdot\frac{2}{3}}{\frac{9}{14}\cdot\frac{2}{3}+\frac{10}{14}\cdot\frac{1}{3}} \\ &= \frac{9}{14} \end{aligned}$$

This result admits a simple interpretation in terms of reduced sample space: given that the second ball is white is equivalent to selecting the first ball randomly among a reduced set of 14 balls containing 9 white and 5 black, hence the result.

Warning: Although effective in this simple example, the use of a reduced sample space approach to solve more complex conditional probability problems requires great care, or it may lead to an erroneous solution. The use of a deductive approach (e.g. 4.22) is recommended. ◄

*Remarks:*

- In (4.22), events $B$ and $B^c$ form a partition of the sample space $S$.
- As for the law of total probability, (4.22) may be generalized to an arbitrary partition $B_1, B_2, \ldots, B_n$ of $S$.

*Theorem 4.4:* Suppose $B_1, B_2, ..., B_n$ is a partition of $S$ with $P(B_i) > 0$ for $i = 1, ..., n$. Let $A$ be any event with $P(A) > 0$. Then, for any $k \in \{1, ..., n\}$

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} \tag{4.23}$$

*Proof:* From the definition of conditional probability, we have

$$P(B_k|A) = \frac{P(B_k A)}{P(A)}$$

Using the law of multiplication, the numerator can be expressed as

$$P(B_k A) = P(A|B_k)P(B_k)$$

Using the law of total probability, the denominator can be expanded as

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

The desired result follows by combining the above expressions. □

*Remarks:*

- Theorem 4.4 is also known as *Bayes' formula*.
- Bayes formula is useful when the probabilities $P(B_i)$ and $P(A|B_i)$ are known for $i = 1, ..., n$, and it is desired to compute $P(B_k|A)$ for one or more values of $k$.
- In statistical applications of this formula, the following special terminology is often used:
  - The events $B_i$ $(i = 1, ..., n)$, which form a partition of $S$, are called hypotheses.
  - The probabilities $P(B_i)$ $(i = 1, ..., n)$ are called *a priori* probabilities (i.e., before knowing that $A$ occurred).
  - The conditional probabilities $P(B_k|A)$ $(k = 1, ..., n)$ are called *a posteriori* probabilities (i.e., after knowing that $A$ occurred).

*Example 4.7: Car renting problem, again*

► A car rental agency has a fleet of 1000 Ford vehicules: 400 Escorts, 400 Taurus and 200 Explorers. These are equipped with either Firestone or Goodyear tires in the following proportions (in %):

| | Firestone | Goodyear |
|---|---|---|
| Escort | 35 % | 65 % |
| Taurus | 55 % | 45 % |
| Explorer | 40 % | 60 % |

A customer selects a car at random: given that the car is equiped with Firestone tires, what is the probability that it is an Explorer?

*Solution:* Define the events

$$\begin{aligned} A &= \{\text{Firestones tires}\} \\ B_1 &= \{\text{Escort}\} \\ B_2 &= \{\text{Taurus}\} \\ B_3 &= \{\text{Explorer}\} \end{aligned}$$

where $B_1, B_2, B_3$ form a partition of the sample space. We seek $P(B_3|A)$. Using Bayes' formula, we find

$$\begin{aligned} P(B_3|A) &= \frac{P(A|B_3)P(B_3)}{\sum_{i=1}^{3} P(A|B_i)P(B_i)} \\ &= \frac{40\% \times \frac{200}{1000}}{0.44} \\ &= 0.18 \end{aligned}$$

where the value of 0.44 for the denominator has already been computed in example 4.5. ◄

## 4.3 Independence

*Introduction:*

- Consider a random experiment in which a fair coin is tossed twice:

$$S = \{HH, HT, TH, TT\}$$

- Consider the two events:

$$\begin{aligned} A = \{\text{heads up on first toss}\} &= \{HH, HT\} \implies P(A) = 1/2 \\ B = \{\text{heads up on second toss}\} &= \{HH, TH\} \implies P(B) = 1/2 \end{aligned}$$

- How does knowledge that $B$ occurred affects the likelihood of $A$?

$$AB = A \cap B = \{HH\} \quad \Longrightarrow \quad P(AB) = 1/4$$

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

- Since $P(A|B) = P(A)$, we conclude that the occurrence of $B$ has no effect on the likelihood of $A$. We say that $A$ is *independent* of $B$

*Discussion:*

- If $A$ is independent of $B$, as defined above, then:

$$\begin{aligned} P(A|B) = P(A) \;&\Rightarrow\; \frac{P(AB)}{P(B)} = P(A) \\ &\Rightarrow\; P(AB) = P(A)P(B) \end{aligned} \tag{4.24}$$

- In turns, it follows from (4.24) that (assume $P(A) > 0$):

$$P(B|A) = \frac{P(BA)}{P(A)} = P(B) \tag{4.25}$$

- Thus, $A$ independent of $B$ implies that $B$ independent of $A$. We say that independence is a *symmetric* relation.

- Because of this symmetry, it is more natural (and practical) to define independence directly in terms of (4.24).

*Definition:* Two events $A$ and $B$ are called independent iff

$$P(AB) = P(A)P(B) \tag{4.26}$$

*Remarks:*

- This definition is valid even when $P(A) = 0$ or $P(B) = 0$.
- If $A$ and $B$ are independent, with $P(B) > 0$, then

  $$P(A|B) = \frac{P(AB)}{P(B)} = P(A) \tag{4.27}$$

  so that occurrence of $B$ does not affect likelihood of $A$. Similarly, assuming $P(A) > 0$, we have $P(B|A) = P(B)$.
- Independence conveys the idea of an absence of a causal relation between events $A$ and $B$.
- It is not always obvious that two events $A$ and $B$ are independent.

*Example 4.8:*

► A card is drawn randomly form a 52-card deck. Consider the events $A = \{\text{getting a heart}\}$ and $B = \{\text{getting an ace}\}$. Here, we have:

$$P(A) = 13/52 = 1/4$$
$$P(B) = 4/52 = 1/13$$
$$P(AB) = P(\{\text{ace of heart}\}) = 1/52$$

Since $P(AB) = P(A)P(B)$, we conclude that $A$ and $B$ are independent. ◄

*Example 4.9:*

► An urn contains 10 white balls and 5 black balls. Suppose that two balls are picked at random from the urn. Let $W_1$ and $W_2$ denote the events that the first and second ball is white, respectively. Determine whether or not these two events are independent if (a) the balls are selected with replacement and (b) without replacement?

*Solution:* ◄

*Theorem 4.5:* If $A$ and $B$ are independent, then so are the pairs:

(a) $A$ and $B^c$ .

(b) $A^c$ and $B$ .

(c) $A^c$ and $B^c$ .

*Proof:* From Theorem 3.5, we have

$$P(A) = P(AB) + P(AB^c) \tag{4.28}$$

Thus, using the fact that $A$ and $B$ are independent,

$$\begin{aligned} P(AB^c) &= P(A) - P(AB) \\ &= P(A) - P(A)P(B) \\ &= P(A)(1 - P(B)) \\ &= P(A)P(B^c) \end{aligned} \tag{4.29}$$

This shows that $A$ and $B^c$ are independent. By symmetry, it follows that $A^c$ and $B$ are independent, and from that, we finally deduce that $A^c$ and $B^c$ are independent. □

*Definition:* The events $A_1, A_2, ..., A_n$ are called (mutually) independent iff all the relations below hold:

$$P(A_iA_j) = P(A_i)P(A_j) \quad \text{for all } i < j \tag{4.30}$$

$$P(A_iA_jA_k) = P(A_i)P(A_j)P(A_k) \quad \text{for all } i < j < k \tag{4.31}$$

$$\dots$$

$$P(A_1A_2...A_n) = P(A_1)P(A_2)...P(A_n) \tag{4.32}$$

*Remarks:*

- (4.30) is called pairwise independence.
- (4.30)-(4.32) is called mutual independence (much stronger).
- For example, consider three events $A$, $B$ and $C$:
  - mutual independence implies that $A$ and $B \cup C$ are independent;
  - pairwise independence DOES NOT.
- In applications, mutual independence is often put forward as an assumption (probability model) in the solution of a complex problem.

*Example 4.10:*

► Consider the electric circuit shown below in which each of the four switches, labelled $S1, S2, S3$ and $S4$, is independently closed or open with probability $p$ and $1-p$, respectively. If a voltage is applied at the input, what is the probability that it is transmitted at the output? ◄

## 4.4 Product of independent experiments

*Introduction:*

- Often, a random experiment may be viewed as a compounded experiment, made up of "smaller", independent experiments that occur sequentially and/or concurrently in time.
- Some examples include:
  - flipping a coin, then rolling a die,
  - drawing $N$ cards with replacement from a 52-card deck,
  - transmitting a sequence of $10^8$ bits over a digital link.
- In this section:
  - a basic mathematical framework is developed to deal with such situations;
  - a special case of particular interest (i.e. Bernouilli trials) is then investigated.

### 4.4.1 The basic model

*Definition:* We say that random experiment $(S, \mathcal{F}, P)$ is the product of $n$ random experiments $(S_i, \mathcal{F}_i, P_i)$ if

(a) $S = S_1 \times S_2 \times ... \times S_n$, where $\times$ denotes the cartesian product.

(b) $\mathcal{F}$ is the smallest $\sigma$-algebra containing all cartesian products of the type $A_1 \times A_2 \times ... \times A_n$, with $A_i \in \mathcal{F}_i$.

(c) For any $A_i \in \mathcal{F}_i$, $i = 1, \ldots, n$, we have

$$P(S_1 \times \ldots \times S_{i-1} \times A_i \times S_{i+1} \times \ldots \times S_n) = P_i(A_i) \tag{4.33}$$

*Definition:* Within this framework, we say that the sub-experiments $(S_i, \mathcal{F}_i, P_i)$ are independent (also called independent trials) if conditions (a), (b) and (c) above are satisfied and if

$$P(A_1 \times A_2 \times ... \times A_n) = P_1(A_1)P_2(A_2)...P_n(A_n) \tag{4.34}$$

*Remarks:*

- Note that condition (4.34) supersedes condition (4.33).
- As a consequence of (4.34), the probability of any event associated to the product experiment may be computed from the individual probability functions associated to the sub-experiments.
- The use of (4.34) as a probability model is extremely useful. However, it must be justified on physical and/or experimental grounds, as it will not be valid for all kinds of combined experiments.

*Example 4.11:*

► A random experiment consist in the following sequence of two sub-experiments, each one characterized by its own probability space:

- Firstly, flipping a fair coin once:
  - Sample space: $S_1 = \{H, T\}$
  - Probability function: $P_1(\{H\}) = P_1(\{T\}) = 1/2$
- Secondly, rolling a fair die once:
  Sample space: $S_2 = \{1, 2, 3, 4, 5, 6\}$
  Probability: $P_2(\{i\}) = 1/6$ for $i = 1, ..., 6$

In this type of situation, and in the absence of further information, it is reasonable to assume that the sub-experiments are independent. The sample space of the product experiment is then

$$\begin{aligned} S &= S_1 \times S_2 \\ &= \{ai : a \in \{H, T\} \text{ and } i \in \{1, 2, 3, 4, 5, 6\}\} \\ &= \{H1, H2, ..., H6, T1, T2, ..., T6\} \end{aligned}$$

For any $A_1 \subseteq S_1$ and $A_2 \subseteq S_2$, we have by assumption:

$$P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$$

For example, define the events

$$\begin{aligned} A_1 &= \{\text{coin shows heads}\} = \{H\} \\ A_2 &= \{\text{die shows even}\} = \{2, 4, 6\} \end{aligned}$$

The probability of the compound event $A_1 \times A_2 = \{\text{heads followed by even}\}$ in the product experiment can be obtained as:

$$P(A_1 \times A_2) = P_1(A_1)P_2(A_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

In fact, since for any outcome $ai \in S$, we have $\{ai\} = \{a\} \times \{i\}$, it follows that

$$P(\{ai\}) = P_1(\{a\})P_2(\{i\}) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12} \tag{4.35}$$

In this example, the assumption of independent experiments is equivalent to an equiprobable model for the combined experiments, which is intuitively satisfying.
◄

*Theorem 4.6:* Let $(S, \mathcal{F}, P)$ be the product of $n$ random experiments $(S_i, \mathcal{F}_i, P_i)$, $i = 1, ..., n$. The product space $(S, \mathcal{F}, P)$ is equiprobable if and only if the sub-experiments $(S_i, \mathcal{F}_i, P_i)$ are independent and equiprobable.

*Remarks:*

- The theorem provides a nice generalization Example 4.11.
- It gives one particular set of conditions under which independence of the sub-experiments is satisfied.
- The next theorem provides a link between the notions of independent experiments and mutually independent events.

*Theorem 4.7:* Let $(S, \mathcal{F}, P)$ be a product of $n$ independent experiments $(S_i, \mathcal{F}_i, P_i)$. Suppose that events $\mathcal{A}_1, \ldots, \mathcal{A}_n$ in $\mathcal{F}$ are such that the occurrence of $\mathcal{A}_i$ only depends on the result of the $i$th experiment. The events $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are then mutually independent.

*Remarks:*

- The statement that the occurrence of event $\mathcal{A}_i$ only depends on result of the $i$th experiment is equivalent to

$$\mathcal{A}_i = S_1 \times \ldots \times S_{i-1} \times A_i \times S_{i+1} \times \ldots \times S_n \tag{4.36}$$

  for some $A_i \in S_i$.

- The proof of Theorem 4.7 amounts to showing that

$$P(\mathcal{A}_1 \cap \mathcal{A}_2 \cap ... \cap \mathcal{A}_n) = P(\mathcal{A}_1)P(\mathcal{A}_2)...P(\mathcal{A}_n) \tag{4.37}$$

  This follows easily by using (4.36) in combination with (4.34).

*Proof of Theorem 4.7 (optional reading):* To simplify the presentation, assume $n = 2$. The statement that $\mathcal{A}_i$ only depends on the result of the $i$th experiment is equivalent to

$$\mathcal{A}_1 = A_1 \times S_2 \quad \text{where} \quad A_1 \subseteq S_1 \tag{4.38}$$

$$\mathcal{A}_2 = S_1 \times A_2 \quad \text{where} \quad A_2 \subseteq S_2 \tag{4.39}$$

From (4.34), it follows that

$$P(\mathcal{A}_1) = P(A_1 \times S_2) = P_1(A_1)P_2(S_2) = P_1(A_1) \tag{4.40}$$

$$P(\mathcal{A}_2) = P(S_1 \times A_2) = P_1(S_1)P_2(A_2) = P_2(A_2) \tag{4.41}$$

Now, since

$$\mathcal{A}_1 \cap \mathcal{A}_2 = (A_1 \times S_2) \cap (S_1 \times A_2) = A_1 \times A_2,$$

we obtain

$$\begin{aligned} P(\mathcal{A}_1 \cap \mathcal{A}_2) &= P(A_1 \times A_2) \\ &= P_1(A_1)P_2(A_2) = P(\mathcal{A}_1)P(\mathcal{A}_2) \end{aligned} \tag{4.42}$$

This shows that $\mathcal{A}_1$ and $\mathcal{A}_2$ are independent. □

*Further remarks (optional reading):*

- To simplify notations, it is common practice to identify the events $\mathcal{A}_i \subseteq S$ with the corresponding events $A_i \subseteq S_i$ and to write (4.34) in the form

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)P(A_2)...P(A_n) \tag{4.43}$$

- Note that in (4.43), the $P(.)$ on the LHS and RHS really represent different functions; also, the intersections on the LHS represents cartesian products.

### 4.4.2 Sequence of Bernouilli trials

*Definition:* A Bernouilli trial is a random experiment $(S, \mathcal{F}, P)$ in which a particular event $A \in \mathcal{F}$ has been identified and assigned a probability $p = P(A)$, $0 \leq p \leq 1$.

*Remarks:*

- Clearly, event $A$ may or not occur during the trial.
- Event $A$ is called a success, while its complement $A^c$ is called a failure.
- $p = P(A) =$ is called probability of success, while $q \triangleq P(A^c) = 1 - p =$ is the probability of failure.

*Example:* Basic examples of Bernouilli trials include:

- Getting heads when flipping a fair coin: $A = \{H\}$, $A^c = \{T\}$, $p = 1/2$.
- Getting five or more when rolling a fair die: $A = \{5, 6\}$, $A^c = \{1, 2, 3, 4\}$, $p = 1/3$.

*Definition:* A sequence of Bernouilli trials is a product experiment that consists of $n$ identical and independent Bernouilli trials, each with probability of success $p$.

*Examples:*

- Sequence of 10 independent flips of a coin.
- Independent transmission of $10^{12}$ bits over a digital communication link.

*Theorem 4.8 (Bernouilli law):* The probability of $k$ successes in a sequence of $n$ Bernouilli trials, with individual probability of success $p$, is given by:

$$P(k,n) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.44}$$

*Proof:* Think of the outcome of a sequence of $n$ Bernouilli trials as an $n$-tuple of the form $a_1 a_2 \ldots a_n$ where each $a_i$ may take the value $S$ for success or $F$ for failure. In this setting, obtaining $k$ successes corresponds to an $n$-tuple containing letter $S$ in exactly $k$ positions (and letter $F$ in the remaining $n-k$ positions). Note the following:

- Because of the independence assumption, the probability of occurrence of each such $n$-tuple of $k$ successes is given by $p^k(1-p)^{n-k}$.
- The number of different $n$-tuples containing letter $S$ in exactly $k$ position is given by $\binom{n}{k}$.

Hence, from the additivity property of probability (Axiom 3), the probability of $k$ successes is obtained as (4.44). □

*Example 4.12:*

▶ Consider a biased (unfair) coin with $P(H) = 0.6$ and $P(T) = 0.4$. What is the probability of exactly two heads in 5 independent throws? ◀

# Problems

1. We draw 8 cards at random from a 52-card deck, without replacement. Given that at least 3 out of the 8 cards are spades, what is the probability that all 8 cards are spades?

2. A box contains 95 good resistors and 5 bad ones. To find the bad one, a student decides to test them one by one (without replacement). If the first 15 resistors are all good, what is the probability that the next one is defectuous? (Hint: you may to try a reduced sample space approach).

3. Consider the independent transmission of binary digits (or bits) over a noisy communication channel. Suppose that for each bit, the probability of making an error during transmission (e.g.: send 1 and receive 0) is equal to $p$ ($0 < p \ll 1$). Consider the transmission of $n$ consecutive bits. Find the probability (a) of making no errors; (b) of making 1 error and (c) of making 2 errors. Evaluate numerically for $p = 10^{-4}$ and $n = 10^8$.

# Chapter 5

# Introduction to Random Variables

Consider a random experiment described by a triplet $(S, \mathcal{F}, P)$. In applications of probabilities, we are often interested in numerical quantities derived from the experimental outcomes. These quantities may be viewed as functions from the sample space $S$ into the set of real numbers $\mathbb{R}$, as in:

$$s \in S \rightarrow X(s) \in \mathbb{R}$$

Provided certain basic requirements are satisfied, these quantities are generally called *random variables*.

As an example, consider the sum of the two numbers showing up when rolling a fair die twice:

- The set of all possible outcomes is $S = \{(i, j) : i, j \in \{1, 2, ..., 6\}\}$.
- The sum of the two numbers showing up may be represented by the functional relationship

$$s = (i, j) \rightarrow X(s) = i + j.$$

- Note that the function $X(s)$ may be used in turn to define more complex events. For instance, the event that the sum is greater or equal to 11 may be expressed concisely as $A = \{s \in S : X(s) \geq 11\}$

The terminology *random variable* is appropriate in this type of situations because:

- The value $X(s)$ depends on the experimental outcome $s$.
- The outcome $s$ of a particular realization of the random experiment (i.e. a trial) is unknown beforehand, and so is $X(s)$.
- Each experimental trial may lead to a different value of $X(s)$

Random variables are extremely important in engineering applications. They are often used to model physical quantities of interest that cannot be predicted exactly due to uncertainties. Some examples include:

- Voltage and current measurements in an electronic circuit.
- Number of erroneous bits per second in a digital transmission.
- Instantaneous background noise amplitude at the output of an audio amplifier.

Modelization of such quantities as random variables allows the use of probability in the design and analysis of these systems.

This and the next few Chapters are devoted to the study of random variables, including: definition, characterization, standard models, properties, and a lot more...

In this Chapter, we give a formal definition of a random variable, we introduce the concept of a cumulative distribution function and we introduce the basic types of random variables.

## 5.1 Preliminary notions

*Function from $S$ into $\mathbb{R}$:*

- Let $S$ denote a sample space of interest.
- A function from $S$ into $\mathbb{R}$ is a mapping, say $X$, that associate to every outcome in $S$ a unique real number $X(s)$:

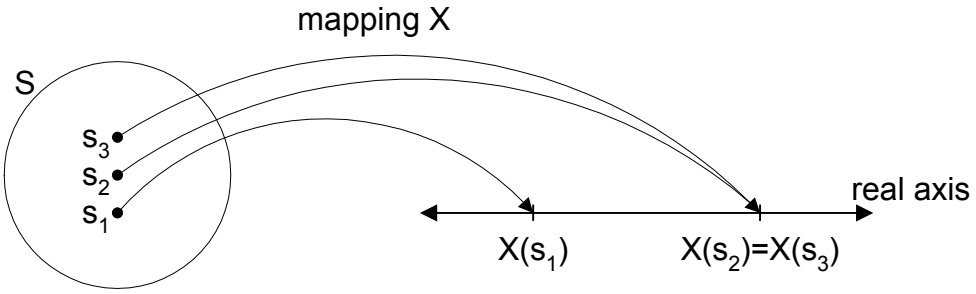



Figure 5.1: Illustration of a mapping $X$ from $S$ into $\mathbb{R}$.

- The following notation is often used to convey this idea:

$$X : s \in S \rightarrow X(s) \in \mathbb{R}. \tag{5.1}$$

- We refer to the sample space $S$ as the *domain* of the function $X$.
- The *range* of the function $X$, denoted $\mathcal{R}_X$, is defined as

$$\mathcal{R}_X = \{X(s) : s \in S\} \subseteq \mathbb{R} \tag{5.2}$$

  That is, $\mathcal{R}_X$ is the of all possible values for $X(s)$, or equivalently, the set of all real numbers that can be "reached" by the mapping $X$.

*Inverse function:*

- Let $X$ be a function from $S$ into $\mathbb{R}$.
- We define the inverse function $X^{-1}$ as follows: for any subset $D$ of $\mathbb{R}$,

$$X^{-1}(D) = \{s \in S : X(s) \in D\} \tag{5.3}$$

- That is, $X^{-1}(D)$ is the subset of $S$ containing all the outcomes $s$ (possibly more than one) such that $X(s)$ is in $D$. This is illustrated below.

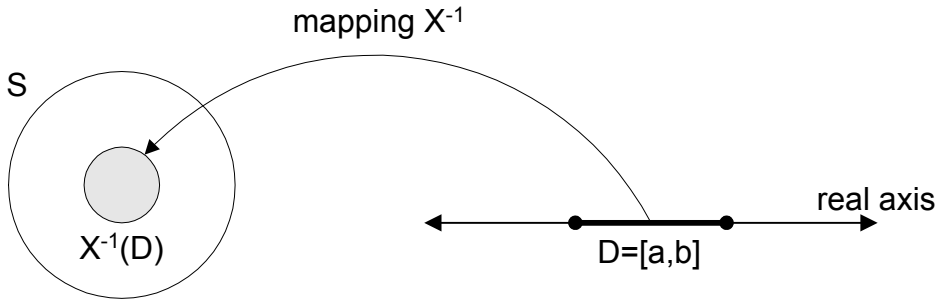



Figure 5.2: Illustration of inverse mapping $X^{-1}$

- This definition of an inverse is very general; it applies even in the case when $X$ is many-to-one.

*Left semi-infinite intervals:*

- To every $x \in \mathbb{R}$, we associate a left semi-infinite interval $I_x$, defined as

$$I_x \triangleq (-\infty, x] = \{y \in \mathbb{R} : y \leq x\} \tag{5.4}$$

- The inverse image of $I_x$ under mapping $X$ is given by

$$X^{-1}(I_x) = \{s \in S : X(s) \leq x\} \tag{5.5}$$

## 5.2 Definition of a random variable

*Definition:* Let $(S, \mathcal{F}, P)$ be a probability space. A function $X : s \in S \rightarrow X(s) \in \mathbb{R}$ is called a *random variable* (RV) if

$$X^{-1}(I_x) = \{s \in S : X(s) \leq x\} \in \mathcal{F}, \quad \text{for all } x \in \mathbb{R} \tag{5.6}$$

*Discussion:*

- According to this definition, $X$ defines a mapping from sample space $S$ into $\mathbb{R}$, as illustrated in Figure 5.1
- However, $X$ is not arbitrary: we require that for any $x \in \mathbb{R}$, the inverse image $X^{-1}(I_x)$, as illustrated in Figure 5.3, be a valid event.

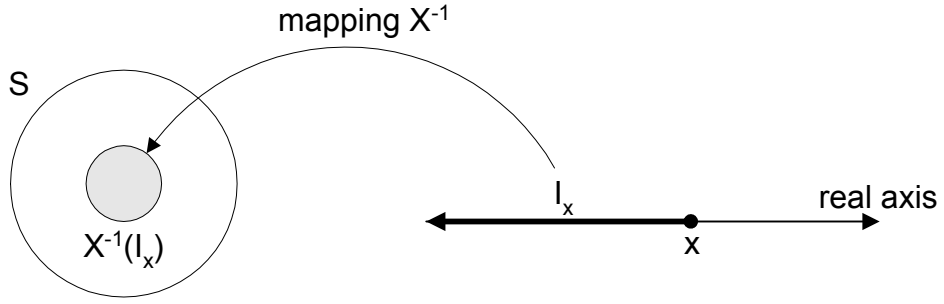



Figure 5.3: Inverse image $X^{-1}(I_x)$

- This condition ensures that $P(\{s \in S : X(s) \leq x\})$, i.e. the probability that $X(s)$ belong to the interval $I_x$, is well-defined.

*Example 5.1:*

► A fair coin is flipped twice. Let random variable $X$ represent the number of tails observed in this experiment. Here, the sample space may be defined as

$$S = \{HH, HT, TH, TT\}$$

Since this is a finite set, a proper choice of event algebra is

$$\mathcal{F} = \mathcal{P}_S = \{\emptyset, \{HH\}, \ldots, S\}$$

Note that $\mathcal{F}$ contains $2^4 = 16$ events. According to the problem statement, the function $X : S \to \mathbb{R}$ may be computed as follows:

$$\begin{aligned} s = HH &\rightarrow X(s) = 0 \\ s = HT \text{ or } TH &\rightarrow X(s) = 1 \\ s = TT &\rightarrow X(s) = 2 \end{aligned}$$

so that its range is $\mathcal{R}_X = \{0, 1, 2\}$. This is illustrated below:

For any $x \in \mathbb{R}$, we have

$$X^{-1}(I_x) = \{s \in S : X(s) \leq x\} \in \mathcal{F}$$

since $I_x$ is a subset of $S$ and $\mathcal{F} = \mathcal{P}_S$ contains all the possible subsets of $S$. This shows that $X$ is a valid random variable. For instance:

$$\begin{aligned} x < 0 &\Rightarrow X^{-1}(I_x) = \emptyset \in \mathcal{F} \\ 0 \leq x < 1 &\Rightarrow X^{-1}(I_x) = \{HH\} \in \mathcal{F} \\ 1 \leq x < 2 &\Rightarrow X^{-1}(I_x) = \{HH, HT, TH\} \in \mathcal{F} \\ 2 \leq x &\Rightarrow X^{-1}(I_x) = S \in \mathcal{F} \end{aligned}$$

.

◄

*Remarks on condition (5.6):*

- By definition, $X^{-1}(I_x) = \{s \in S : X(s) \leq x\}$ is a subset of $S$. Therefore, whenever $\mathcal{F} = \mathcal{P}_S$, condition (5.6) is satisfied and we need not worry about it. This is the case when $S$ is finite or countably infinite.

- When $S$ is uncountably infinite and $\mathcal{F} = \mathcal{B}_S$, there exist functions $X : S \to \mathbb{R}$ that do not satisfy condition (5.6) and for which $P(\{s \in S : X(s) \leq x\})$ is not defined. It is precisely to avoid this situation that (5.6) is included in the definition of a random variable.

- In applications, we will want to compute probabilities of the type

$$P(\{s \in S : X(s) \in D\})$$

  where $D \subseteq \mathbb{R}$ represents any practical subset of real numbers. This includes, for example, intervals of the type $[a, b]$, for any $a \leq b \in \mathbb{R}$, as well as unions, intersections and/or complements of such intervals.

- As a consequence of condition (5.6), it can be shown that

$$\{s \in S : X(s) \in D\}) \in \mathcal{F}$$

  for any $D \in \mathcal{B}_{\mathbb{R}}$, so that $P(\{s \in S : X(s) \in D\})$ is well-defined for any practical subset of real-numbers (see next page for more on this).

- In this course, we shall work with relatively simple functions $X$ and shall *always* assume that the condition $X^{-1}(I_x) \in \mathcal{F}$ is satisfied.

*Further remarks (optional reading):*

- The condition $X^{-1}(I_x) \in \mathcal{F}$ in the definition of random variable $X$ ensures that

$$X^{-1}(I_x) = \{s \in S : X(s) \leq x\}$$

  is a valid event, for which the probability $P(X^{-1}(I_x))$ is well-defined.

- More importantly, the condition ensures that for practical subsets of real numbers encountered in applications of probability, i.e. for any $D \in \mathcal{B}_{\mathbb{R}}$, the Borel field of $\mathbb{R}$, the set of experimental outcomes

$$X^{-1}(D) = \{s \in S : X(s) \in D\}$$

  is also a valid event for which a probability can be computed.

- While the detail of the proof are beyond the scope of this course, the justification of this statement involves three basic steps:
  - Any real number subset $D \in \mathcal{B}_{\mathbb{R}}$ can be expressed as a combination of unions, intersections and complements of basic intervals of the type $I_x$.
  - Because $X^{-1}(I_x) \in \mathcal{F}$ for any $x$ and because $\mathcal{F}$ is a $\sigma$-algebra (closed under union, intersection and complement), it follows that $X^{-1}(D)$ is also in $\mathcal{F}$.
  - Finally, $X^{-1}(D) \in \mathcal{F}$ implies that $P(X^{-1}(D))$ is well-defined.

- In the next section, we will see how $P(X^{-1}(I_x))$ can actually be used in the computation of $P(X^{-1}(D))$.

*Simplified notations:*

- Let $D \subseteq \mathbb{R}$. The following notations for $X^{-1}(D)$ are equivalent:

$$X^{-1}(D) = \{s \in S : X(s) \in D\} = \{X \in D\} \tag{5.7}$$

  The notation $\{X \in D\}$ is generally preferred. When using it, you should keep in mind that it really refers to an event, i.e. a subset of $S$.

- When referring to the probability of the event $\{X \in D\}$, we often drop the curly brackets. The following notations are thus equivalent:

$$P(\{s \in S : X(s) \in D\}) = P(\{X \in D\}) = P(X \in D) \tag{5.8}$$

- The notation $X(.)$ is used to represent the function $X : S \to \mathbb{R}$, while the notation $X(s)$ means the value of $X(.)$ at the point $s$. In probability textbooks, $X$ is often used to denote both $X(.)$ and $X(s)$; the interpretation is context dependent.

## 5.3 Cumulative distribution function

*Introduction:*

- Let $X$ be a random variable defined on $(S, \mathcal{F}, P)$.
- According to the definition of a random variable, this implies that the probability $P(X \leq x)$ is well-defined for any real-number $x$.
- If $P(X \leq x)$ is known for all $x \in \mathbb{R}$, it is possible (in theory) to compute $P(X \in D)$ for essentially any subset $D \subseteq \mathbb{R}$ of practical interest.
- For this reason, the quantity $P(X \leq x)$, seen as a function of $x$, plays a very important role in probability and is thus given a special name.

*Definition:* The function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) \triangleq P(X \leq x), \quad \text{for all } x \in \mathbb{R} \tag{5.9}$$

is called the cumulative distribution function (CDF) of $X$.

*Remarks:*

- One should bear in mind that $F(x) = P(X \leq x)$ really means

$$F(x) = P(\{s \in S : X(s) \leq x\})$$

- We say cumulative because as $x$ increases, the set $\{s \in S : X(s) \leq x\}$ includes more and more possible outcomes $s \in S$.

*Example 5.2:*

► A random experiment consists in flipping two fair coins. Let RV $X$ represents the number of tails. Find the CDF of $X$?

*Solution:*

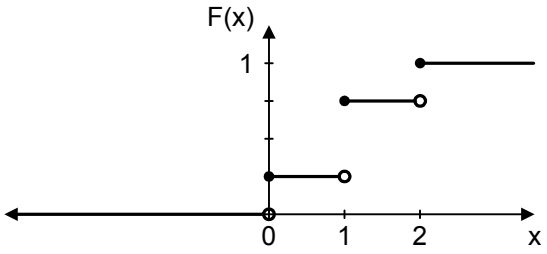
- Sample space: $S = \{HH, HT, TH, TT\}$
- Values of interest for $X$: $s \in S \Rightarrow X(s) \in \{0, 1, 2\} = \mathcal{R}_X$
- Corresponding probabilities:

$$\begin{aligned} P(X=0) &= P(\{HH\}) = 1/4 \\ P(X=1) &= P(\{HT, TH\}) = 1/2 \\ P(X=2) &= P(\{TT\}) = 1/4 \end{aligned}$$

- Distribution function:

$$\begin{aligned} x < 0 &\Rightarrow F(x) = P(X \leq x) = 0 \\ 0 \leq x < 1 &\Rightarrow F(x) = P(X \leq x) = P(X=0) = 1/4 \\ 1 \leq x < 2 &\Rightarrow F(x) = P(X \leq x) = P(X=0) + P(X=1) = 3/4 \\ 2 \leq x &\Rightarrow F(x) = P(X \leq x) = 1 \end{aligned}$$

- Graphical representation of $F(x)$:




◄

*Theorem 5.1:* The CDF $F(x)$ satisfies the following basic properties:

(a)
$$a < b \Rightarrow F(a) \leq F(b) \tag{5.10}$$

(b)
$$F(\infty) \equiv \lim_{x\to\infty} F(x) = 1 \tag{5.11}$$

(c)
$$F(-\infty) \equiv \lim_{x\to-\infty} F(x) = 0 \tag{5.12}$$

(d)
$$F(a^+) \equiv \lim_{x\to a+} F(x) = F(a) \tag{5.13}$$

(e)
$$F(a^-) \equiv \lim_{x\to a-} F(x) = F(a) - P(X = a) \tag{5.14}$$

*Remarks:*

- According to (a), $F(x)$ is non-decreasing.
- From properties (b) and (c), it follows that $F(x)$ is lower bounded by 0 in the limit $x \to -\infty$ and upper bounded by 1 in the limit $x \to \infty$.
- According to (d), $F(x)$ is right continuous.
- However, from (e), we conclude that $F(x)$ is not necessarily left-continuous:
  - if not, the size of the jump at $x = a$, i.e. the difference $F(a) - F(a^-)$, is equal to the probability $P(X = a)$.
  - if $F(x)$ is continuous at $x = 1$ (i.e. no jump), then $P(X = a) = 0$.

*Proof:*

(a) Let $a$ and $b$ be arbitrary real numbers with $a < b$. We have

$$\begin{aligned} a < b &\Rightarrow \{X \leq a\} \subseteq \{X \leq b\} \\ &\Rightarrow P(X \leq a) \leq P(X \leq b) \\ &\Rightarrow F(a) \leq F(b) \end{aligned}$$

where the second line follows from Theorem 3.3.

(b) For $n \in \mathbb{N}$, the sets $A_n \triangleq \{s \in S : X(s) \leq n\}$ define an increasing sequence of events with $\lim_{n\to\infty} A_n = \cup_{n=1}^{\infty} A_n = S$, where $S$ is the sample space. Thus, making use of Theorem 3.6, we have

$$\begin{aligned} \lim_{n\to\infty} F(n) &= \lim_{n\to\infty} P(\{s \in S : X(s) \leq n\}) \\ &= \lim_{n\to\infty} P(A_n) \\ &= P(\lim_{n\to\infty} A_n) \\ &= P(S) = 1 \end{aligned} \tag{5.15}$$

Because $F(x)$ is non-decreasing and $\leq 1$, (6.11) this in turns imply $\lim_{x\to\infty} F(x) = 1$.

(c)-(e) Left as optional exercise for the student. □

*Remarks:*

- Any function $F : \mathbb{R} \to [0, 1]$ satisfying properties (a)-(d) above is generally called a cumulative distribution function.
- In theory, if the CDF $F(x) = P(X \leq x)$ is known for all $x \in \mathbb{R}$, it can be used to compute the probability of any event of the type $\{s \in S : X(s) \in D\}$ where $D \subseteq \mathcal{B}_{\mathbb{R}}$.
- This is illustrated in the example below.

*Example 5.3:*

▶ Let $a$ and $b$ be arbitrary real numbers such that $a < b$. Express the following probabilities in terms of the CDF $F(x)$: $P(X > a)$, $P(X = a)$, $P(X < a)$ and $P(a < X \leq b)$.

*Solution:* Since $\{X > a\}$ is the complement of $\{X \leq a\}$, we have:

$$\begin{aligned} P(X > a) &= 1 - P(X \leq a) \\ &= 1 - F(a) \end{aligned} \tag{5.16}$$

From Theorem 5.1, (e):

$$P(X = a) = F(a) - F(a^-) \tag{5.17}$$

Note that $\{X < a\} = \{X \leq a\} - \{X = a\}$, where $\{X = a\} \subseteq \{X \leq a\}$. Therefore, using Theorem 3.3, we have:

$$\begin{aligned} P(X < a) &= P(X \leq a) - P(X = a) \\ &= F(a) - (F(a) - F(a^-)) \\ &= F(a^-) \end{aligned} \tag{5.18}$$

Since $\{a < X \leq b\} = \{X \leq b\} - \{X \leq a\}$, with $\{X \leq a\} \subseteq \{X \leq b\}$, we have similarly:

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F(b) - F(a) \end{aligned} \tag{5.19}$$

◀

*List of formulae:*

- A more complete list of such properties is given below:

| Event | Expression |
|---|---|
| $X \leq a$ | $F(a)$ |
| $X < a$ | $F(a^-)$ |
| $X > a$ | $1 - F(a)$ |
| $X \geq a$ | $1 - F(a^-)$ |
| $X = a$ | $F(a) - F(a^-)$ |
| $a < X \leq b$ | $F(b) - F(a)$ |
| $a \leq X \leq b$ | $F(b) - F(a^-)$ |
| $a < X < b$ | $F(b^-) - F(a)$ |
| $a \leq X < b$ | $F(b^-) - F(a^-)$ |

- The student should not try to remember this list. Instead, he/she should be able to reconstruct it starting from basic properties of $F(x)$.

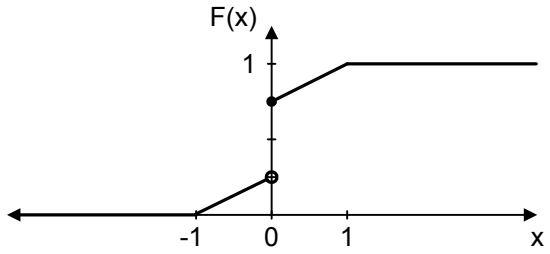
 

*Example 5.4:*

► Consider the following CDF:

$$F(x) = \begin{cases} 0 & x < -1, \\ (x+1)/4 & -1 \le x < 0, \\ (x+3)/4 & 0 \le x < 1, \\ 1 & 1 \le x \ . \end{cases}$$

whose graph is illustrated below: Using above formulae, we have, for example:




$$\begin{aligned} P(X < 0) &= F(0^-) = 1/4 \\ P(X \le 0) &= F(0) = 3/4 \\ P(X = 0) &= F(1) - F(1^-) = 1/2 \\ P(0 \le X \le 1) &= F(1) - F(0^-) = 3/4 \end{aligned}$$

etc.

◄

## 5.4 Classification of random variables

For the purpose of studying their properties, it is convenient to classify random variables according to the behavior of their CDF $F(x)$. Specifically:

- We say that a random variable $X$ is discrete if its CDF $F(x)$ is flat, except for a finite or countably infinite number of discontinuities (i.e. jumps). A basic example is the number of tails when flipping a coin twice, as discussed in Example 5.2.
- We say that a random variable $X$ is continuous if its CDF $F(x)$ is an (absolutely) continuous function of $x$. A basic example of this would be the waiting time of a person at a bus stop.
- Otherwise we say that a random variable has a mixed behavior. The CDF $F(x)$ in Example 5.4 fits into this category.

In the next two Chapters, we study discrete and continuous RVs separately. In each case, we include:

- A formal definition of the class and related special properties of the CDF.
- Definition of the *expectation* operation, which provides an extension to the intuitive notion of averaging and plays a major role in applications.
- Study of particular RVs of interest.

This is followed by another Chapter which provides a unifying framework for the study of the above three types of RVs. The important concept of moment generating function is also covered.

# Chapter 6

# Discrete Random Variables

This Chapter focusses on discrete random variables, including:

- Formal definition
- Probability mass function
- Expected value and variance
- Standard discrete RVs of interest (Binomial, Poisson, etc.)

## 6.1 Basic concepts

*Definition:* Let $X$ be a random variable defined on probability space $(S, \mathcal{F}, P)$. We say that $X$ is *discrete* if its CDF is a *step function*. That is, there exists a finite or countably infinite set of points, say $\{x_i\}$, such that

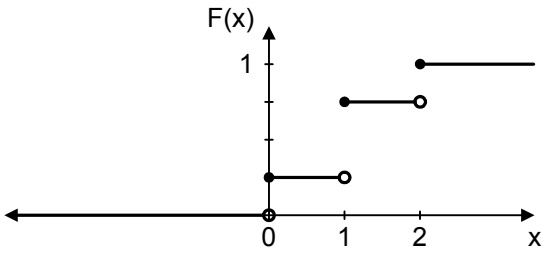
(a) $F(x)$ admits a positive step (or jump) at any point $x_i$;

(b) $F(x)$ remains constant (flat) between any two consecutive jump points, that is, the derivative

$$F'(x) = 0, \quad \text{for all } x \in \mathbb{R} - \{x_i\}. \tag{6.1}$$

*Example 6.1:*

▶ Let the random variable $X$ represents the number of tails obtained when flipping a fair coin twice. Its cumulative distribution function $F(x)$, derived in Example 5.2, is reproduced below for convenience.




It is clear that $F(x)$ is a step function and accordingly, $X$ qualifies has a discrete random variable. Specifically, there are three points of discontinuity at $x = 0$, 1 and 2. The corresponding jumps are

$$F(0) - F(0^-) = P(X = 0) = 1/4$$
$$F(1) - F(1^-) = P(X = 1) = 1/2$$
$$F(2) - F(2^-) = P(X = 2) = 1/4$$

Between consecutive points of discontinuity, the function $F(x)$ remains constant.

◀

*Remarks:*

- Discrete random variables are easily identifiable:
  - Any RV $X$ defined over a discrete (i.e. finite or countably infinite) sample space $S$ is necessarily discrete.
  - More generally, any RV $X$ with a discrete range $\mathcal{R}_X$ must be discrete.
- Invoking Theorem 5.1 (e), the value of the jump in $F(x)$ at the point $x_i$ is given by

$$P(X = x_i) = F(x_i) - F(x_i^-) > 0 \tag{6.2}$$

  At any other value of $x \in \mathbb{R} - \{x_i\}$, the function $F(x)$ is continuous and therefore

$$P(X = x) = F(x) - F(x^-) = 0 \tag{6.3}$$

- We refer to $\{x_i\}$ as the *set of possible values* for $X$. While $\{x_i\} \subseteq \mathcal{R}_X$, the converse is not necessarily true. However, since $P(X = x) = 0$ for any $x \in \mathcal{R}_X$ that is not in $\{x_i\}$, it can be seen that for the purpose of computing probabilities, both sets are equivalent.
- In the sequel, we assume that $\mathcal{R}_X$ is discrete and we identify $\{x_i\} \equiv \mathcal{R}_X$.

*Definition:* Let $X$ be a discrete RV. The function $p : \mathbb{R} \to [0, 1]$ defined by

$$p(x) = P(X = x), \quad \text{for all } x \in \mathbb{R} \tag{6.4}$$

is called the *probability mass function* (PMF) of $X$.

*Remarks:*

- $p(x)$ is sometimes called the *discrete probability function*.
- From (6.2) and (6.3), we immediately obtain:

$$p(x_i) = F(x_i) - F(x_i^-) > 0 \tag{6.5}$$

$$p(x) = 0 \quad \text{for all } x \notin \mathcal{R}_X \tag{6.6}$$

- It should be clear that knowledge of the CDF $F(x)$ is sufficient to construct the PMF $p(x)$, and vice versa. In particular

$$\begin{aligned} F(x) &= \sum_{\text{all } i} p(x_i)u(x - x_i) \\ &= \sum_{x_i \leq x} p(x_i) \end{aligned} \tag{6.7}$$

  where $u(x)$ is the unit step function defined by

$$u(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6.8}$$

- While both functions $F(x)$ and $p(x)$ convey the same information, it is often preferable to work with $p(x)$ in applications, as it usually simplifies the computation of probabilities.
- Knowledge of $p(x)$ is extremely important from the viewpoint of computing probabilities of events related to RV $X$. Indeed, as we will shortly explain, any probability of the type $P(X \in D)$, where $D \subseteq \mathbb{R}$, can be expressed in terms of $p(x)$.

*Example 6.2:*

► A random experiment consists in rolling a fair die twice. Let $X$ represent the sum of the two numbers so obtained. Find the discrete probability function of $X$, i.e. $p(x)$.

*Solution:* An appropriate sample space is

$$S = \{(i,j) \in \mathbb{N}^2 : 1 \leq i,j \leq 6\}$$

which contains $N(S) = 36$ outcomes. We take the power set of $S$ as event algebra: $\mathcal{F} = \mathcal{P}_S$. Because the die is assumed to be fair, we use an equiprobable model. Thus, for any individual outcome $(i,j) \in S$, we have

$$P(\{(i,j)\}) = \frac{1}{N(S)} = \frac{1}{36}$$

Let $X$ be the random variable representing the sum of the two numbers. $X$ is a function from $S$ into $\mathbb{R}$, defined by

$$X(i,j) = i + j, \quad \text{for all } (i,j) \in S$$

We note that $(i,j) \in S \Rightarrow 2 \leq i+j \leq 12$. Thus, the range of $X$, or equivalently, the set of its possible values, is given by

$$\mathcal{R}_X = \{2, 3, ..., 12\}$$

Values of the PMF $p(x)$ may be computed as follows:

$$\begin{aligned} p(2) = P(X=2) &= P(\{(i,j) \in S : X(i,j) = i+j = 2\}) \\ &= P(\{(1,1)\}) \\ &= 1/36 \\ p(3) = P(X=3) &= P(\{(i,j) \in S : X(i,j) = i+j = 3\}) \\ &= P(\{(1,2),(2,1)\}) \\ &= 2/36 \end{aligned}$$

Proceeding in this way for the other possible values of $X$, we obtain:

| $x$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $p(x)$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

Finally, note that

$$p(x) = 0 \text{ if } x \notin \{2, 3, \ldots, 12\}$$

◄

*Theorem 6.1:* Let $X$ be a discrete RV with range $\mathcal{R}_X = \{x_1, x_2, ...\}$. The PMF of $X$ obeys the following properties:

$$(a) \qquad p(x) = 0 \quad \text{for all } x \notin \mathcal{R}_X \tag{6.9}$$

$$(b) \qquad p(x) \geq 0 \quad \text{for all } x \in \mathcal{R}_X \tag{6.10}$$

$$(c) \qquad \sum_{x \in \mathcal{R}_X} p(x) \equiv \sum_{\text{all } i} p(x_i) = 1 \tag{6.11}$$

*Proof:* Properties (a) and (b) are merely a restatement of (6.5)-(6.6). Part (c) may be proved by combining (6.7) and Theorem 5.1 (b) as follows:

$$\begin{aligned} 1 &= \lim_{x \to \infty} F(x) \\ &= \lim_{x \to \infty} \sum_{\text{all } i} p(x_i) u(x - x_i) = \sum_{\text{all } i} p(x_i) \quad \square \end{aligned}$$

*Remarks:*

- Generally, any function $p(x)$ satisfying properties (a)-(c) above is called a probability mass function.
- One may view $\mathcal{R}_X$ together with the numbers $p_i \triangleq p(x_i)$ as defining a *simplified* probability space, adequate for the study of RV $X$.
- For any real number subset $D \subseteq \mathbb{R}$, the probability $P(X \in D)$ can be expressed as a sum of numbers $p(x_i)$. Specifically:

$$P(X \in D) = \sum_{i \,:\, x_i \in D} P(X = x_i) = \sum_{i \,:\, x_i \in D} p(x_i) \tag{6.12}$$

  where the sum is over all $i$ such that $x_i \in D$.
- In particular, if $D \cap \mathcal{R}_X = \emptyset$, then $P(X \in D) = 0$

*Example 6.3:*

► A random experiment consists in flipping a fair coin until heads shows up; assume that each flip is an independent sub-experiment. Let $X$ represent the number of necessary flips. Find the PMF of $X$ and compute the probability that $X$ is even. ◄

## 6.2 Function of a random variable

*Introduction:*

- We often have to deal with two or more RVs that are related to each other by simple functional relationships, or transformations.
- Here, we briefly expose the concept of a transformation from one discrete RV, say $X$, to another RV, say $Y$.
- Special attention is given to the relationship between the PMFs of $X$ and $Y$.

*Transformation of a single RV:*

- Let $X : S \to \mathbb{R}$ be a discrete RV defined on probability space $(S, \mathcal{F}, P)$ and let $\mathcal{R}_X$ denote the range of $X$ (assumed to be discrete).
- Let $h : \mathbb{R} \to \mathbb{R}$ be a real-valued function. The composition $Y = h \circ X$ is a function from $S$ into $\mathbb{R}$, defined by

$$Y(s) = h(X(s)) \text{ for all } s \in S \tag{6.13}$$

- Let $\mathcal{R}_Y$ denote the range of $Y$. Since $\mathcal{R}_X$ is discrete and $\mathcal{R}_Y = h(\mathcal{R}_X)$, it follows that $\mathcal{R}_Y$ is also discrete. Thus, the function $Y = h \circ X$ defines a discrete RV.
- Typical examples include: $Y = X^2$, $Y = \cos(X)$, etc.

*Theorem 6.2:* Let $X$ be a discrete RV with range $\mathcal{R}_X = \{x_1, x_2, \ldots\}$ and PMF $p_X(x)$. Consider the discrete RV $Y = h(X)$ with range $\mathcal{R}_Y = \{y_1, y_2, \ldots\}$ and let $p_Y(y) = P(Y = y)$ denote its PMF. We have

$$p_Y(y_j) = \sum_{h(x_i)=y_j} p_X(x_i) \tag{6.14}$$

where the summation is over all integer $i$ such that $h(x_i) = y_j$.

*Proof:* For any $y_j \in \mathcal{R}_Y$, the event $\{Y = y_j\}$ may be expressed as a union of mutually exclusive events as follows (please think...):

$$\{Y = y_j\} = \bigcup_{h(x_i)=y_j} \{X = x_i\} \tag{6.15}$$

Therefore, we have

$$P(Y = y_j) = \sum_{h(x_i)=y_j} P(X = x_i) \tag{6.16}$$

which is equivalent to (6.14). □

*Example 6.4:*

► A fair die is rolled twice. Let $X$ represent the sum of the two numbers so obtained and define $Y = |X - 7|$. Find the probability mass function (PMF) of $Y$. ◄

## 6.3 Expectation of a discrete RV

*Motivation:* Consider a lottery game in which the probability of winning each one of three possible dollar prizes is:

$$P(\text{Winning } x \text{ dollars}) = 0.25/x, \quad x = 1, 10, 100$$

How much would you be willing to pay for the price of a ticket?

An answer to this question may be obtained via the relative frequency interpretation of probabilities. Let $X$ represent our gain in dollars each time we buy a ticket. $X$ may be viewed as a discrete RV with range $\mathcal{R}_X = \{1, 10, 100\}$ and PMF

$$p(x) = P(X = x) = \begin{cases} 0.25/x, & x = 1, 10, 100 \\ 0, & \text{otherwise} \end{cases}$$

Suppose we were allowed to play the game a large number of times, say $n$. For $x \in \mathcal{R}_X$, let

$$n_x = \text{number of times we win } x \text{ dollars}$$

out of the $n$ trials. The *arithmetic average* of the observed values of the gain $X$ in $n$ trials of the game can be computed as

$$\begin{aligned} \text{Average gain} &= (1\,n_1 + 10\,n_{10} + 100\,n_{100})/n \\ &= 1\,(n_1/n) + 10\,(n_{10}/n) + 100\,(n_{100}/n) \\ &\approx 1\,p(1) + 10\,p(10) + 100\,p(100) \qquad (6.17) \\ &= 1\,(0.25) + 10\,(0.025) + 100\,(0.0025) = 0.75 \end{aligned}$$

where we have used the approximation $p(x) \approx n_x/n$.

C onclusion: we should not pay more than 75 cents for the price of a ticket. On average, we are going to lose money if we pay more than that!

Note how in (6.17), the average gain is expressed in terms of the PMF of random variable $X$ as

$$\text{Average gain } \approx \sum_{x \in \mathcal{R}_X} x\, p(x) \tag{6.18}$$

where the summation is over all $x \in \mathcal{R}_X = \{1, 10, 100\}$. This leads naturally to the following definition.

*Definition:* Let $X$ be a discrete RV with set of possible values $\mathcal{R}_X = \{x_1, x_2, ...\}$ and PMF $p(x)$. The expected value of $X$ is defined as

$$E(X) = \sum_{x \in \mathcal{R}_x} x\, p(x) = \sum_{\text{all } i} x_i\, p(x_i) \tag{6.19}$$

provided the series converges absolutely.

*Remarks:*

- $E(X)$ is also called mean, expectation or mathematical expectation; it is often denoted simply by $\mu$ or $\mu_X$.
- Knowledge of $E(X)$ is very useful in decision making processes (e.g.: should you play a game or not). In fact, the solution of many engineering problems amounts to optimizing an expected value (e.g. minimum mean square-error design of a digital radio receiver).
- Formally, $E(X)$ is only defined if the series (6.19) converges absolutely, that is $\sum_{x \in \mathcal{R}_X} |x|\, p(x) < \infty$:
  - when $\mathcal{R}_X$ is finite, this condition is always satisfied;
  - when $\mathcal{R}_X$ is countably infinite, it may not be satisfied.

*Example 6.5:*

► In a 6-49 lottery, players pick 6 different integers in $\{1, 2, ..., 49\}$. The lottery commission also picks 6 of these numbers randomly as the winning combination. A player wins the

- Grandprize of \$2,400,000 if all 6 numbers match winning combination
- 2nd prize of \$1600 if 5 out of 6 matching numbers
- 3rd prize of \$70 if 4 out of 6 matches

What is the expected value of the amount a player can win in this game?

*Solution:* Let RV $X$ denote the gain in \$. $X$ may take the following values with corresponding probabilities:

$$x_1 = 2.4 \times 10^6 \quad \text{with} \quad p(x_1) = P(X = x_1) = 1/\binom{49}{6} = 7.1 \times 10^{-8}$$

$$x_2 = 1600 \quad \text{with} \quad p(x_2) = \binom{6}{5}\binom{43}{1}/\binom{49}{6} = 1.85 \times 10^{-5}$$

$$x_3 = 70 \quad \text{with} \quad p(x_3) = \binom{6}{4}\binom{43}{2}/\binom{49}{6} = 9.69 \times 10^{-4}$$

Using the above figures, the expected gain can be computed as

$$E(X) = x_1 p(x_1) + x_2 p(x_2) + x_3 p(x_3) = 0.26\$$$

In practice, this value of $E(X)$ would be used by the lottery commission to set the price of a ticket. ◄

### 6.3.1 Properties of expectation:

In this Section, we state important properties of the expectation operator $E(.)$ in the form of individual Theorems. These properties in handy when performing computations involving $E(.)$; we shall often use and/or refer to them. In the Theorem statements, the following is implicitly assumed

- $X$ is a discrete RV
- $\mathcal{R}_X = \{x_1, x_2, ...\}$ is the range of $X$ (i.e. set of possible values)
- $p(x) = P(X = x)$ is the probability mass function (PMF) of $X$.

*Theorem 6.3:* If $X$ is constant with probability one, that is, if $P(X = c) = 1$ for some constant $c$, then $E(X) = c$.

*Proof:* Recall that $\sum_{x\in\mathcal{R}_X} p(x) = 1$. Here, since $p(c) = P(X = c) = 1$, it follows that $p(x) = 0$ if $x \neq c$. Then:

$$\begin{aligned} E(X) &= \sum_{x\in\mathcal{R}_X} x\, p(x) \\ &= c\, p(c) + \sum_{x\in\mathcal{R}_X, x\neq c} x\, p(x) = c \cdot 1 + 0 = c \quad \square \end{aligned}$$

*Theorem 6.4:* Let $h : \mathbb{R} \to \mathbb{R}$ is be a real-valued function:

$$E(h(X)) = \sum_{x\in\mathcal{R}_X} h(x)p(x) = \sum_{\text{all } i} h(x_i)p(x_i) \tag{6.20}$$

*Proof:* Define discrete random variable $Y = gh(X)$ and let $\mathcal{R}_Y = \{y_1, y_2, ...\}$ denote its corresponding set of possible values. Invoking the definition of expectation (6.19) and Theorem 6.2, we have:

$$\begin{aligned} E(h(X)) &= E(Y) \\ &= \sum_{\text{all } j} y_j \, p_Y(y_j) \\ &= \sum_{\text{all } j} y_j \sum_{h(x_i)=y_j} p(x_i) \\ &= \sum_{\text{all } j} \sum_{h(x_i)=y_j} h(x_i) \, p(x_i) \\ &= \sum_{\text{all } i} h(x_i) \, p(x_i) \quad \square \end{aligned}$$

*Why is Theorem 6.4 useful?*

- Let $Y = h(X)$. Then, from the definition of expectation:

$$E(h(X)) = E(Y) = \sum_j y_j \, p_Y(y_j) \tag{6.21}$$

  where $p_Y(y) = P(Y = y)$ is the PMF of $Y$.

- According to Theorem 6.4, we do not need to know $p_Y(y)$ explicitly to evaluate $E(h(X))$. Knowledge of $p(x)$, the PMF of $X$, is sufficient.

- A common mistake for students at this stage is to assume that $E(h(X)) = h(E(X))$. This is not true in general, as we will see in examples.

*Corollary 6.4:* Let $h_1, h_2, \ldots, h_n$ be real-valued functions and $\alpha_1, \alpha_2, \ldots, \alpha_n$ be real constants:

$$E(\sum_{k=1}^{n} \alpha_k h_k(X)) = \sum_{k=1}^{n} \alpha_k E(h_k(X)) \tag{6.22}$$

*Proof:* Applying Theorem 6.4 to the left-hand side of (6.22), we have

$$\begin{aligned} E(\sum_{k=1}^{n} \alpha_k h_k(X)) &= \sum_{x \in \mathcal{R}_X} (\sum_{k=1}^{n} \alpha_k h_k(x))\, p(x) \\ &= \sum_{k=1}^{n} \alpha_k \sum_{x \in \mathcal{R}_X} h_k(x)\, p(x) \\ &= \sum_{k=1}^{n} \alpha_k E(h_k(X)) \quad \square \end{aligned}$$

*Remarks:*

- $E(.)$ acts as a linear operator on the RVs $h_k(X)$.
- Let $\alpha$ and $\beta$ be arbitrary real constants. As a special case of (6.22):

$$E(\alpha X + \beta) = \alpha E(X) + \beta \tag{6.23}$$

*Example 6.6:*

▶ Equation (6.22) is typically used to break down the computation of an expectation into simpler parts, as in

$$\begin{aligned} E((X+1)^3) &= E(X^3 + 3X^2 + 3X + 1) \\ &= E(X^3) + 3E(X^2) + 3E(X) + 1 \end{aligned}$$

$$E(2\cos(X) + e^{-X}) = 2E(\cos(X)) + E(e^{-X})$$

◀

*Example 6.7:*

► Let $X$ be a randomly selected integer from the set $\{0, 1, ..., N\}$, where $N$ is a given positive integer. Find the expected value of

$$Y = X(N - X)$$

Also, verify that $E(Y) \neq E(X)(N - E(X))$.

*Solution:* In the absence of further a priori knowledge, we assume an equiprobable model for $X$. Thus, its PMF is given by

$$p(x) = P(X = x) = \begin{cases} \frac{1}{N+1}, & x \in \mathcal{R}_X = \{0, 1, ..., N\} \\ 0, & \text{otherwise.} \end{cases}$$

To compute the expected value of $Y$, we proceed as follows:

$$\begin{aligned} E(X) &= \sum_{x=0}^{N} xp(x) = \frac{1}{N+1} \sum_{x=0}^{N} x \\ &= \frac{1}{N+1} \frac{N(N+1)}{2} = \frac{N}{2} \end{aligned}$$

$$\begin{aligned} E(X^2) &= \sum_{x=0}^{N} x^2 p(x) = \frac{1}{N+1} \sum_{x=0}^{N} x^2 \\ &= \frac{1}{N+1} \frac{N(N+1)(2N+1)}{6} = \frac{N(2N+1)}{6} \end{aligned}$$

$$\begin{aligned} E(X(N - X)) &= E(NX - X^2) \\ &= NE(X) - E(X^2) \\ &= N\frac{N}{2} - \frac{N(2N+1)}{6} = \frac{N(N-1)}{6} \end{aligned}$$

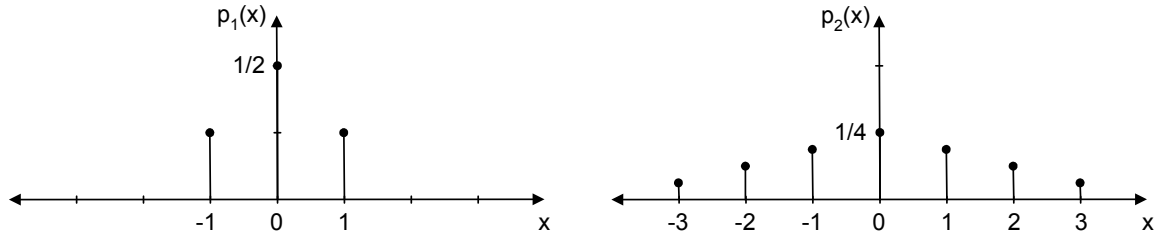
Note that

$$E(X)(N - E(X)) = \frac{N}{2}(N - \frac{N}{2}) = \frac{N^2}{4} \neq E(Y) = \frac{N(N-1)}{6}$$

◄

## 6.4 Variance of a discrete random variable

*Introduction:* Consider 2 discrete RVs, say $X_1$ and $X_2$, with PMF $p_1(x)$ and $p_2(x)$, respectively, and identical mean, say $\mu = E(X_1) = E(X_2)$. Although both $X_1$ and $X_2$ have the same mean, their statistical behavior around $\mu$, as characterized by the size and frequency of the deviation $X_i - \mu$, may be quite different. To illustrate this point, consider the PMF illustrated below:




Clearly, the likelihood that $X_2$ be found far away from its mean $\mu = 0$ is larger than that for $X_1$.

In many applications, the deviation of a RV about its mean is of great significance. For example, in the above example,

- Suppose $X_1$ and $X_2$ represent the distribution of voltage measurements across an open circuit using two different digital instruments, say I and II, respectively.
- The measurements $X_i$ $(i = 1, 2)$ have a random nature due to the inherent errors generated within the instruments.
- Based on the above PMF, we can affirm that instrument I is superior to instrument II, as its measurements are less likely to deviate from the mean value, assumed equal to the true voltage.

For the purpose of comparison, it is important to introduce a quantitative measure of the spread of a PMF. The variance fulfills this role.

*Definition:* Let $X$ be a discrete RV with range $\mathcal{R}_X = \{x_1, x_2, ...\}$, PMF $p(x)$ and mean value $E(X) = \mu$. The variance of $X$ is defined as

$$Var(X) = E[(X - \mu)^2] \tag{6.24}$$

The standard deviation of $X$ is defined as

$$\sigma_X = \sqrt{Var(X)} \tag{6.25}$$

*Remarks:*

- The variance really is a characteristic of the PMF $p(x)$:

$$Var(X) = \sum_{x \in \mathcal{R}_X} (x - \mu)^2 p(x) = \sum_i (x_i - \mu)^2 p(x_i) \tag{6.26}$$

- $Var(X)$ measures the dispersion, or spread, of $X$ about its mean $\mu$.
- Difference between $Var(X)$ and $\sigma_X$:
  - $Var(X)$ is in unit of $X^2$ while
  - $\sigma_X$ is in unit of $X$.
- An alternative measure of spread is $E[|X - \mu|]$. However, because of the absolute values, this measures is less mathematically tractable. The measure $E[(X - \mu)^2]$ is preferred in practice.

*Theorem 6.5:* The variance of $X$ satisfies:

$$Var(X) = E(X^2) - \mu^2 \tag{6.27}$$

*Proof:* Recall that $\mu = E(X)$. We have

$$\begin{aligned} Var(X) &= E((X-\mu)^2) \\ &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2 \quad \square \end{aligned} \tag{6.28}$$

*Remarks:*

- It is often simpler to evaluate $E(X^2)$ than $E((X-\mu)^2)$. Theorem 6.5 simply offers an alternative way of computing $Var(X)$.
- Since $Var(X) \geq 0$, it follows from Theorem 6.5 that $E(X^2) \geq (E(X))^2$.

*Theorem 6.6:* $Var(X) = 0$ if and only if $P(X = \mu) = 1$

*Proof:* Left as an exercise to the student.

*Theorem 6.7:* For any real constants $a$ and $b$,

$$Var(aX + b) = a^2\, Var(X) \tag{6.29}$$

*Proof:* Using the definition (6.24) of the variance, we have:

$$\begin{aligned} Var(aX+b) &= E[(aX+b) - E(aX+b))^2] \\ &= E[(aX + b - aE(X) - b)^2] \\ &= E[a^2(X - E(X))^2] \\ &= a^2 E[(X - E(X))^2] = a^2 Var(X) \quad \square \end{aligned} \tag{6.30}$$

*Remarks:*

- From (6.29), we conclude that $Var(.)$ is not a linear operation.
- In particular, $Var(aX + b) \neq aVar(X) + b$ in general

*Example 6.8:*

▶ Let $X$ be a randomly selected integer from the set $\{-N, ..., -1, 0, 1, ..., N\}$. Find the standard deviation of $X$. ◀

## 6.5 Discrete RVs in repeated experiments

*Introduction:*

- In this and the next Section, we study some common discrete RVs of interest in science and engineering.
- These RVs should be viewed as basic building blocks when developing probability models for specific problems and applications.
- In this Section, we study discrete RVs that relate to sequences of identical and independent random experiments, i.e. Binomial and geometric RVs.
- In the next Section, we look at the Poisson RV.

### 6.5.1 Bernouilli RV

*Recall:* A Bernouilli trial is a random experiment in which a particular event $A$, that may or not occur, has been identified and assigned a probability

$$p \triangleq P(A), \quad 0 \leq p \leq 1. \tag{6.31}$$

Event $A$ is called a success and its complement $A^c$ is called a failure. The number $p = P(A)$ is called the probability of success and

$$q \triangleq P(A^c) = 1 - p \tag{6.32}$$

is called probability of failure.

*Definition:* A random variable $X$ is called Bernouilli with parameter $p$ if there exists an event $A$ with probability $p = P(A)$ such that:

$$X(s) = \begin{cases} 1, & s \in A \\ 0, & s \notin A. \end{cases} \tag{6.33}$$

where $s$ denotes an arbitrary experimental outcome in the sample space $S$. In other words, $X = 1$ if event $A$ occurs, and $X = 0$ otherwise.

*Probability mass function of $X$:*

- $X$ is a discrete RV with only two possible values: $\mathcal{R}_X = \{0, 1\}$.
- The probability mass function of $X$ is given by

$$p(0) = P(X = 0) = P(A^c) = 1 - p = q \tag{6.34}$$

$$p(1) = P(X = 1) = P(A) = p \tag{6.35}$$

$$p(x) = 0 \text{ if } x \notin \{0, 1\} \tag{6.36}$$

- Graph of $p(x)$:

*Expected value:*

$$\begin{aligned} \mu = E(X) &= \sum_{i=0}^{1} i\, p(i) \\ &= 0 \cdot q + 1 \cdot p = p \end{aligned} \tag{6.37}$$

*Variance:*

$$\begin{aligned} E(X^2) &= \sum_{i=0}^{1} i^2\, p(i) \\ &= 0^2 \cdot q + 1^2 \cdot p = p \end{aligned} \tag{6.38}$$

$$\begin{aligned} \sigma^2 = Var(X) &= E(X^2) - \mu^2 \\ &= p - p^2 = p\,(1-p) = p\,q \end{aligned} \tag{6.39}$$

*Remarks:*

- The Bernouilli RV is one of the simplest RV that can be imagined.
- It is of limited use by itself, but extremely useful as a building block in the development of models for repeated experiments.

### 6.5.2 Binomial RV

*Definition:* Consider a sequence of $n$ identical and independent Bernouilli trials with probability of success $p$. The RV $X$ defined by

$$X = \text{ number of successes in the } n \text{ trials} \tag{6.40}$$

is called binomial with parameters $n$ and $p$, or simply $B(n,p)$.

*Remarks:*

- The notation $X \sim B(n,p)$ is also used.
- Each Bernouilli trial is a random experiment with sample space $S$ and $P(A) = p$ for some selected event $A \subseteq S$ ($A \equiv$ success).
- The sample space of the product experiment, i.e. the sequence of $n$ Bernouilli trials, is the cartesian product $S^n$.
- RV $X$ is a function from $S^n \rightarrow \mathcal{R}_X = \{0, 1, ..., n\}$.
- Basic examples of Binomial RVs include:
  - Number of heads in a sequence of 10 independent flips of a coin.
  - Number of defective IC chips in a production sample of size $n$.

*Theorem 6.8:* Let $X$ be $B(n,p)$. The PMF of $X$ is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, ..., n \tag{6.41}$$

and $p(x) = 0$ otherwise.

*Proof:* $X$ represents the number of successes in a sequence of $n$ Bernouilli trials; it is a discrete RV with range $\mathcal{R}_X = \{0, 1, ..., n\}$. As an immediate application of Theorem 4.8, we have for any $x \in \mathcal{R}_X$:

$$\begin{aligned} p(x) &= P(X = x) \\ &= P(\{x \text{ successes in } n \text{ trials}\}) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \end{aligned}$$

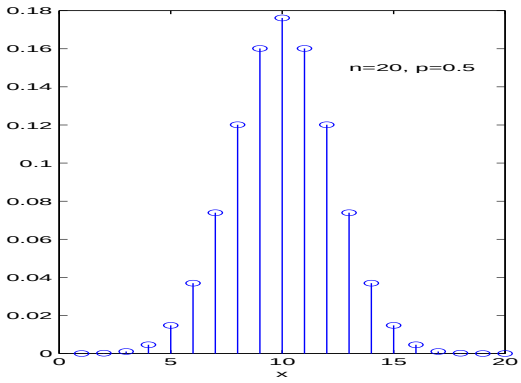
For any $x \notin \mathcal{R}_X$, $p(x) = P(X = x) = 0$. □

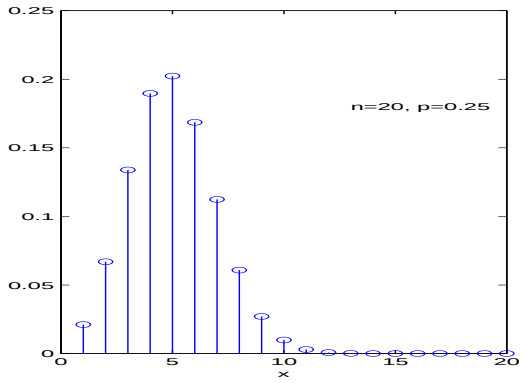
*Example 6.9:*

▶ Consider an unmanned space program in which the probability of a successful launch of a certain type of rocket has been evaluated to $p = 0.975$. Consider a sequence of 50 independent launches. Find the probability that all launches are sucessful. Find the probability of a single unsuccessful launch. ◀

*Remarks on the function $p(x)$:*

- The probability mass function (6.41) is sometimes called binomial law or binomial distribution.
- In the special cases $p = 0$ or $p = 1$ (i.e. $q = 1 - p = 0$), (6.41) remains valid provided one assumes $0^0 = 1$.
- Typical plots of $p(x)$:




- In the special case $p = q = 1/2$, the function $p(x)$ is symmetrical:

$$p(x) = p(n - x), \ x \in \{0, 1, ..., n\} \tag{6.42}$$

- In the case $0 < p < 1$, the function $p(x)$ attains a maximum at the largest integer less than or equal to $(n + 1)p$, sometimes denoted as $\lfloor (n + 1)p \rfloor$. To verify this, first note that

$$\frac{p(x)}{p(x-1)} = \frac{\binom{n}{x} p^x (1-p)^{n-x}}{\binom{n}{x-1} p^{x-1} (1-p)^{n-x+1}} = \frac{(n+1)p - xp}{x - xp} \tag{6.43}$$

From (6.42) it follows that

$$\begin{aligned} x < (n+1)p &\Rightarrow p(x) > p(x-1) \\ x = (n+1)p &\Rightarrow p(x) = p(x-1) \\ x > (n+1)p &\Rightarrow p(x) < p(x-1) \end{aligned}$$

*Theorem 6.9:* Let $X$ be $B(n,p)$. Then

$$E(X) = np \tag{6.44}$$

$$Var(X) = np(1-p) = npq \tag{6.45}$$

*Proof:* First consider $E(X)$:

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} \\ &= \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x} \\ &= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!(n-1-(x-1))!} p\, p^{x-1} q^{n-1-(x-1)} \end{aligned} \tag{6.46}$$

Making the change of variable $y = x-1$ and $m = n-1$ in (6.46), we obtain

$$\begin{aligned} E(X) &= np \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^y q^{m-y} \\ &= np(p+q)^m = np \end{aligned} \tag{6.47}$$

where we have made use of Theorem 2.10. The proof for the variance is left as an exercise. □.

### 6.5.3 Geometric RV

*Definition:* Consider a sequence of independent Bernouilli trials, each with probability of success $p$. The RV $X$ defined by

$$X = \text{ number of trials until first success} \tag{6.48}$$

is called geometric with parameter $p$.

*Remarks:*

- The set of possible values for $X$ is the set of positive integers, that is $\mathcal{R}_X = \{1, 2, \ldots\} = \mathbb{N}$.
- Thus, $X$ is a discrete RV with a countably infinite set of possible values.

*Theorem 6.10:* Let $X$ be geometric with parameter $p$. The PMF of $X$ is

$$p(x) = (1-p)^{x-1}p, \quad x = 1, 2, 3, ... \tag{6.49}$$

and $p(x) = 0$ otherwise.

*Proof:* Consider the following tree diagram of the underlying random experiment, where letters $S$ and $F$ indicate success and failure, respectively:

For any $x \in \mathbb{N}$, the event $X = x$ is equivalent to a succession of $x - 1$ failures followed by a success, as represented by the outcome $FF...FS$. The corresponding probability is therefore

$$p(x) = P(X = x) = P(\{FF \dots FS\}) = q^{x-1}p \tag{6.50}$$

where the independence assumption as been used in the last equality. □

*Theorem 6.11:* Let $X$ be geometric with parameter $p$. Then

$$E(X) = \frac{1}{p} \tag{6.51}$$

$$Var(X) = \frac{1-p}{p^2} \tag{6.52}$$

*Proof:* We make use of the following basic relations:

$$\sum_{k=1}^{\infty} k\rho^k = \frac{\rho}{(1-\rho)^2}, \qquad \sum_{k=1}^{\infty} k^2\rho^k = \frac{\rho(\rho+1)}{(1-\rho)^3} \tag{6.53}$$

valid for any number $\rho$ with $|\rho| < 1$. First consider $E(X)$:

$$\begin{aligned} E(X) &= \sum_{x=1}^{\infty} x\, p\, q^{x-1} = \frac{p}{q}\sum_{x=1}^{\infty} x\, q^x \\ &= \frac{p}{q}\,\frac{q}{(1-q)^2} = \frac{1}{p} \end{aligned}$$

The proof for the variance is left as an exercise. □.

## 6.6 Poisson RV

*Definition:* A discrete random variable $X$ is called Poisson with parameter $\lambda > 0$ if its discrete probability function takes the form

$$p(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, ... \tag{6.54}$$

and $p(x) = 0$ otherwise.

*Remarks:*

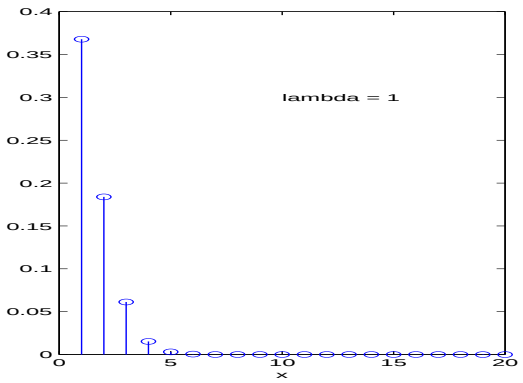
- The function $p(x)$ (6.55) is a valid PMF: $p(x) \geq 0$ for all $x$ and

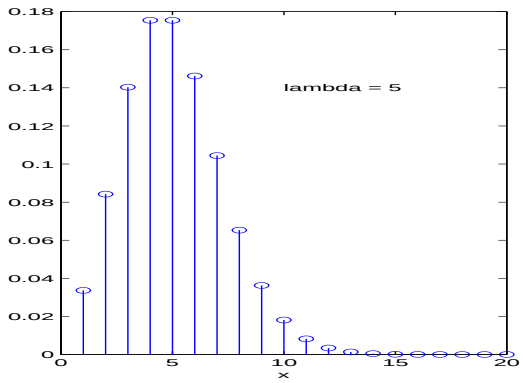
$$\sum_{x=0}^{\infty} p(x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1 \tag{6.55}$$

- Also known as the Poisson distribution, it first appeared as an approximation to the Binomial distribution (Poisson, 1837).

- It is one of the most commonly used distribution for discrete RVs. As we will see, it is the model of choice for many practical situations.

- Typical plots of the Poisson distribution:



*Theorem 6.12:* Let $X$ be a Poisson RV with parameter $\lambda$. Then

$$E(X) = \lambda \tag{6.56}$$

$$Var(X) = \lambda \tag{6.57}$$

*Proof:* We leave it to the reader to first demonstrate the following identities:

$$\sum_{x=0}^{\infty} x \, \frac{\lambda^x}{x!} = \lambda e^{\lambda} \tag{6.58}$$

$$\sum_{x=0}^{\infty} x^2 \, \frac{\lambda^x}{x!} = (\lambda^2 + \lambda) e^{\lambda} \tag{6.59}$$

Using (6.58), we have for the expected value of $X$:

$$E(X) = \sum_{x=0}^{\infty} x \, p(x) = \sum_{x=0}^{\infty} x \, \frac{\lambda^x}{x!} e^{-\lambda} = \lambda$$

To find the variance of $X$, first evaluate $E(X^2)$ using (6.59):

$$E(X^2) = \sum_{x=0}^{\infty} x^2 \, \frac{\lambda^x}{x!} e^{-\lambda} = \lambda^2 + \lambda \tag{6.60}$$

Finally,

$$Var(X) = E(X^2) - \lambda^2 = \lambda \quad \square$$

*Remarks:* The Poisson distribution occurs naturally in two very important classes of problems:

- Approximation to the binomial (presented below)
- Poisson processes (to be discussed in connection with random processes)

### 6.6.1 Poisson's approximation to the binomial:

*Historical perspective:*

- Consider the binomial distribution with parameters $n$ and $p$:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, ..., n \tag{6.61}$$

- Before the advent of electronic calculators, the evaluation of the binomial PMF for large $n$ and small $p$ posed a significant challenge.
- The following result, due to French mathematician Poisson (1837), provides a means for approximating a binomial PMF by the more tractable Poisson PMF, provided certain basic requirements are satisfied.

*Theorem 6.13:* For a fixed value of $x$, consider the limit of $p(x)$ in (6.61) when $n \to \infty$ and $p = \mu/n \to 0$, so that $np = \mu = E(X)$ remains constant. We have:

$$\lim_{n\to\infty} p(x) = \frac{e^{-\mu}\mu^x}{x!} \tag{6.62}$$

*Proof:* First express the binomial PMF as follows:

$$\begin{aligned} p(x) &= \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\ &= \frac{n(n-1)...(n-x+1)}{x!}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x} \\ &= (1-\frac{1}{n})(1-\frac{2}{n})...(1-\frac{x-1}{n})\,\frac{\mu^x}{x!}\,\frac{(1-\frac{\mu}{n})^n}{(1-\frac{\mu}{n})^x} \end{aligned} \tag{6.63}$$

Taking the limit as $n \to \infty$ and recalling (from basic calculus) that $\lim_{n\to\infty}(1-\frac{\mu}{n})^n = e^{-\mu}$, we obtain:

$$\lim_{n\to\infty} p(x) = \frac{e^{-\mu}\mu^x}{x!} \quad \square$$

*Basic approximation and notes:*

- Based on Theorem 6.13, we conclude that for $n \gg x$ and $p \ll 1$, the binomial PMF (6.61) may be approximated by a Poisson law with parameter $\lambda = \mu = np$:

$$p(x) = \binom{n}{x}p^x(1-p)^{n-x} \approx \frac{e^{-\lambda}\lambda^x}{x!} \tag{6.64}$$

- An important application of the Poisson law is indeed as an approximation of the binomial when $n$ is large and $p$ is small.

- Not only is the Poisson distribution easier to compute in this case, but its use often leads to important analytical simplifications.

- Typical examples of binomial RVs for which the Poisson approximation is particularly adequate include the following:
  - Number of persons affected by a rare disease (low $p$) in a large population (large $n$).
  - Number of bits received in error when transmitting a large binary file over a memoryless communication channel.
  - Number of misprints or typos in a document page.

*Example 6.10:*

▶ Consider the transmission of a binary packet of length $n = 1024$ bits over a noisy channel. Assume that each bit is transmitted independently of the others, with a probability of error of $p = 10^{-2}$. Let $X$ denote the total number of bits received in error. Evaluate the probability of an error free packet transmission. ◀

## Problems

1. Find the CDF of a Binomial random variable and sketch it.

2. Complete the proof of Theorem 6.9. That is, show that the variance of a $B(n,p)$ random variable $X$ is given by $Var(X) = npq$ where $q = 1-p$.

3. Prove the following identities, used in the proof of Theorem 6.12:

$$\sum_{x=0}^{\infty} x\,\frac{\lambda^x}{x!} = \lambda e^{\lambda}, \quad \sum_{x=0}^{\infty} x^2\,\frac{\lambda^x}{x!} = (\lambda^2+\lambda)e^{\lambda}$$

# Chapter 7

# Continuous Random Variables

In many applications of probability, we encounter random variables whose characteristics differ significantly from those associated to discrete RVs, as studied in Chapter 6. Specifically, denoting such a RV by $X$, we find that:

- The range of $X$ is not countable;
- There is no concentration of probability in the sense that $P(X = x) = 0$ for all $x \in \mathbb{R}$;
- The CDF $F(x)$ is smoothly increasing from 0 to 1 (i.e. no jump).

Typical examples of this include: the time interval between two random phone calls; the measurement error when using an analog instrument.

This Chapter focusses on the study of such RVs, commonly called continuous RVs. The following topics are covered:

- Definition and the probability density function;
- Transformation of continuous RV ($Y = g(X)$);
- Expectation and variance;
- Continuous RVs of interest, including Gaussian, uniform, exponential, etc.

## 7.1 Basic concepts

*Definition:* Let $X$ be a random variable defined on probability space $(S, \mathcal{F}, P)$. We say that $X$ is continuous if its CDF is *absolutely continuous*. That is:

(a) $F(x)$ is continuous everywhere, i.e. for all $x \in \mathbb{R}$;

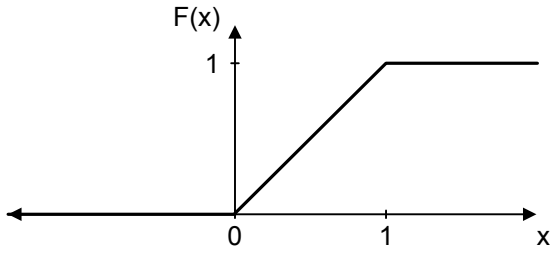
(b) The derivative $F'(x)$ exists everywhere, except possibly at a finite or countably infinite set of points in $\mathbb{R}$.

*Example 7.1:*

▶ Let $X$ denote a randomly selected point from the interval $[0, 1] \subseteq \mathbb{R}$. The CDF of $X$ is easily obtained as

$$F(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases} \tag{7.1}$$

The graph of $F(x)$ is shown below:




Clearly, $F(x)$ is a continuous function of $x$. Its derivative is obtained as

$$F'(x) = \begin{cases} 0, & x < 0 \\ 1, & 0 < x < 1 \\ 0, & x > 1 \end{cases} \tag{7.2}$$

It is defined everywhere except at the points $x = 0$ and $x = 1$. Thus, $F(x)$ is absolutely continuous and $X$ is a continuous RV. ◀

*Remark:* Absolute continuity implies continuity in the conventional sense. Thus, if $X$ is a continuous random variable, $F(x)$ is continuous everywhere and invoking Theorem 5.1 (e), we have that:

$$P(X = x) = F(x) - F(x^-) = 0, \quad \text{all } x \in \mathbb{R} \tag{7.3}$$

Consequently, for a continuous RV, the concept of a probability mass is meaningless.

*Definition:* Let $X$ be a continuous RV with CDF $F(x)$. The function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = F'(x) = \frac{dF(x)}{dx} \tag{7.4}$$

is called the *probability density function* (PDF) of $X$.

*Remarks:*

- The PDF $f(x)$ is uniquely defined and continuous everywhere, except at the points of discontinuity of $F'(x)$. In practice, we find that the particular value assigned to $f(x)$ at those isolated points is irrelevant.
- Knowledge of the CDF $F(x)$ enables one to construct the associated PDF $f(x)$. Conversely, if $f(x)$ is known, it is possible to recover $F(x)$ (see Theorem below).

*Theorem 7.1:*

$$F(x) = \int_{-\infty}^{x} f(t)\, dt, \quad \text{for all } x \in \mathbb{R}. \tag{7.5}$$

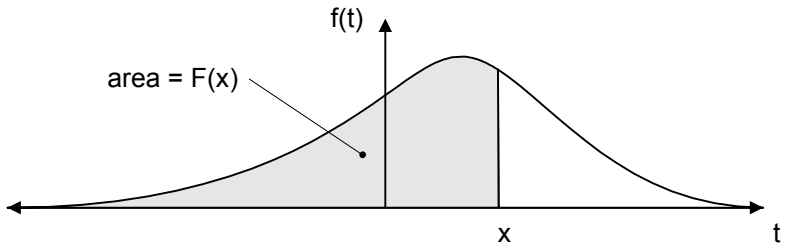
*Proof:* Invoking fundamental theorem from Calculus, it follows from (7.4) that for any real number $c$, we can write:

$$F(x) = F(c) + \int_{c}^{x} f(t)\, dt$$

Taking the limit as $c \to -\infty$ and using Theorem 5.1 (c), we obtain the desired result.

*Remarks:*

- According to (7.5), $F(x) = P(X \leq x)$ is equal to the area under the graph of $f(t)$ from $t = -\infty$ to $t = x$:




- While both $F(x)$ and $f(x)$ convey the same information, it is often preferable to work with the PDF in the computation of probabilities related to RV X.
- The PDF $f(x)$ plays a central role in the theory and application of continuous RVs. As we will see, any probability of the type $P(X \in A)$, where $A$ is some practical subset of real numbers, can be expressed in terms of $f(x)$.

*Theorem 7.2:* The PDF $f(x)$ satisfies the following basic properties:

(a) Non-negativity:

$$f(x) \geq 0, \quad x \in \mathbb{R} \tag{7.6}$$

(b) Normalization condition:

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{7.7}$$

*Proof:* To prove (a), note from Theorem 5.1 (a) that $F(x)$ is a non-decreasing function and therefore,

$$f(x) = \frac{dF(x)}{dx} \geq 0.$$

Property (b) follows from (7.5) and Theorem 5.1 (b):

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{x\to\infty} F(x) = 1 \quad \square$$

*Remarks:*

- In the theory of probability, any function $f(x)$ satisfying properties (a) and (b) is called a probability density function.
- According to (7.7), the area under the graph of $f(x)$ from $x = -\infty$ to $\infty$ is equal to one.

*Theorem 7.3:* Let $X$ be a continuous RV with PDF $f(x)$. For any real numbers $a \leq b$, we have:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx \tag{7.8}$$

*Proof:* Since $P(X = a) = 0$, we have $P(a \leq X \leq b) = P(a < X \leq b)$. Now, using (5.19) and (7.5), we find

$$\begin{aligned} P(a < X \leq b) &= F(b) - F(a) \\ &= \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx = \int_a^b f(x)\,dx \quad \square \end{aligned}$$

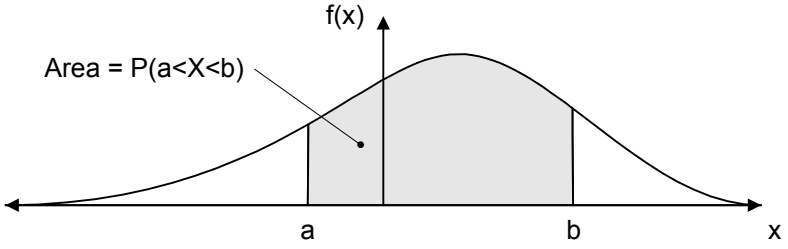
*Remarks:*

- Since $P(X = a) = P(X = b) = 0$, it should be clear that the formula (7.8) can also be used to compute

$$\begin{aligned} P(a < X \leq b) &= P(a \leq X \leq b) \\ &= P(a \leq X < b) = P(a < X < b) \end{aligned}$$

  That is, it does not matter whether or not the end-points $a$ and $b$ are taken into account.

- According to (7.8), the probability that $a \leq X \leq b$ is equal to the area under the graph of $f(x)$ over the interval $[a, b]$.



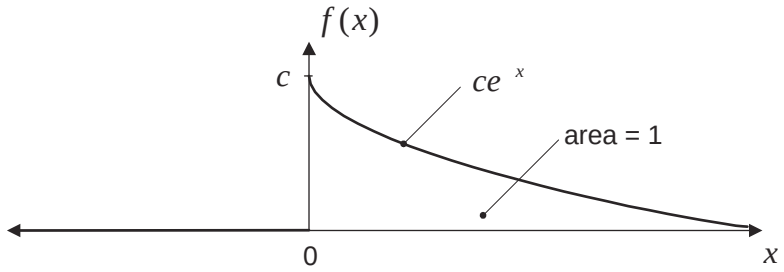

*Example 7.2:*

▶ *Example:* Let $X$ be a continuous RV with PDF

$$f(x) = \begin{cases} c\, e^{-x}, & 0 \leq x \\ 0, & x < 0. \end{cases} \tag{7.9}$$

(a) Determine the constant $c$ and sketch $f(x)$.
(b) Determine and sketch $F(x)$.
(c) Compute $P(-1 \leq X \leq 1)$.

*Solution:* The graph of $f(x)$ is illustrated below:




(a) To find the constant $c$, we simply require that the area under the graph of $f(x)$ be equal to 1:

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x)\,dx = c\int_{0}^{\infty} e^{-x}\,dx \\ &= c\left(-e^{-x}\right)\Big|_{0}^{\infty} = c \end{aligned}$$

(b) The CDF $F(x)$ is obtained by applying formula (7.5)

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

We must consider two cases:

$$\begin{aligned} x < 0 \;&\Rightarrow\; F(x) = \int_{-\infty}^{x} 0\,dt = 0 \\ x \geq 0 \;&\Rightarrow\; F(x) = \int_{0}^{x} e^{-t}\,dt = 1 - e^{-x} \end{aligned}$$

(c) Finally, we have

$$P(-1 \leq X \leq 1) = F(1) - F(-1^{-}) = 1 - e^{-1}$$

◀

*Interpretation of the PDF:*

- By itself, the PDF $f(x)$ is not a probability, and should not be identified with $P(X = x)$, which is 0 for all $x \in \mathbb{R}$.

- A proper interpretation of $f(x)$ can be developed as follows. For $\epsilon$ a small, positive number, we have

$$\begin{aligned} P(|X - x| < \epsilon) &= P(x - \epsilon < X < x + \epsilon) \\ &= \int_{x-\epsilon}^{x+\epsilon} f(t)\, dt \\ &\approx 2\epsilon f(x) \end{aligned} \tag{7.10}$$

  or equivalently,

$$f(x) \approx \frac{P(|X - x| < \epsilon)}{2\epsilon} \tag{7.11}$$

- Equation (7.11) explicitly shows $f(x)$ as a measure of the local probability "density" at the point $x$. In fact, (7.11) becomes an identity in the limit $\epsilon \to 0$.

- Alternatively, we may say that $f(x)$ is proportional to the likelihood that $X$ be found in a small neighborhood (of fixed size $2\epsilon$) around the point $x$.

- Thus, if $f(x_1) > f(x_2)$ for some $x_1 > x_2$, it is more likely that $X$ falls within $x_1 \pm \epsilon$ than within $x_2 \pm \epsilon$ (provided $\epsilon$ is small).

*Theorem 7.4:* For any real number subset $A \in \mathcal{B}_{\mathbb{R}}$, we have

$$P(X \in A) = \int_A f(x)\, dx \tag{7.12}$$

*Remarks:*

- Recall that $\mathcal{B}_{\mathbb{R}}$ is the Borel field of $\mathbb{R}$: i.e. the set of all subsets of $\mathbb{R}$ that can be constructed from intervals via a countable number of basic set operations, i.e. union, intersection and complement (see Section 3.4).
- From a practical viewpoint, $\mathcal{B}_{\mathbb{R}}$ contains all subsets $A \subseteq \mathbb{R}$ that may be of interest in engineering applications.
- The Theorem simply states that for any such subset $A$, the probability that $X \in A$ can be obtained as the area under the graph of $f(x)$ over the region $A$.
- The proof of the theorem is beyond the scope of this course.

## 7.2 PDF of a transformed RV

*Problem statement and overview:*

- Suppose $X$ is a continuous RV with known PDF $f(x)$
- Let $Y = h(X)$ where $h : \mathbb{R} \to \mathbb{R}$ is a real-valued function.
- Then, what is the PDF of the transformed RV $Y = h(X)$?
- In this section, we present two methods for evaluating the PDF $Y$:
  - Method of distribution (Section 7.2.1)
  - Method of transformation (Section 7.2.2)

### 7.2.1 Method of distributions

*Notations:*

- $X$ is continuous RV with known PDF $f(x)$.
- $Y = h(X)$, where $h : \mathbb{R} \to \mathbb{R}$.
- $g(y)$ and $G(y)$ denote the PDF and CDF of $Y$, respectively.

*Principle of the method:*

(1) For every $y \in \mathbb{R}$, find a real number subset $A_y$ such that

$$Y \leq y \Longleftrightarrow X \in A_y \tag{7.13}$$

(2) Find the CDF of $Y$ by integrating $f(x)$ over $A_y$:

$$\begin{aligned} G(y) &= P(Y \leq y) \\ &= P(X \in A_y) \\ &= \int_{A_y} f(x)\, dx \end{aligned} \tag{7.14}$$

(3) Find the PDF of $Y$ by differentiating $G(y)$:

$$g(y) = \frac{d}{dy} G(y) = G'(y) \tag{7.15}$$

*Remarks:*

- In many problems, we do not need to evaluate the integral in step (2) explicitly, because of the subsequent differentiation in step (3)
- In this respect, the following formula, known as Leibnitz's rule, is extremely useful in implementing step 3:

$$\frac{d}{dy}\int_{\alpha(y)}^{\beta(y)} f(x)\,dx = f(\beta(y))\beta'(y) - f(\alpha(y))\alpha'(y) \tag{7.16}$$

*Example 7.3:*

▶ Let $X$ be a continuous RV with PDF:

$$f(x) = \begin{cases} 1, & -\frac{1}{2} \le x \le \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{7.17}$$

Find the PDF of $Y = aX + b$ (assume $a > 0$).

*Solution:* We first note that

$$Y \le y \Longleftrightarrow aX + b \le y \Longleftrightarrow X \le \frac{y-b}{a}$$

Therefore, the CDF of $Y$ can be expressed in terms of the CDF of $X$ as follows:

$$\begin{aligned} G(y) &= P(Y \le y) \\ &= P(X \le \frac{y-b}{a}) = F(\frac{y-b}{a}) \end{aligned}$$

Finally, to obtain the PDF of $Y$, we simply take the derivative with respect to $y$:

$$\begin{aligned} g(y) &= G'(y) \\ &= \frac{d}{dy}F(\frac{y-b}{a}) \\ &= \frac{1}{a}F'(\frac{y-b}{a}) = \frac{1}{a}f(\frac{y-b}{a}) \end{aligned} \tag{7.18}$$

where the chain rule of derivation has been used on the last line. The relationship between $f(x)$ and $g(y)$ is illustrated below. Actually, results (7.18) is general and applies to any PDF $f(x)$ (i.e. not only the one in (7.17)). ◀

### 7.2.2 Method of transformations

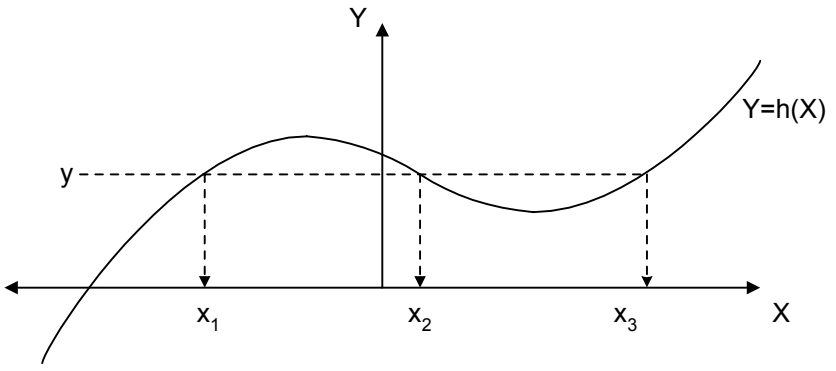
*Introduction:*

- This method may be viewed as a generalization of the operations involved in the method of distribution.
- This generalization takes the form of a theorem (or formula), for computing $g(y)$ directly from the knowledge of $f(x)$ and the transformation $h(.)$, without the need to explicitly compute $G(y)$.
- Thus, the method of transformation simply amounts to the direct application of the theorem below to compute $g(y)$.

*Theorem 7.5:* Let $X$ be a continuous RV with PDF $f(x)$. Let $Y = h(X)$ where $h$ is a differentiable real-valued function. For every $y \in \mathbb{R}$, let $x_i \equiv x_i(y)$ $(i = 1, 2, ...)$ denote the distinct real roots of the equation $y = h(x)$. Then

$$g(y) = \sum_i f(x_i) \left| \frac{dx_i}{dy} \right| \tag{7.19}$$

*Proof:* Consider the graph of $Y = h(X)$:



For an arbitrary value of $y$, let $x_1$, $x_2$,... denote the distinct real roots of the equation $y = h(x)$. From the graph of $Y = h(X)$, we note that

$$\begin{aligned} G(y) &= P(Y \leq y) \\ &= P(X \leq x_1) + P(x_2 \leq X \leq x_3) + ... \\ &= F(x_1) - F(x_2) + F(x_3) + ... \end{aligned} \tag{7.20}$$

We also note that the roots $x_i$ are functions of $y$, i.e. $x_i = x_i(y)$, with

$$\frac{dx_1}{dy} > 0, \quad \frac{dx_2}{dy} < 0, \quad \frac{dx_3}{dy} > 0, \quad ... \tag{7.21}$$

Taking the derivative of $G(y)$ and using this information, we have

$$\begin{aligned} g(y) &= \frac{dG(y)}{dy} \\ &= F'(x_1)\frac{dx_1}{dy} - F'(x_2)\frac{dx_2}{dy} + F'(x_3)\frac{dx_3}{dy} - ... \\ &= f(x_1)\left|\frac{dx_1}{dy}\right| + f(x_2)\left|\frac{dx_2}{dy}\right| + f(x_3)\left|\frac{dx_3}{dy}\right| + ... \end{aligned} \tag{7.22}$$

*Remarks:*

- To apply the theorem, we must first find the distinct real roots of the equation $y = h(x)$ as a function of $y$, denoted $x_i = x_i(y)$ $(i = 1, 2, \ldots)$. The number of such roots $x_i$ may depend on the specific value of $y$.
- Once the roots are known, we must compute the derivatives $\frac{dx_i}{dy}$ $(i = 1, 2, \ldots)$ and use them in (7.19) to evaluate the desired PDF $g(y)$.
- If the equation $y = h(x)$ has no real root for a given value of $y$, then formula (7.19) is interpreted as meaning $g(y) = 0$.

*Example 7.4:*

► Let $Y = X^2$ where $X$ is a continuous RV with PDF

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}$$

Find the PDF of $Y$ using the method of transformation.

*Solution:* In this type of problem, we find it useful to sketch the graph of the transformation $Y = h(X) = X^2$:

The number of roots of the equation $y = h(x) = x^2$ depends on the value of $y$:

- In the case $y < 0$, the equation $y = x^2$ has no real solution, and therefore

$$g(y) = 0, \quad y < 0$$

- In the case $y > 0$, the equation $y = x^2$ has two roots, namely:

$$x_1 = \sqrt{y} \;\Rightarrow\; \frac{dx_1}{dy} = \frac{1}{2\sqrt{y}}$$
$$x_2 = -\sqrt{y} \;\Rightarrow\; \frac{dx_2}{dy} = -\frac{1}{2\sqrt{y}}$$

At this point, a direct application of Theorem 7.4 gives

$$\begin{aligned} g(y) &= f(x_1)\left|\frac{dx_1}{dy}\right| + f(x_2)\left|\frac{dx_2}{dy}\right| \\ &= \frac{1}{\pi(1+x_1^2)} \cdot \frac{1}{2\sqrt{y}} + \frac{1}{\pi(1+x_2^2)} \cdot \frac{1}{2\sqrt{y}} \\ &= \frac{1}{\pi\sqrt{y}(1+y)}, \quad y > 0 \end{aligned}$$

◄

## 7.3 Expectation and variance

*Introduction:*

- Recall the definition of the expectation of a discrete RV $X$, with set of possible values $\{x_1, x_2, ...\}$ and PMF $p(x) = P(X = x)$:

$$E(X) \triangleq \sum_i x_i \, p(x_i) \tag{7.23}$$

- This definition cannot be applied to a continuous RV, because in this latter case:
  - the set of possible values is not discrete and
  - $P(X = x) = 0$ for all $x \in \mathbb{R}$.

- In this Section:
  - We present an alternate definition of expectation that is suitable for continuous RVs and study its properties.
  - We also extend the concept of variance to continuous RVs.

### 7.3.1 Definition of Expectation for continuous RVs

*Definition:* Let $X$ be a continuous RV with PDF $f(x)$. Provided the integral $\int_{-\infty}^{\infty} |x| f(x) dx$ is finite, the expected value of $X$ is defined as

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx. \tag{7.24}$$

*Remarks:*

- $E(X)$ is also called mean or expectation; it is often denoted by $\mu$ or $\mu_x$
- The condition $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$ is imposed for mathematical convenience. One can find continuous RVs for which this integral diverges; in this case, we say that $E(X)$ does not exist.
- Interpretations of $E(X)$:
  - Relative frequency: Consider $N$ independent observations of the RV X. $E(X)$ may be interpreted as the limiting value of the arithmetic average of these measurements when $N$ goes to infinity.
  - Circuit application: Let $X$ represent the measured voltage across a resistor in a DC circuit. Due to noise, interference and measurement error, this voltage is subject to small random fluctuations. Here, $E(X)$ may be interpreted as the true DC value of the voltage.

*Connection between the two definitions:*

- Let us investigate the connection between (7.24) and (7.23).
- For simplicity, assume that $f(x)$ has finite support $[a, b]$, i.e. $f(x) = 0$ if $x \notin [a, b]$.
- Then, from the definition of the Riemann integral, (7.24) gives

$$\begin{aligned} E(X) &= \int_a^b x\, f(x)\, dx \\ &= \lim_{N\to\infty} \sum_{i=1}^{N} x_i\, f(x_i)\, \Delta x \end{aligned} \tag{7.25}$$

  where $\Delta x = (b-a)/N$ and $x_i = a + (i - \frac{1}{2})\Delta x$.

- For $\Delta x$ small, $f(x_i)\Delta x \approx P(|X - x_i| \leq \Delta x/2)$ is the probability that $X$ lies in a small neighborhood of size $\Delta x$ centered at $x_i$. Thus, we have

$$E(X) \approx \sum_{i=1}^{N} x_i\, P(|X - x_i| \leq \Delta x/2) \tag{7.26}$$

  which is in agreement with (7.23).

*Example 7.5:*

▶ Let $X$ be a continuous RV with PDF

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases} \tag{7.27}$$

for some real numbers $a < b$. The expected value of $X$ is computed as follows:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\, f(x)\, dx = \frac{1}{b-a} \int_a^b x\, dx \\ &= \frac{1}{b-a} \left[ \frac{x^2}{2} \right]_a^b = \frac{(b^2 - a^2)}{2(b-a)} = \frac{a+b}{2} \end{aligned} \tag{7.28}$$

Thus, $E(X)$ is equal to the midpoint of the interval $[a, b]$. ◀

*Example 7.6:*

▶ Let $X$ be a continuous RV with PDF

$$f(x) = \begin{cases} \frac{2}{\pi} \frac{1}{1+x^2} & x \geq 0 \\ 0 & x < 0. \end{cases} \tag{7.29}$$

Here, we find

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\, f(x)\, dx \\ &= \frac{1}{\pi} \int_0^{\infty} \frac{2x}{1+x^2}\, dx \\ &= \frac{1}{\pi} \ln(1+x^2)|_0^{\infty} = \infty \end{aligned} \tag{7.30}$$

Thus, $E(X)$ does not exist. ◀

### 7.3.2 Properties of expectation

*Theorem 7.6:* Suppose that the PFD $f(x)$ is symmetric with respect to some real number $a$, i.e. $f(a-x) = f(a+x)$ for all $x \in \mathbb{R}$. Then

$$E(X) = a. \tag{7.31}$$

*Remarks:*

- The proof of the theorem amounts to a manipulation of the integral in (7.24). This is left as an exercise.
- An illustration of this theorem is given by Example 7.5, where $f(x)$ is symmetric with respect to the midpoint $(a+b)/2$.

*Theorem 7.7:* Let $h : \mathbb{R} \to \mathbb{R}$ be a real-valued function:

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) f(x)\, dx \tag{7.32}$$

*Remarks:*

- The proof of the theorem is more involved than in the discrete case. The student is referred to the textbook for additional details.
- Importance of the theorem:
  - Let $Y = h(X)$. From the definition of the expectation, we have $E(h(X)) = E(Y) = \int y\, g(y)\, dy$, where $g(y)$ denotes the PDF of $Y$.
  - According to Theorem 7.7, $g(y)$ is not explicitly needed to compute $E(h(X))$; knowledge of $f(x)$ is sufficient.

*Example 7.7:*

▶ The length $X$ of the side of a square is a RV with PDF

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{7.33}$$

Find the expected value of the square's area. ◀

*Corollary 7.7:* Let $h_1$, $h_2$,...,$h_n$ be real-valued functions and let $\alpha_1$, $\alpha_2$,...,$\alpha_n$ be real numbers. Then

$$E(\sum_{k=1}^{n} \alpha_k h_k(X)) = \sum_{k=1}^{n} \alpha_k E(h_k(X)) \tag{7.34}$$

*Remarks:*

- The proof is similar to that of Corollaries 6.4, with $\sum$ and $p(x)$ being replaced by $\int$ and $f(x)dx$, respectively.
- According to Corollary 7.7, the expectation $E(.)$ acts as a linear operation on its RV arguments $h_k(X)$.
- As a special case, we note that for any real numbers $\alpha$ and $\beta$,

$$E(\alpha X + \beta) = \alpha E(X) + \beta \tag{7.35}$$

*Example 7.8:*

▶ Let $X$ be a random angle with PDF

$$f(x) = \begin{cases} \frac{1}{2\pi}, & -\pi \leq x \leq \pi \\ 0, & \text{otherwise.} \end{cases} \tag{7.36}$$

... ◀

### 7.3.3 Variance of a continuous RV

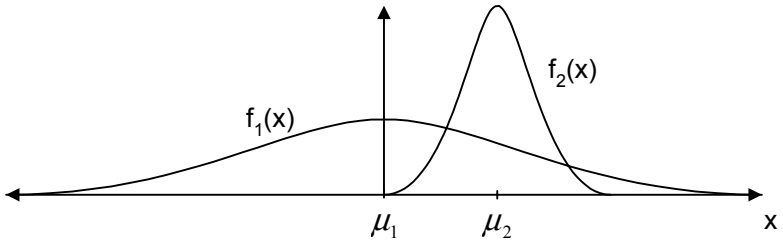
*Definition:* Let $X$ be a continuous RV with expectation $E(X) = \mu$ and PDF $f(x)$. The variance and standard deviation of $X$ are respectively defined as

$$Var(X) \triangleq E((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \tag{7.37}$$

$$\sigma_X \triangleq \sqrt{Var(X)} \tag{7.38}$$

*Remarks:*

- From (7.367), it should be clear that $Var(X) \geq 0$. In fact, if $X$ is a continuous RV, then $Var(X) > 0$ (always).
- $Var(X)$ (or equivalently $\sigma_X$), measures the spread of the PDF $f(x)$ about its mean $\mu$.
- This is illustrated in the figure below where $Var(X_1) > Var(X_2)$:




*Theorem 7.8:* The following relations hold, where $a$ and $b$ are arbitrary real numbers:

$$Var(X) = E(X^2) - \mu^2 \tag{7.39}$$

$$Var(aX+b) = a^2 Var(X) \tag{7.40}$$

*Remarks:*

- The proofs are identical to those of Theorem 6.5 and 6.7.
- In terms of the standard deviation, (7.40) is equivalent to

$$\sigma_{aX+b} = |a|\,\sigma_X \tag{7.41}$$

- Equation (7.39) is particularly useful when it is easier to compute $E(X^2)$ than $E((X-\mu)^2)$.

*Example 7.9:*

▶ Let $X$ be a continuous RV with PDF

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise.} \end{cases} \tag{7.42}$$

for some real numbers $a < b$. Find the variance of $X$

*Solution:* We have

$$\begin{aligned} Var(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \\ &= \frac{1}{b-a}\int_a^b (x-\mu)^2\,dx \\ &= \frac{1}{b-a}\left[\frac{(x-\mu)^3}{3}\right]_a^b \end{aligned} \tag{7.43}$$

Substituting $\mu = (a+b)/2$ in (7.43), we finally obtain

$$\begin{aligned} Var(X) &= \frac{1}{3(b-a)}\left[\frac{(b-a)^3}{8} - \frac{(a-b)^3}{8}\right] \\ &= \frac{(b-a)^2}{12} \end{aligned} \tag{7.44}$$

◀

## 7.4 The normal RV

In this and the next Section, we look at special continuous random variables of interest. These appear frequently in the application of probability theory. They may be used directly, or as basic building blocks to derive more advanced probabilistic models.

This Section is devoted to the study of the normal random variable, which is possibly the most important one. In the following Section, we shall study other RVs of interest, including uniform, exponential, Gamma and Rayleigh.

### 7.4.1 The standard normal

*Definition:* A continuous RV $X$ is called standard normal if its PDF takes the special form

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \text{all } x \in \mathbb{R} \tag{7.45}$$

*Remarks:*

- The use of the special notation $\phi(x)$ (instead of $f(x)$) is motivated by the importance of the PDF in (7.45), also called standard normal PDF.
- The following properties of $\phi(x)$ (7.45) may be verified easily:
  - Symmetry about $x = 0$:

$$\phi(-x) = \phi(x) \tag{7.46}$$

  - Absolute maximum at $x = 0$:

$$\phi(x) \leq \phi(0) = \frac{1}{\sqrt{2\pi}} \tag{7.47}$$

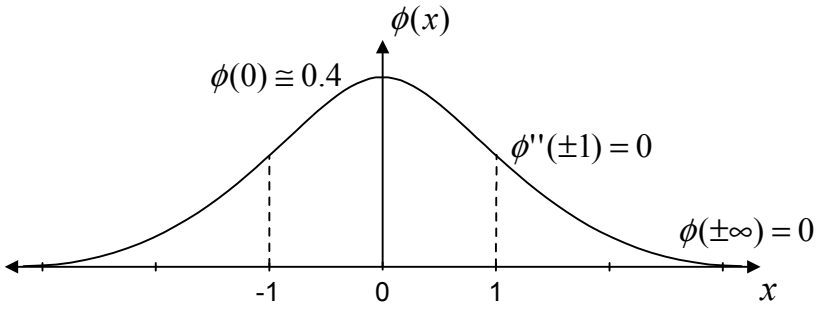
- Inflection points at $x = \pm 1$:

$$\phi''(-1) = \phi''(1) = 0 \tag{7.48}$$

- Asymptotic behavior:

$$\lim_{x \to \pm\infty} \phi(x) = 0 \tag{7.49}$$

- The graph of the standard normal PDF is illustrated below. It is characterized by a bell shape, consistent with basic properties (7.46)-(7.49).




- Finally, it can be verified that the area under the graph of $\phi(x)$ is equal to one, that is:

$$\int_{-\infty}^{\infty} \phi(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}dx = 1 \tag{7.50}$$

*Cumulative distribution function:* The CDF of the standard normal, denoted $\Phi(x)$, is given by

$$\Phi(x) = P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}dt \tag{7.51}$$

It is equal to the area under the graph of the standard normal PDF $\phi(t)$ (7.45) to the left of the point $t = x$.

*Remarks:*

- Again, the use of a special notation for the CDF (i.e. $\Phi(x)$ instead of $F(x)$) is motivated by the central role it plays in probability theory.

- Due to the symmetry of the standard normal PDF $\phi(t)$ about the origin $t = 0$, it follows that for any $x \in \mathbb{R}$,

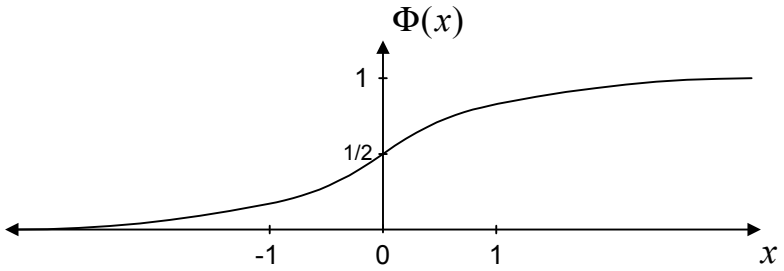
$$\begin{aligned} P(X \leq -x) &= P(X \geq x) \\ &= 1 - P(X < x) = 1 - P(X \leq x) \end{aligned}$$

  or equivalently,

$$\Phi(-x) = 1 - \Phi(x) \tag{7.52}$$

  Setting $x = 0$ in (7.52), we deduce that $\Phi(0) = 1/2$.

- The graph of $\Phi(x)$ is illustrated below:




- Unfortunately, no closed form expression exists for the CDF $\Phi(x)$ in (7.51); that is, the function $e^{-x^2/2}$ has no simple anti-derivative. Thus, the evaluation of $\Phi(x)$ requires the use of numerical integration or other kind of approximations.

- In practice, two simple approaches can be used for evaluating of $\Phi(x)$:
  - Use of a Table of values of $\Phi(x)$.
  - Use of scientific calculator and/or computer software.

*Use of tables:*

- Table of values of the function $\Phi(x)$ are available from many sources. Such a simplified Table is presented below:

| $x$ | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8810 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9772 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |

- These tables usually list values of $\Phi(x)$ for non-negative $x$ only. For example, from the above table, we read:

$$\Phi(0.75) = 0.7734$$

- Values of $\Phi(x)$ for $x < 0$ can be obtained from the relation (7.52). For example:

$$\Phi(-0.75) = 1 - \Phi(0.75) = 0.2266$$

*Use of calculator: error functions*

- Nowadays, many scientific calculators and computer softwares are available for the efficient computation of $\Phi(x)$, or closely related functions.
- In particular, we mention the error function:

$$\text{erf}(x) \triangleq \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{7.53}$$

  and the complementary error function:

$$\begin{aligned} \text{erfc}(x) &\triangleq \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt \\ &= 1 - \text{erf}(x) \end{aligned} \tag{7.54}$$

- The standard normal CDF $\Phi(x)$ can be expressed in terms of both types of error functions as follows:

$$\Phi(x) = 1 - \frac{1}{2}\text{erfc}(\frac{x}{\sqrt{2}}) \tag{7.55}$$

$$= \frac{1}{2} + \frac{1}{2}\text{erf}(\frac{x}{\sqrt{2}}) \tag{7.56}$$

*Probability calculations:*

- Let $X$ be a standard normal RV. For any practical real number subset, say $A \subseteq \mathbb{R}$, the probability $P(X \in A)$ can be expressed as a linear function of values of $\Phi(x)$.
- In particular, some frequently occurring events and their probability in terms of $\Phi(x)$ are listed below, where it is assumed that $x \geq 0$.

| Event | Probability |
|---|---|
| $X \leq x$ | $\Phi(x)$ |
| $X \leq -x$ | $1 - \Phi(x)$ |
| $\lvert X \rvert \leq x$ | $2\,\Phi(x) - 1$ |
| $\lvert X \rvert \geq x$ | $2\,(1 - \Phi(x))$ |

*Example 7.10:*

► ◄

*Theorem 7.9* Let $X$ be a standard normal RV. Then

$$E(X) = 0 \tag{7.57}$$

$$Var(X) = 1 \tag{7.58}$$

*Proof:* Since the standard normal PDF $\phi(x)$ (7.45) is symmetric about $x = 0$, it follows immediately from Theorem 7.56 that $E(X) = 0$. For the variance, note from (7.45) that

$$\phi'(x) = \frac{-x}{\sqrt{2\pi}} e^{-x^2/2} = -x\phi(x) \tag{7.59}$$

$$\phi''(x) = x^2\phi(x) - \phi(x) \tag{7.60}$$

Thus, using (6.74), we have

$$\begin{aligned} Var(X) &= \int_{-\infty}^{\infty} x^2\phi(x)\,dx \\ &= \int_{-\infty}^{\infty} \phi(x)\,dx + \int_{-\infty}^{\infty} \phi''(x)\,dx \\ &= 1 + \phi'(x)|_{-\infty}^{\infty} \\ &= 1 \end{aligned} \tag{7.61}$$

since $\phi'(\pm\infty) = 0$. □

### 7.4.2 The normal RV

*Definition:* A continuous RV $X$ is called normal (or Gaussian) with parameters $\mu$ and $\sigma$, or equivalently $X \sim N(\mu, \sigma^2)$, if its PDF is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R} \tag{7.62}$$

where $\mu$ and $\sigma$ are real numbers and $\sigma > 0$.

*Remarks:*

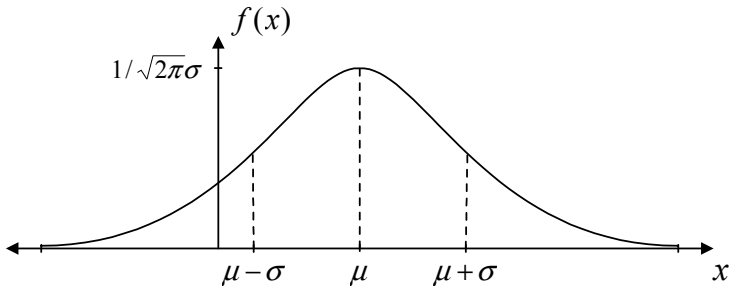
- We also say that $X$ is normally *distributed*.
- Note the relationship between the PDF $f(x)$ (7.62) and the standard normal PDF $\phi(x)$ (7.45):

$$f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$$

- In the special case $\mu = 0$ and $\sigma = 1$, $f(x) = \phi(x)$. Thus, the notation $N(0, 1)$ is synonymous of a standard normal distribution.
- The effects of the parameters $\mu$ and $\sigma$ in (7.62) are as follows:
  - $\mu$ is a translational parameter ($\mu > 0 \Rightarrow$ shift to the right)
  - $\sigma$ is a scaling parameter ($\sigma > 1 \Rightarrow$ dilation)
- The following properties of $f(x)$ (7.62) may be verified easily:
  - Symmetry about $x = \mu$: $f(\mu - x) = f(\mu + x)$
  - Absolute maximum at $x = \mu$ with $f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}$,
  - Inflection points at $\mu \pm \sigma$: $f''(\mu \pm \sigma) = 0$

- The graph of the normal PDF $f(x)$ (7.62) is shown below. It is characterized by a bell shape centered at $x = \mu$ with inflection points at $\mu \pm \sigma$.




- It can be verified that the PDF (7.62) is properly normalized. Using the change of variable $y = (x - \mu)/\sigma$, $dy = dx/\sigma$, we have:

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = 1 \tag{7.63}$$

  where the last equality follows from (7.50).

*Cumulative distribution function:* The CDF of a normal random variable $X \sim N(\mu, \sigma^2)$ is given by

$$F(x) \triangleq P(X \leq x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2} \, dt \tag{7.64}$$

*Remarks:*

- While no closed-form expression exists for the integral (7.64), $F(x)$ can be expressed in terms of the standard normal CDF $\Phi(x)$ in (7.51).

- Indeed, making the change of variable $y = (t - \mu)/\sigma$, $dy = dt/\sigma$, in (7.64), we obtain:

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-y^2/2} \, dy = \Phi(\frac{x - \mu}{\sigma}) \tag{7.65}$$

- Thus, any table or computer program available for the evaluation of $\Phi(x)$ may also be used to compute $F(x)$.

*Theorem 7.10* Let $X \sim N(\mu, \sigma^2)$. Then

$$E(X) = \mu, \qquad Var(X) = \sigma^2 \tag{7.66}$$

*Proof:* Since $f(x)$ (7.62) is symmetric about $x = \mu$, it follows from Theorem 7.6 that $E(X) = \mu$. For the variance, we have

$$\begin{aligned} Var(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2} dx \\ &= \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy \\ &= \sigma^2 \int_{-\infty}^{\infty} y^2 \phi(y) dy \\ &= \sigma^2 \end{aligned} \tag{7.67}$$

where the second equality follows from the change of variable $y = (x-\mu)/\sigma$, $dy = dx/\sigma$, and the last equality follows from (7.61). □

*Theorem 7.11:* Let $X \sim N(\mu, \sigma^2)$. Then the RV

$$Z = \frac{X - \mu}{\sigma} \tag{7.68}$$

is a standard normal RV, that is $Z \sim N(0, 1)$.

*Proof:* The proof is an application of the method of transformation in Section 7.2.2. Let $f(x)$ and $g(z)$ denote the PDFs of $X$ and $Z$, respectively, with $f(x)$ given by (7.62). The equation $z = (x-\mu)/\sigma$ has a single root, i.e. $x = \sigma z+\mu$, with derivative $dx/dz = \sigma$. Therefore, using Theorem 7.5, we obtain

$$\begin{aligned} g(z) &= f(x)\left|\frac{dx}{dz}\right| = \sigma f(\sigma z+\mu) \\ &= \sigma\frac{1}{\sqrt{2\pi}\sigma}e^{-[(\sigma z+\mu)-\mu]^2/2\sigma^2} = \frac{1}{\sqrt{2\pi}}e^{-z^2/2} = \phi(x) \end{aligned}$$

which shows that $Z$ is a standard normal RV. □

*Probability computations for the normal RV:*

- Let $X \sim N(\mu, \sigma^2)$ and suppose we want to compute $P(X \in A)$, for some $A \subseteq \mathbb{R}$.
- According to Theorem 7.10, RV $Z = (X-\mu)/\sigma$ is a standard normal.
- Thus, to compute $P(X \in A)$, we may proceed as follows:
  1. Find an equivalent subset $B \subseteq \mathbb{R}$ such that:
  $$X \in A \Longleftrightarrow Z \in B$$
  2. Compute $P(X \in A) = P(Z \in B)$ using techniques available for the standard normal (i.e. tables, calculator, etc.)

*Example 7.11:*

► Let $X \sim N(\mu, \sigma)$ with $\mu = 65$ and $\sigma = 15$. Find the probability that $X \geq 80$. ◄

*Special cases of interest:*

- Cumulative distribution function of $X$:

$$\begin{aligned} P(X \leq x) &= P(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) \\ &= P(Z \leq x') \\ &= \Phi(x') \end{aligned} \tag{7.69}$$

  where $x' \triangleq (x-\mu)/\sigma$ is a "standardized" value of $x$.

- Probability over interval:

$$\begin{aligned} P(a \leq X \leq b) &= P(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}) \\ &= P(a' \leq Z \leq b') \\ &= \Phi(b') - \Phi(a') \end{aligned} \tag{7.70}$$

  where $a' = (a-\mu)/\sigma$ and $b' = (b-\mu)/\sigma$.

*Example 7.12:*

▶ The value $X$ of a certain type of resistors is $N(\mu, \sigma^2)$ with $\mu = 10\Omega$ and $\sigma = 0.2\Omega$. If we buy 100 such resistors, what is the probability that $|X - 10\Omega| \leq 0.4\Omega$ for all resistors?

*Solution:* ◀

## 7.5 Other continuous RVs

We present below several continuous RVs of interest. Following a common trend, we shall often refer to these RV models as *distributions*.

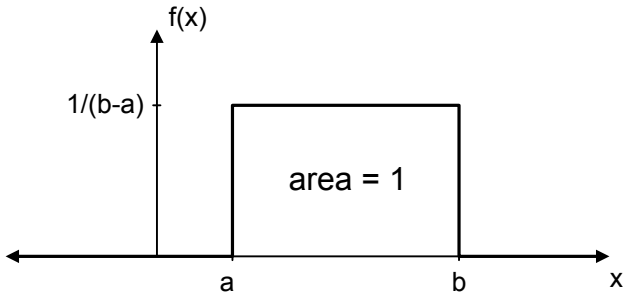
### 7.5.1 Uniform RV

*Definition:* A continuous RV $X$ is called uniform over the interval $(a, b)$, or equivalently $X \sim U(a, b)$, where $a < b$ are real numbers, if its PDF

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b, \\ 0, & \text{otherwise.} \end{cases} \tag{7.71}$$

*Remarks:*

- Graph of the PDF $f(x)$:




- The values of $f(x)$ at $x = a$ and $x = b$ are irrelevant.
- Consider sub-interval $A = (\alpha, \beta)$, where $a \leq \alpha < \beta \leq b$:

$$P(X \in A) = \int_{\alpha}^{\beta} f(x)\, dx = \frac{\beta - \alpha}{b - a} = \frac{\text{length of } (\alpha, \beta)}{\text{length of } (a, b)} \tag{7.72}$$

- The uniform RV is thus equivalent to earlier concept of random selection of a point from an interval (see Section 3.4.1)

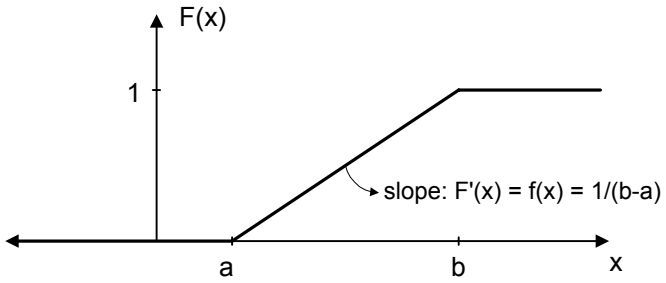
*Cumulative distribution function (CDF):* Recall the definition of the CDF:

$$F(x) \triangleq P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt \tag{7.73}$$

There are 3 cases to consider in the evaluation of $F(x)$: if $x < a$, then $f(t) = 0$ in (7.73) and $F(x) = 0$; if $a \leq x \leq b$, then $F(x) = \int_a^x \frac{1}{b-a}\, dt = \frac{x-a}{b-a}$; finally, if if $x > b$, then $F(x) = \int_a^b f(t)\, dt = 1$. We can summarize the results as

$$F(x) = \begin{cases} 0 & x \leq a, \\ \dfrac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases} \tag{7.74}$$

The graph of $F(x)$ is illustrated below:




*Theorem 7.12:* Let $X \sim U(a, b)$. Then

$$\mu = E(X) = \frac{a+b}{2} \tag{7.75}$$

$$\sigma^2 = Var(X) = \frac{(b-a)^2}{12} \tag{7.76}$$

*Proof:* Equation (7.75) is derived in Example 7.5 while (7.76) is derived in Example 7.9. □

### 7.5.2 The exponential RV

*Definition:* A continuous RV $X$ is called exponential with parameter $\lambda > 0$ if its PDF is given by

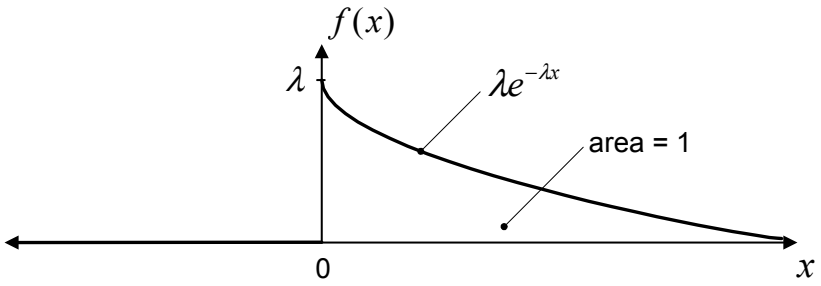
$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x < 0 \end{cases} \tag{7.77}$$

*Features of $f(x)$:*

- The value of $f(x)$ at $x = 0$ is irrelevant.
- $f(x)$ is properly normalized:

$$\int_0^\infty \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_0^\infty = (0 + 1) = 1 \tag{7.78}$$

- Graph of $f(x)$:



*Cumulative distribution function:* We need to evaluate the integral

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt \tag{7.79}$$

When $x \leq 0$, it follows from (7.77) that $F(x) = 0$; When $x > 0$, we have $F(x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}$. Therefore:

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{7.80}$$

The graph of $F(x)$ is shown below

*Theorem 7.13:* Let $X$ be exponential with parameter $\lambda > 0$:

$$E(X) = \frac{1}{\lambda}, \qquad Var(X) = \frac{1}{\lambda^2} \tag{7.81}$$

*Proof:* Left as an exercise to the student (use integration by parts). □

### 7.5.3 Laplacian RV

*Definition:* We say that continuous RV $X$ is Laplacian with parameters $\alpha > 0$ if its PDF takes the form

$$f(x) = \frac{\alpha}{2} e^{-\alpha|x|}, \quad x \in \mathbb{R} \tag{7.82}$$

*Remarks:* The Laplacian PDF (7.82) finds application in speech signal processing where it is used to model the random distribution of speech signal amplitudes.

*Theorem 7.14:* Let $X$ be Laplacian with parameter $\alpha$:

$$E(X) = 0, \quad Var(X) = \frac{2}{\alpha^2} \tag{7.83}$$

### 7.5.4 Rayleigh RV

*Definition:* We say that continuous RV $X$ is Rayleigh with parameter $\beta > 0$ if its PDF takes the form

$$f(x) = \begin{cases} \frac{x}{\beta^2} e^{-x^2/2\beta^2}, & x > 0 \\ 0, & x < 0 \end{cases} \tag{7.84}$$

*Remarks:* The Rayleigh distribution is used to model the statistics of signals transmitted through radio channels, as in e.g. mobile radio applications.

*Theorem 7.15:* Let $X$ be Rayleigh with parameter $\sigma$:

$$E(X) = \beta\sqrt{\pi/2}, \quad Var(X) = (2 - \pi/2)\beta^2 \tag{7.85}$$

### 7.5.5 Gamma RV

*Definition:* We say that continuous RV $X$ is Gamma with parameters $\beta > 0$ and $\lambda > 0$ if its PDF takes the form

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\beta-1}}{\Gamma(\beta)}, & x > 0 \\ 0, & x < 0 \end{cases} \tag{7.86}$$

where

$$\Gamma(\beta) \triangleq \int_0^\infty x^{\beta-1} e^{-x} dx \tag{7.87}$$

is the so-called Gamma function.

*Remarks:*

- This is an example of a family of PDFs characterized by two parameters.
- Gamma RVs find applications in e.g. queueing theory and reliability analysis.
- Note that in the special case $\beta = 1$, the Gamma PDF (7.86) reduces to the exponential PDF (7.77).

*Theorem 7.16:* Let $X$ be Gamma with parameter $\beta$ and $\lambda$:

$$E(X) = \frac{\beta}{\lambda}, \quad Var(X) = \frac{\beta}{\lambda^2} \tag{7.88}$$

## Problems

1. Provide a proof of Theorem 7.5.

2. Proove equation (7.50). (Hint: define $I = \int_{-\infty}^{\infty} \phi(x)dx$ and evaluate $I^2$, using polar coordinates.)

# Chapter 8

# Mixed RVs and moments

*Introduction*

- In the previous chapters, we defined and investigated the properties of discrete and continuous RVs.
- In applications, we often encounter RVs that are neither discrete nor continuous. Such RVs are generally called mixed.
- In this chapter:
  - We define the concept of a mixed RV.
  - We introduce a unifying notations suitable for all kinds of RVs.
  - Within this unifying framework, we study the concepts of moments and moment generating functions .

## 8.1 Mixed RVs

*Definition:* We say that RV $X$ is of a mixed type if its CDF can be expressed in the form

$$F(x) = \alpha F_d(x) + \beta F_c(x) \tag{8.1}$$

where $F_d(x)$ is a discrete CDF, $F_c(x)$ is a continuous CDF and $\alpha$ and $\beta$ are non-negative real numbers such that $\alpha + \beta = 1$.

*Remarks:*

- In (8.1), $F_d(x)$ must be a step function (see Section 6.1) and $F_d(x)$ must be absolutely continuous (see Section 7.1)
- It is not difficult to verify that $F(x)$ defined as above is a valid CDF, i.e. it satisfies properties (a)-(d) in Theorem 5.1.
- Discrete and continuous RVs are included as special cases of mixed RVs with the choice $\alpha = 1$ and $\alpha = 0$, respectively.

*Example 8.1:*

► Let RV $X$ denote the waiting time of a student at a registration desk. Assume that $X = 0$ if a clerk is available, and $X$ is exponential with parameter $\lambda$ if all the clerks are busy. Let $p$ denote the probability of a clerk being available. Find the CDF of $X$ and show that $X$ is a mixed RV.

*Solution:* We have

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= P(X \leq x|C)P(C) + P(X \leq x|C^c)P(C^c) \end{aligned}$$

where $C$ denotes the event that a clerk is available. Recall the definition of the unit step function:

$$u(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{8.2}$$

Given a clerk is available, the waiting time is 0 and we have:

$$P(X \leq x|C) = u(x)$$

Given a clerk is not available, $X$ is exponential and we have:

$$P(X \leq x|C^c) = (1 - e^{-\lambda x})u(x)$$

Finally, we obtain

$$F(x) = pu(x) + (1-p)(1 - e^{-\lambda x})u(x)$$

We note that $X$ is a mixed RV: its CDF can be expressed in the form (8.1) with

$$\begin{aligned} \alpha = p \qquad & F_d(x) = u(x) \\ \beta = 1 - p \qquad & F_c(x) = (1 - e^{-\lambda x})u(x) \end{aligned}$$

◄

*Definition:* Let $X$ be a mixed random variable with CDF $F(x)$. The PDF of $X$ is defined as

$$f(x) = F'(x) = \frac{dF(x)}{dx} \tag{8.3}$$

*Remarks:*

- Clearly, one should exercise care in the use of above definition since for mixed RVs, the CDF $F(x)$ will exhibit discontinuities in its graph.
- At the points of discontinuity of $F(x)$, the derivative in (7.6) does not exist in the conventional sense and the the PDF $f(x)$ (8.3) will contain singularities.
- However, (8.3) remains a valid operation if we extend the class of permissible PDFs $f(x)$ to include generalized functions.

*Derivative of unit step function:*

- The derivative of the unit step function is a generalized function called the unit impulse, and denoted

$$\delta(x) = \frac{du(x)}{dx} \tag{8.4}$$

- It may be viewed as an infinitely narrow pulse with area of one:

$$\delta(x) = 0, \quad \text{for all } x \neq 0 \tag{8.5}$$

$$\int_{-\infty}^{\infty} \delta(x)dx = 1 \tag{8.6}$$

- We recall the sifting property of $\delta(x)$: For any function $g(x)$ which is continuous at $x = 0$, we have

$$\int_{-\infty}^{\infty} g(x)\delta(x)dx = g(0) \tag{8.7}$$

*Example 8.2:*

► Find the PDF of the waiting time in the previous example.

*Solution:* In Example 8.1, we found that

$$F(x) = pu(x) + (1-p)(1-e^{-\lambda x})u(x)$$

Taking the derivative on both sides, we find:

$$\begin{aligned} f(x) &= p\delta(x) + (1-p)[\lambda e^{-\lambda x}u(x) + (1-e^{-\lambda x})\delta(x)] \\ &= p\delta(x) + (1-p)\lambda e^{-\lambda x}u(x) \end{aligned}$$

This PDF is illustrated below:

We invite the reader to verify that $\int_{-\infty}^{\infty} f(x)dx = 1$. ◄

*General form of the PDF:*

- Consider a mixed RV with CDF $F(x) = \alpha F_d(x) + \beta F_c(x)$.
- According to (6.7), since $F_d(x)$ is a discrete CDF, there exit numbers $\{x_1, x_2, \ldots\}$ with corresponding probabilities $p(x_i)$ such that

$$F_d(x) = \sum_{\text{all } i} p(x_i)u(x - x_i) \tag{8.8}$$

The corresponding PDF is

$$f_d(x) = \frac{dF_d(x)}{dx} = \sum_{\text{all } i} p(x_i)\delta(x - x_i) \tag{8.9}$$

- Making use of (8.9), the general form of the PDF of a mixed RV is immediately obtained as:

$$f(x) = \alpha \sum_{\text{all } i} p(x_i)\delta(x - x_i) + \beta f_c(x) \tag{8.10}$$

where $f_c(x) = F_c'(x)$ is the continuous PDF associated to $F_c(x)$.

## 8.2 Unifying framework

*Introduction:*

- In Chapters 6 and 7, respectively, we separately studied and derived important relations for discrete and continuous RVs.
- Using the extended definition of the PDF in (8.3), it is possible to recast most of these relations into a single form applicable to discrete, continuous and mixed RVs. This is considered below.

*Properties of $f(x)$:*

- $f(x) \geq 0$; $\int_{-\infty}^{\infty} f(x)dx = 1$
- If $f(x)$ is known, the CDF may be recovered from

$$F(x) = \int_{-\infty}^{x^+} f(t)\, dt \tag{8.11}$$

  where the upper limit $x^+$ means that a singularity at the point $t = x$ is covered by the interval of integration.

- For any real number subset $A$, we have $P(X \in A) = \int_A f(x)dx$

*Definition:* Let $X$ be a RV mixed with PDF $f(x)$. Provided the integral $\int_{-\infty}^{\infty} |x| f(x) dx$ is finite, the expected value of $X$ is defined as

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx. \tag{8.12}$$

*Remarks:*

- In the special case of a discrete RV,

$$f(x) = \sum_i p(x_i)\delta(x - x_i) \tag{8.13}$$

and therefore

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \,(\sum_i p(x_i)\delta(x - x_i))\, dx \\ &= \sum_i p(x_i) \int_{-\infty}^{\infty} x\, \delta(x - x_i)\, dx = \sum_i x_i\, p(x_i) \end{aligned} \tag{8.14}$$

which is identical to (6.19).

*Example 8.3:*

▶ Find the expected value of waiting time $X$ in Example 8.1.

*Solution:* We previously found that

$$f(x) = p\delta(x) + (1 - p)\lambda e^{-\lambda x} u(x)$$

The expected value of $X$ is obtained as follows:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx \\ &= p \int_{-\infty}^{\infty} x\delta(x) dx + (1 - p) \int_{-\infty}^{\infty} x e^{-\lambda x} u(x) dx \\ &= p \times 0 + (1 - p) \int_0^{\infty} x e^{-\lambda x} dx \\ &= (1 - p)\frac{1}{\lambda} \end{aligned}$$

where we recognize the last integral has the expected value of an exponential RV with parameter $\lambda$, which is equal to $1/\lambda$. ◀

*Properties of expectation:*

- Properties of the expectation derived for discrete and continuous RVs remain valid for the extended definition of expectation given above.
- For example (see Th. 7.5), if the PFD $f(x)$ is symmetric about $a$, i.e. $f(a-x) = f(a+x)$ for all $x \in \mathbb{R}$, then $E(X) = a$
- Also (see Th. 7.6), if $Y = h(X)$, then

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) f(x)\, dx \tag{8.15}$$

- Algebraic properties of the expectation, as stated in Corollary 7.6, also remain valid.

*Variance:*

- The definition of the variance is unchanged, that is:

$$Var(X) \triangleq E((X-\mu)^2), \qquad \sigma_X \triangleq \sqrt{Var(X)} \tag{8.16}$$

- Properties of the variance derived perviously for discrete and continuous RVs are generally applicable to mixed RVs.

## 8.3 Moments of a RV

*Definition:* Let $X$ be a RV with $\mu = E(X)$. Let $n$ be a non-negative integer, $r$ be a non-negative real number and $c$ be an arbitrary real number. We define:

(a) $E(X^n) = n$th moment of $X$

(b) $E(|X|^r) = r$th absolute moment of $X$

(c) $E[(X-c)^n] = n$th moment of $X$ about $c$

(d) $E[|X-c|^r] = r$th absolute moment of $X$ about $c$

*Remarks:*

- When $c = \mu$, moments about $c$ are called central moments.
- Moments are useful in applications:
  - For $n = 1$, $E(X^n) = E(X) =$ mean
  - For $n = 2$ and $c = \mu$, $E[(X-c)^2] = Var(X) =$ variance (2nd central moment)
  - The 3rd central moment provides information about the asymmetry of $f(x)$, etc.
- Note: some of these moments may not exist ($= \infty$).

*Example 8.4:*

► Let $X \sim N(\mu, \sigma^2)$. Find the $n$th central moments of $X$, where $n$ is a positive integer. ◄

*Remarks:* In the study of RV, and specially when comparing different RVs, it is often desirable to remove the effects of the mean and the variance. This is achieved by properly normalizing the RVs of interest.

*Definition:* Let $X$ be a RV with mean $\mu = E(X)$ and variance $Var(X) = \sigma^2$. The random variable

$$Z = \frac{X - \mu}{\sigma} \tag{8.17}$$

is called the standardized $X$.

*Theorem 8.1:* The RV $Z$ defined above has zero mean and unit variance:

$$E(Z) = 0, \quad Var(Z) = 1 \tag{8.18}$$

*Proof:* Using the definition (8.13), we have

$$E(Z) = \frac{1}{\sigma}(E(X) - \mu) = 0$$

$$Var(Z) = Var(\frac{X}{\sigma} - \frac{\mu}{\sigma}) = \frac{1}{\sigma^2}Var(X) = 1$$

*Example 8.5:*

► Suppose that the mean and standard deviation of all grades in the probability course are 65 and 15, respectively, while the corresponding quantities for the digital circuit course are 80 and 10. Mary has obtained 75 in probability and 85 in digital circuit. In what course is she doing better?

*Solution:* Let $X_1$ with $\mu_1 = 65$ and $\sigma_1 = 15$ denote the grade of a probability student. Similarly, let $X_2$ with $\mu_2 = 80$ and $\sigma_2 = 10$ denote the grade of a digital circuit student.

To determine in which course Mary did best, we compare her *standardized grades* in both courses:

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1} = \frac{2}{3}$$

$$Z_2 = \frac{X_2 - \mu_2}{\sigma_2} = \frac{1}{2}$$

Since $Z_2 > Z_1$, we conclude that Mary did better in the probability course. ◄

## 8.4 Characteristic function

*Introduction:*

- The characteristic function provides an alternative characterization of the PDF of a random variable.
- It is useful in several different ways, and specially:
  - in the computation of higher order moments of a RV
  - in evaluating the PDF of certain combinations of multiple RVs (e.g. sum of independent RVs).
  - in proving fundamental limit theorems in probability.
- The characteristic functions is a special types of so-called generating functions that come in different flavors:
  - Probability generating function
  - Moment generating function
  - Characteristic function
- In these notes, we focus on the Characteristic function.

### 8.4.1 Definition and properties

*Definition:* Let $X$ be a random variable with PDF $f(x)$. The Characteristic function (CF) of $X$, denoted by $\psi(\omega)$, where $\omega \in \mathbb{R}$, is defined by

$$\psi(\omega) \triangleq E(e^{-j\omega X}) = \int_{-\infty}^{\infty} f(x)e^{-j\omega x}\, dx, \quad \omega \in \mathbb{R} \tag{8.19}$$

*Remarks:*

- As defined in (8.19), the CF indeed corresponds to the Fourier transform of the PDF $f(x)$.
- In standard probability textbooks, the CF is usually defined without the minus sign in the argument of the exponential function in (8.19). Conceptually, this difference is of no consequences. From a practical perspective, however, the use of the minus sign in (8.19) allows the direct application of various formulas available for the Fourier transform in the calculation of (8.19).
- The integral in (8.19) always converges, regardless of the value of $\omega$. Indeed:

$$\begin{aligned} |\psi(\omega)| &= |\int_{-\infty}^{\infty} f(x)e^{-j\omega x}\, dx| \\ &\leq \int_{-\infty}^{\infty} |f(x)e^{-j\omega x}|\, dx = \int_{-\infty}^{\infty} f(x)\, dx = 1 \end{aligned} \tag{8.20}$$

*Theorem 8.2* The PDF may be expressed in terms of its CF via

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi(\omega) e^{j\omega x} d\omega, \quad x \in \mathbb{R} \tag{8.21}$$

*Remarks:*

- This is merely a restatement of the well-known inverse Fourier transform relationship that you have studied in a Signals and Systems course.
- One important consequence of this result is that the CF $\psi(\omega)$ uniquely characterizes RV X. Indeed, if two RVs $X$ and $Y$ have the same CF, say $\psi_X(\omega) = \psi_Y(\omega)$, then their PDF are identical, that is $f_X(x) = f_Y(x)$ for all $x \in \mathbb{R}$.

*Theorem 8.3:* Let $\psi(\omega)$ denote the CF of RV $X$. We have

$$E(X^n) = j^n \, \psi^{(n)}(0) \tag{8.22}$$

where $\psi^{(n)}(0) \equiv \frac{d^n \psi(\omega)}{d\omega^n}\Big|_{\omega=0}$.

*Proof:* Taking the $n$th derivative of (8.19) with respect to $\omega$, we have

$$\begin{aligned} \psi^{(n)}(\omega) &= \frac{d^n}{d\omega^n} \int_{-\infty}^{\infty} f(x) \, e^{-j\omega x} \, dx \\ &= \int_{-\infty}^{\infty} f(x) \, [\frac{d^n}{d\omega^n} e^{-j\omega x}] \, dx \\ &= \int_{-\infty}^{\infty} (-jx)^n f(x) \, e^{-j\omega x} \, dx \end{aligned}$$

Evaluating at $\omega = 0$, we obtain

$$\psi^{(n)}(0) = (-j)^n \int_{-\infty}^{\infty} x^n f(x) \, dx = (-j)^n E(X^n)$$

from which (8.122) follows immediately. □

*Remarks:*

- The Theorem states that for any arbitrary integer $n$, the $n$th moment of $X$ may be obtained by first computing the $n$th derivative of $\psi(\omega)$ with redpect to $\omega$ and then evaluating the result at $\omega = 0$.
- In other words, if an analytical expression is available for $\psi(\omega)$, we can compute all the moments $E(X^n)$ using only derivatives (and some algebra). This is why we sometimes refer to the characteristic function as a generating function for the moments $E(X^n)$.
- We note the following important special cases of (8.22):

$$\begin{aligned} \mu &= E(X) = j\psi'(0) && (8.23) \\ \sigma^2 &= Var(X) \\ &= E(X^2) - \mu^2 = -\psi''(0) + [\psi'(0)]^2 && (8.24) \end{aligned}$$

*Corollary 8.3:* The McLaurin series of $\psi(\omega)$ is

$$\psi(\omega) = \sum_{n=0}^{\infty} \psi^{(n)}(0)\frac{\omega^n}{n!} = \sum_{n=0}^{\infty} E(X^n)\frac{(-j\omega)^n}{n!} \tag{8.25}$$

*Usefulness:*

- This result provides an alternative way of computing $E(X^n)$ from $\psi(\omega)$.
- Indeed, suppose we already know the power series expansion of $\psi(\omega)$:

$$\psi(\omega) = c_0 + c_1\omega + c_2\frac{\omega^2}{2} + c_3\frac{\omega^3}{3!} + .... \tag{8.26}$$

- Then, we can identify $E(X^n) = j^n c_n$.

### 8.4.2 Characteristic functions of continuous RVs

*Uniform RV:* Let $X \sim U(a,b)$ where $a < b$. The PDF of $X$ is given by

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b, \\ 0, & \text{otherwise.} \end{cases} \tag{8.27}$$

Substituting into (8.15), we obtain

$$\begin{aligned} \psi(\omega) &= \int_{-\infty}^{\infty} f(x) e^{-j\omega x}\, dx \\ &= \frac{1}{b-a} \int_a^b e^{-j\omega x}\, dx \\ &= \frac{e^{-j\omega b} - e^{-j\omega a}}{j\omega(a-b)} \end{aligned} \tag{8.28}$$

*Exponential RV:* Let $X$ be an exponential RV $X$ with parameter $\lambda > 0$. The PDF of $X$ is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x < 0 \end{cases} \tag{8.29}$$

Substituting this expression in (8.15), we obtain the CF of $X$ as follows:

$$\begin{aligned} \psi(\omega) &= \int_0^{\infty} \lambda e^{-\lambda x} e^{-j\omega x}\, dx \\ &= \lambda \int_0^{\infty} e^{-(\lambda + j\omega)x} dx \qquad (8.30) \\ &= \lambda \left[ \frac{e^{-(\lambda+j\omega)x}}{-(\lambda + j\omega)} \right]_0^{\infty} \\ &= \frac{-\lambda}{(\lambda + j\omega)} (e^{-\infty} - e^{0}) \qquad (8.31) \\ &= \frac{\lambda}{\lambda + j\omega} \qquad (8.32) \end{aligned}$$

*Normal RV* Let $X \sim N(\mu, \sigma^2)$, with its PDF given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R} \tag{8.33}$$

Substituting into (8.15), we have

$$\psi(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} e^{-j\omega x}\, dx, \tag{8.34}$$

Upon evaluation of the integral (left as an exercise for the student), we obtain

$$\psi(\omega) = \exp(-j\omega\mu - \frac{\sigma^2\omega^2}{2}) \tag{8.35}$$

### 8.4.3 Characteristic functions of discrete RVs

*Introduction:*

- In this Section, we derive and study the characteristic functions of some of the basic discrete RVs introduced in Chapter 6.
- Here, $X$ is a discrete RV with set of possible values $\mathcal{R}_X = \{x_1, x_2, \ldots\}$ and probability mass function $p(x_i)$.
- In this special case, the CF (8.15) reduces to

$$\psi(\omega) = E(e^{-j\omega X}) = \sum_{x \in \mathcal{R}_X} p(x) e^{-j\omega x} \tag{8.36}$$

*Binomial RV:*

- Let $X \sim B(n,p)$ with $0 \leq p \leq 1$ and $q = 1-p$.
- The PMF of $X$ is given by

$$p(x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, ..., n \tag{8.37}$$

  and $p(x) = 0$ otherwise.

- The CF is obtained as follows:

$$\begin{aligned} \psi(\omega) &= \sum_{x=0}^{\infty} p(x) e^{-j\omega x} \\ &= \sum_{x=0}^{\infty} \binom{n}{x} p^x q^{n-x} e^{-j\omega x} \\ &= \sum_{x=0}^{\infty} \binom{n}{x} (pe^{-j\omega})^x q^{n-x} \\ &= (pe^{-j\omega} + q)^n \end{aligned} \tag{8.38}$$

- Let us apply the moment Theorem 8.3:

$$\psi'(\omega) = -jpe^{-j\omega} n (pe^{-j\omega} + q)^{n-1}$$

$$E(X) = j\psi'(0) = pn(p+q)^{n-1} = np \quad \text{(OK)}$$

- Try to compute $Var(X)$ using (8.24).

*Geometric RV:* For the geometric RV, we have

$$p(x) = p\,q^{x-1}, \quad x = 1, 2, 3, ... \tag{8.39}$$

and $p(x) = 0$ otherwise. The CF is obtained as

$$\begin{aligned} \psi(\omega) &= \sum_{x=1}^{\infty} p(x) e^{-j\omega x} \\ &= \sum_{x=1}^{\infty} p\,q^{x-1} e^{-j\omega x} \\ &= pe^{-j\omega} \sum_{x=0}^{\infty} (qe^{-j\omega})^x \\ &= \frac{pe^{-j\omega}}{1 - qe^{-j\omega}} \end{aligned} \tag{8.40}$$

*Poisson RV:* Let $X$ be Poisson with parameter $\lambda$. Its PMF is given by

$$p(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, ... \tag{8.41}$$

and $p(x) = 0$ otherwise. We leave it as an exercise for the student to verify that

$$\psi(\omega) = \exp(\lambda(e^{-j\omega} - 1)) \tag{8.42}$$

# Chapter 9

# Bivariate distributions

*Introduction and motivation:*

Up to now, our study of RVs has been limited to considering only a single RV, or function thereof, at a time. In many applications of probability in science and engineering, we must deal with several RVs that are simultaneously defined over a common probablity space.

For example, we might want to compute the probability that two RVs, say $X$ and $Y$, respectively belong to real number subsets $A$ and $B$ at the same time, that is: $P(X \in A, Y \in B)$.

In this and subsequent Chapters, the previously developed theory of a single RV, i.e. Chapters 5 to 8, is extended to handle such situations. This leads to the notion of joint distributions.

In this and the next Chapter, we first study in detail the case of two RVs, also known as bivariate distribution. In a subsequent Chapter, we shall consider the general case of $n \geq 2$ RVs.

## 9.1 Bivariate distributions

*Definition:* Let $X$ and $Y$ be two RVs defined on the probability space $(S, \mathcal{F}, P)$. We say that the mapping

$$s \in S \rightarrow (X(s), Y(s)) \in \mathbb{R}^2 \tag{9.1}$$

defines a two-dimensional random variable (or vector).

*Joint events:*

- It can be shown that for any practical subset $D \subseteq \mathbb{R}^2$ of the real plane, the set of outcomes

$$\{(X, Y) \in D\} \triangleq \{s \in S : (X(s), Y(s)) \in D\} \tag{9.2}$$

  is a valid event.[1] We refer to (9.2) as a joint (or bi-variate) event.

- The situation of interest here is illustrated below:

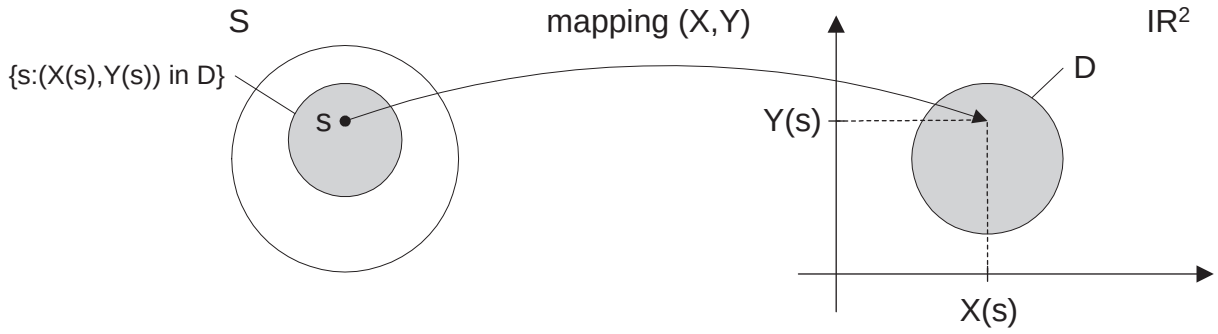



Figure 9.1: Illustration of a mapping $(X, Y)$ from $S$ into $\mathbb{R}^2$.

- Note that in the special case when $D = A \times B$, where $A \subseteq \mathbb{R}$ and $B \subseteq \mathbb{R}$, (9.2) reduces to

$$\{(X, Y) \in D\} = \{X \in A, Y \in B\}$$

[1] Specifically, if $D \in \mathcal{B}_{\mathbb{R}^2}$, the Borel field of $\mathbb{R}^2$ (see Chapter 3), then $\{(X, Y) \in D\} \in \mathcal{F}$.

*Joint probability:*

- Since $\{(X,Y) \in D\}$ is a valid event, the probability that $(X,Y) \in D$ is a well-defined quantity. This probability, denoted

$$P((X,Y) \in D) \equiv P(\{s \in S : (X(s), Y(s)) \in D\}),$$

  is called a joint probability.

- In this and the following sections, we develop tools to efficiently model and compute joint probabilities. Our first step is to introduce a bivariate CDF that generalizes the one defined in Chapter 5.

*Definition:* The joint cumulative distribution function (JCDF) of RVs $X$ and $Y$ is defined as

$$F(x,y) = P(X \leq x, Y \leq y), \quad \text{for all } (x,y) \in \mathbb{R}^2 \tag{9.3}$$

*Remarks:*

- Note that $F(x,y) = P((X,Y) \in C(x,y))$ where we define $C(x,y) = (-\infty, x] \times (-\infty, y]$. Region $C(x,y)$ is sometimes referred to as a *corner*:

- More generally, any practical subset $D \subseteq \mathbb{R}^2$ can be expressed as unions, intersections and/or complements of corners. From the axioms of probability, it follows that for any $D \subseteq \mathbb{R}^2$, the joint probability $P((X,Y) \in D)$ can be expressed in terms of $F(x,y)$.

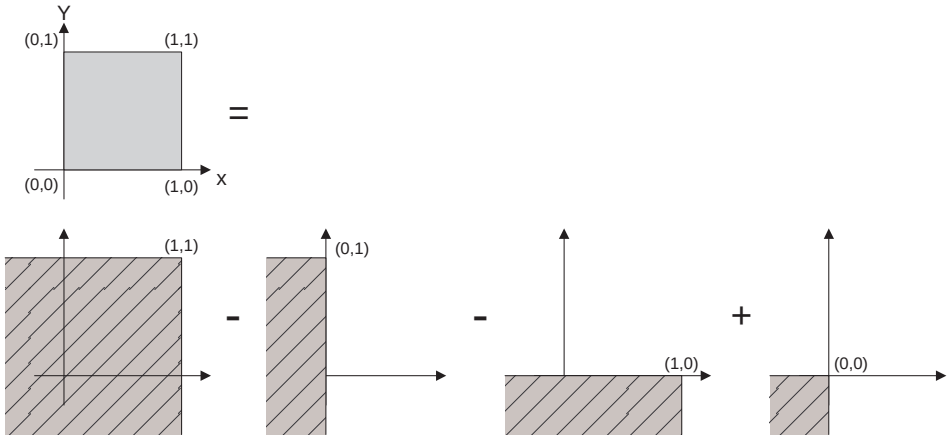
*Example 9.1:*

► Let $D = (0,1] \times (0,1] \subseteq \mathbb{R}^2$. Express $P((X,Y) \in D)$ in terms of the joint CDF of $X$ and $Y$, i.e. $F(x,y)$.

*Solution:* We have

$$\begin{aligned} P(0 < X \leq 1, 0 < Y \leq 1) &= P(0 < X \leq 1, Y \leq 1) - P(0 < X \leq 1, Y \leq 0) \\ &= P(X \leq 1, Y \leq 1) - P(X \leq 0, Y \leq 1) - \\ &\quad [P(X \leq 1, Y \leq 0) - P(X \leq 0, Y \leq 0)] \\ &= F(1,1) - F(0,1) - F(1,0) + F(0,0) \end{aligned}$$

A graphical interpretation of this result is provided in the figure below. ◄




*Theorem 9.1:*

(a) $F(x,y)$ is a non-decreasing function of its arguments $x$ and $y$.

(b) $F(-\infty, y) = F(x, -\infty) = F(-\infty, -\infty) = 0$

(c) $F(x, \infty) = P(X \leq x) = F_X(x)$ (CDF of $X$)
$F(\infty, y) = P(Y \leq y) = F_Y(y)$ (CDF of $Y$)
$F(\infty, \infty) = 1$

(d) $F(x^+, y) = F(x, y^+) = F(x, y)$

(e) If $x_1 < x_2$ and $y_1 < y_2$, then

$$F(x_2, x_2) - F(x_2, y_1) + F(x_1, y_1) - F(x_1, y_2) \geq 0 \tag{9.4}$$

*Remarks:*

- Proof similar to that of Theorem 5.1.
- According to (a), if $y$ is fixed and $x_1 < x_2$, then $F(x_2, y) \geq F(x_1, y)$. Similarly, if $x$ is fixed and $y_1 < y_2$, then $F(x, y_2) \geq F(x, y_1)$.
- In (b) and (c), interpret $\pm\infty$ as a limiting value, e.g.:

$$F(-\infty, y) = \lim_{x \to -\infty} F(x, y).$$

- In (c), $F_X(x)$ and $F_Y(y)$ are the CDF of $X$ and $Y$, respectively, as defined in Chapter 6. Here, they are often called *marginal* CDF.
- In (d), $x^+$ means the limit from the right, i.e.

$$F(x^+, y) = \lim_{t \to x+} F(t, y).$$

  with a similar interpretation for $y^+$.
- Any function $F(x, y)$ satisfying the above properties is called a JCDF.

*What's next?*

- In theory, if the JCDF $F(x, y)$ is known for all $(x, y) \in \mathbb{R}^2$, we can compute any joint probability for $X$ and $Y$. In practice, we find that the JCDF $F(x, y)$ is a bit difficult to handle and visualize.
- For these reasons, we usually work with equivalent but simpler representations of the joint CDF:
  - $X$ and $Y$ discrete $\Rightarrow$ joint PMF (Section 9.2)
  - $X$ and $Y$ continuous $\Rightarrow$ joint PDF (Section 9.3)

## 9.2 Joint probability mass function

*Definition:* Let $X$ and $Y$ be discrete random variables with sets of possible values $\mathcal{R}_X = \{x_1, x_2, ...\}$ and $\mathcal{R}_Y = \{y_1, y_2, ...\}$, respectively. We say that $X$ and $Y$ are *jointly discrete* and we define their joint probability mass function (JPMF) as

$$p(x, y) = P(X = x, Y = y), \quad \text{for all } (x, y) \in \mathbb{R}^2 \tag{9.5}$$

*Theorem 9.2:* The JPMF $p(x, y)$ satisfies the following basic properties:

(a) $0 \leq p(x, y) \leq 1$

(b) $x \notin \mathcal{R}_X$ or $y \notin \mathcal{R}_Y \Rightarrow p(x, y) = 0$

(c) Normalization property:

$$\sum_{x \in \mathcal{R}_X} \sum_{y \in \mathcal{R}_Y} p(x, y) = 1 \tag{9.6}$$

(d) Marginalization:

$$\sum_{y \in \mathcal{R}_Y} p(x, y) = P(X = x) \triangleq p_X(x) \tag{9.7}$$

$$\sum_{x \in \mathcal{R}_X} p(x, y) = P(Y = y) \triangleq p_Y(y) \tag{9.8}$$

*Proof:*

- Results (a) and (b) follow trivially from the definition of the JDPF.
- For (c), observe that the events $\{X = x, Y = y\}$, where $x \in \mathcal{R}_X$ and $y \in \mathcal{R}_Y$, form a partition of the sample space $S$. That is, they are mutually exclusive and

$$\bigcup_{x\in\mathcal{R}_X} \bigcup_{y\in\mathcal{R}_y} \{X = x, Y = y\} = S \tag{9.9}$$

  Using probability Axiom 3, we have

$$\begin{aligned}\sum_{x\in\mathcal{R}_X} \sum_{y\in\mathcal{R}_Y} p(x,y) &= \sum_{x\in\mathcal{R}_X} \sum_{y\in\mathcal{R}_Y} P(X = x, Y = y) \\ &= P\Big(\bigcup_{x\in\mathcal{R}_X} \bigcup_{y\in\mathcal{R}_Y} \{X = x, Y = y\}\Big) = P(S) = 1\end{aligned}$$

- For (d), note that for a given $x$, the events $\{X = x, Y = y\}$ where $y \in \mathcal{R}_Y$, form a partition of $\{X = x\}$. That is, they are mutually exclusive and

$$\bigcup_{y\in\mathcal{R}_Y} \{X = x, Y = y\} = \{X = x\} \tag{9.10}$$

  Again, using probability Axiom 3, we have

$$\begin{aligned}\sum_{y\in\mathcal{R}_Y} p(x,y) &= \sum_{y\in\mathcal{R}_Y} P(X = x, Y = y) \\ &= P\Big(\bigcup_{y\in\mathcal{R}_Y} \{X = x, Y = y\}\Big) = P(X = x) = p_X(x)\end{aligned}$$

  A similar argument holds for $p_Y(y)$. □

*Remarks:*

- $p_X(x)$ and $p_Y(y)$ are the probability mass functions (PMF) of $X$ and $Y$, respectively, as defined in Chapter 6.
- In the present context, they are also called *marginal* DPFs.

*Example 9.2:*

▶ Let $X$ and $Y$ denote the numbers showing up when rolling a magnetically coupled pair of dice. For $(i,j) \in \{1,\ldots,6\}^2$, let the JPMF be given by

$$p(i,j) = \begin{cases} (1+\epsilon)/36 & i = j \\ \alpha/36 & i \neq j. \end{cases}$$

where $0 < \epsilon \ll 1$.

(a) Find the constant $\alpha$.
(b) Find the marginal PMF $p_X(i)$
(c) Find the marginal PMF $p_Y(j)$
(d) Find $P(X = Y)$.

◀

## 9.3 Joint probability density function (JPDF)

*Definition:* We say that RVs $X$ and $Y$ are jointly continuous if there exists an integrable function $f : \mathbb{R}^2 \to [0, \infty)$, such that for any subset $D$ of $\mathbb{R}^2$, we have:

$$P((X, Y) \in D) = \iint_D f(x, y)\, dx\, dy \tag{9.11}$$

The function $f(x, y)$ is called the joint probability density function (JPDF) of $X$ and $Y$.

*Interpretations of $f(x, y)$:*

- Let $\Delta x$ and $\Delta y$ be sufficiently small positive numbers, then:

$$P(|X - x| < \frac{\Delta x}{2}, |Y - y| < \frac{\Delta y}{2}) \approx f(x, y)\, \Delta x\, \Delta y \tag{9.12}$$

- $P((X, Y) \in D)$ is equal to the volume under the graph of $f(x, y)$ over the region $D$ (see figure):

*Particular cases of interest:*

- For any subsets $A$ and $B$ of $\mathbb{R}$:

$$P(X \in A, Y \in B) = \int_A dx \int_B dy\, f(x,y) \tag{9.13}$$

- Let $A = [a,b]$ and $B = [c,d]$:

$$P(a \le X \le b, c \le Y \le d) = \int_a^b dx \int_c^d dy\, f(x,y) \tag{9.14}$$

  Note that the endpoints of the intervals $A$ and $B$ may be removed without affecting the value of the integral. Accordingly,

$$\begin{aligned} P(a \le X \le b, c \le Y \le d) &= P(a \le X < b, c \le Y \le d) \\ &= P(a \le X < b, c \le Y < d) \\ &= \text{etc.} \end{aligned}$$

- Let $C$ be any curve in the plane $\mathbb{R}^2$:

$$P((X,Y) \in C) = \iint_C f(x,y)\, dx\, dy = 0 \tag{9.15}$$

- For any $(a,b) \in \mathbb{R}^2$:

$$P(X = a, Y = b) = 0 \tag{9.16}$$

*Theorem 9.3:* The JPDF $f(x,y)$ satisfies the following properties:

(a) $f(x,y) \geq 0$

(b) Normalization:

$$\iint_{\mathbb{R}^2} f(x,y)\,dx\,dy = 1 \tag{9.17}$$

(c) Marginalization:

$$\int_{-\infty}^{\infty} f(x,y)\,dy = f_X(x) \triangleq \text{ PDF of } X \tag{9.18}$$

$$\int_{-\infty}^{\infty} f(x,y)\,dx = f_Y(y) \triangleq \text{ PDF of } Y \tag{9.19}$$

(d) Connection with JCDF:

$$F(x,y) = \int_{-\infty}^{x} dt \int_{-\infty}^{y} du\, f(t,u) \tag{9.20}$$

$$\frac{\partial^2 F(x,y)}{\partial x\,\partial y} = f(x,y) \tag{9.21}$$

*Note:*

- In the present context, $f_X(x)$ and $f_Y(y)$ are also called marginal PDF of $X$ and $Y$, respectively.

*Proof:*

(a) Follows from the definition of $f(x, y)$.

(b) Using (9.11) with $D = \mathbb{R}^2$, we have

$$\iint_{\mathbb{R}^2} f(x, y)\, dx\, dy = P((X, Y) \in \mathbb{R}^2) = 1$$

(c) From the definition of $f_X(x)$ as the PDF of $X$, we have that

$$P(X \in A) = \int_A f_X(x)\, dx \tag{9.22}$$

From the definition (9.11) of $f(x, y)$, we also have

$$\begin{aligned} P(X \in A) &= P(X \in A, Y \in \mathbb{R}) \\ &= \int_A \left( \int_{-\infty}^{\infty} f(x, y)\, dy \right) dx \end{aligned} \tag{9.23}$$

Both (9.22) and (9.23) being true for any subset $A \subseteq \mathbb{R}$, it follows that

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

(d) From (9.11), we immediately obtain

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} dt \int_{-\infty}^{y} du\, f(t, u)$$

At any continuity point of $f(x, y)$, we have:

$$\begin{aligned} \frac{\partial^2 F(x, y)}{\partial x\, \partial y} &= \frac{\partial}{\partial y} \left( \frac{\partial}{\partial x} \int_{-\infty}^{x} dt \int_{-\infty}^{y} du\, f(t, u) \right) \\ &= \frac{\partial}{\partial y} \int_{-\infty}^{y} du\, f(x, u) \\ &= f(x, y) \qquad \square \end{aligned}$$

*Example 9.3:*

► *Example:* Let $X$ and $Y$ be jointly continuous with JPDF

$$f(x,y) = \begin{cases} cxy & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{9.24}$$

(a) Find the constant $c$.
(b) Find the probability that $Y \ge X$.
(c) Find the marginal PDFs of $X$ and $Y$, i.e. $f_X(x)$ and $f_Y(y)$.

*Solution:* (a) Constant $c$ is obtained from the normalization condition (9.17):

$$\begin{aligned} \int_{-\infty}^{\infty} f(x,y)dxdy &= c\int_0^1\int_0^1 xy\,dx\,dy = c\left(\int_0^1 x\,dx\right)^2 \\ &= c\left(\frac{x^2}{2}\Big|_0^1\right)^2 = \frac{c}{4} = 1 \quad \Rightarrow \quad c = 4 \end{aligned}$$

(b)We seek

$$\begin{aligned} P(Y \ge X) &= P((X,Y) \in D) = \iint_D f(x,y)\,dx\,dy \\ &= \int_0^1 \left(\int_x^1 4xy\,dy\right) dx = \int_0^1 \left(2xy^2\Big|_x^1\right) dx \\ &= 2\int_0^1 (x - x^3)\,dx = 2\left(\frac{x^2}{2} - \frac{x^4}{4}\right)\Big|_0^1 \\ &= 2\left(\frac{1}{2} - \frac{1}{4}\right) = \frac{1}{2} \end{aligned}$$

(c) ... ◄

### 9.3.1 Uniform distribution

*Definition:* $X$ and $Y$ are jointly uniform over region $D \subseteq \mathbb{R}^2$, or equivalently $(X, Y) \sim U(D)$, if their joint PDF takes the form

$$f(x,y) = \begin{cases} c & \text{for all } (x,y) \in D \\ 0 & \text{otherwise} \end{cases} \tag{9.25}$$

where $c$ is a constant.

*Remarks:*

- The value of the constant $c$ is obtained from the requirement that $f(x,y)$ be properly normalized, that is:

$$\iint_D f(x,y)\,dx\,dy = 1 \quad \Rightarrow \quad c = \frac{1}{\iint_D dx\,dy} = \frac{1}{\text{Area}(D)} \tag{9.26}$$

- If $(X, Y) \sim U(D)$, then for any subset $E \subseteq \mathbb{R}^2$:

$$\begin{aligned} P((X,Y) \in E) &= \iint_E f(x,y)\,dx\,dy \\ &= c \iint_{E \cap D} dx\,dy = \frac{\text{Area}(E \cap D)}{\text{Area}(D)} \end{aligned} \tag{9.27}$$

- The above concept is equivalent to the random selection of points from a region $D \subseteq \mathbb{R}^2$, as previously discussed in Chapter 3.

*Example 9.4:*

► Bill and Monica decide to meet for dinner between 20:00 and 20:30 in a restaurant lounge. Assuming that they arrive at random during this time, find the probability that the waiting time of any one of them be more than 15 minutes?

*Solution:* Let $X$ and $Y$ respectively denote the arrival time of Bill and Monica in minutes after 20:00. Assume $(X, Y) \sim U(D)$ where

$$D = \{(x, y) : 0 \leq x \leq 30 \text{ and } 0 \leq y \leq 30\}$$

The event that the waiting time of Bill or Monica is more than 15 minutes can be expressed as

$$E = \{(x, y) \in S : |x - y| \geq 15\}$$

This event is illustrated below:

The desired probability is

$$\begin{aligned} P((X, Y) \in E) &= \frac{\text{Area}(E)}{\text{Area}(D)} \\ &= \frac{2 \times (15 \times 15)/2}{30 \times 30} = \frac{1}{4} \end{aligned}$$

◄

### 9.3.2 Normal Distribution

*Definition:* RVs $X$ and $Y$ are jointly normal if their joint PDF can be expressed in the form

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp[-\frac{1}{2}Q(\frac{x-\mu_X}{\sigma_X}, \frac{y-\mu_Y}{\sigma_Y})], \quad (x,y) \in \mathbb{R}^2 \tag{9.28}$$

where

$$Q(u,v) = \frac{1}{1-\rho^2}(u^2 - 2\rho uv + v^2) \tag{9.29}$$

and the parameters $\mu_X$ and $\mu_Y \in \mathbb{R}$, $\sigma_X$ and $\sigma_Y > 0$ and $-1 < \rho < 1$.

*Remarks:*

- We also refer to (9.28) as the bivariate Gaussian distribution.
- Compact notation: $(X,Y) \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$
- It can be shown that $f(x,y)$ is properly normalized, that is:

$$\iint_{\mathbb{R}^2} f(x,y)\,dx\,dy = 1 \tag{9.30}$$

- The precise meaning of the parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ and $\rho$ will become clear as we progress in our study.
- The normal distribution (and its multi-dimensional extension) is one of the most important joint PDF in applications of probability.

*Shape of $f(x, y)$:*

- The function $Q(u, v)$ in (9.29) is a positive definite quadratic form:

$$Q(u, v) \geq 0, \qquad \text{for all } (u, v) \in \mathbb{R}^2$$

  with equality iff $(u, v) = (0, 0)$. Therefore, $f(x, y)$ (9.28) attains its absolute maximum at the point $(\mu_X, \mu_Y)$.

- In the limit $u \to \pm\infty$ and/or $v \to \pm\infty$, the function $Q(u, v) \to +\infty$. Accordingly, $f(x, y) \to 0$ in the limit $x \to \pm\infty$ and/or $y \to \pm\infty$.

- A study of the quadratic form (9.29) shows that its level contour curves, i.e. the locus defined by $Q(u, v) = c$ for positive constants $c$, are ellipses centered at $(0, 0)$, with the orientation of the principal axis depending on $\rho$.

- Accordingly, the graph of the function $f(x, y)$ has the form of a bell-shaped surface with elliptic cross-sections:
  - The bell is centered at the point $(\mu_X, \mu_Y)$ where $f(x, y)$ attains its maximum value.
  - The level contours of the function $f(x, y)$ have the form of ellipses whose exact shape depend on the parameter $\sigma_X$, $\sigma_Y$ and $\rho$.

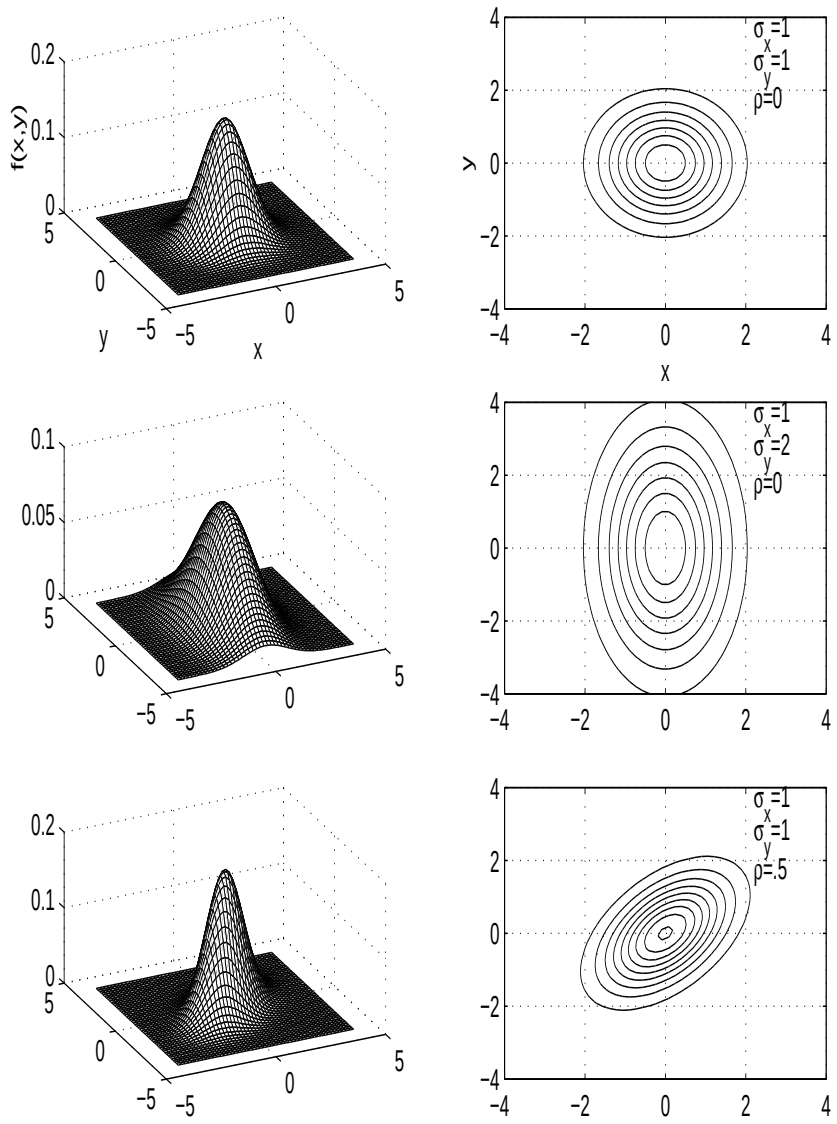



Figure 9.2: The bivariate normal PDF.

*Theorem 9.4:* Let $(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$. The marginal PDF of $X$ and $Y$ are given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-(x-\mu_X)^2/2\sigma_X^2}, \qquad f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-(y-\mu_Y)^2/2\sigma_Y^2} \tag{9.31}$$

That is, $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$

*Remarks:*

- The proof is left as an exercise for the student.
- According to Theorem 9.4, joint normality of RVs $X$ and $Y$ implies that each one of them is normal when considered individually.
- The converse to this statement is not generally true: $X$ and $Y$ being normal when taken in isolation does not imply that they are jointly normal in general.
- Theorem 9.4 provides a complete explanation for the meaning of the parameters $\mu_X$, $\sigma_X$, $\mu_Y$ and $\sigma_Y$:
  - $\mu_X = E(X) =$ expected value of $X$
  - $\sigma_X^2 = Var(X) =$ variance of $X$
  - with similar interpretations for $\mu_Y$ and $\sigma_Y$
- The exact signification of the parameter $\rho$, called correlation coefficient, will be explained in the next chapter.

## 9.4 Conditional distributions

It is often of interest to compute the conditional probability that RV $X$ belong to a real number subset $B \subseteq \mathbb{R}$, given that a certain event, say $A$, has occurred, that is:

$$P(X \in B \,|A)$$

In this Section, we develop the necessary theory to handle this kind of probability computations. Special emphasis is given to the case where the event $A$ is itself defined in terms of a second RV, say $Y$.

### 9.4.1 Arbitrary event $A$

*Definition:* Let $X$ be a RV (discrete or continuous) and let $A$ be some event with $P(A) > 0$. The conditional CDF of $X$ given $A$ is defined as

$$F(x \,|A) \triangleq P(X \leq x \,|A) = \frac{P(X \leq x, A)}{P(A)}, \quad \text{all } x \in \mathbb{R} \tag{9.32}$$

*Remarks:*

- The function $F(x \,|A)$ is a valid CDF, in the sense that it satisfies all the basic properties of a CDF (see Theorem 5.1).
- In theory, the function $F(x \,|A)$ can be used to compute any probability of the type $P(X \in B \,|A)$. In practice, it is preferable to work with closely related functions as defined below.

*Definition:* Let $X$ be a discrete RV with set of possible values $\mathcal{R}_X = \{x_1, x_2, ...\}$. The conditional PMF of $X$ given $A$ is defined as

$$p(x\,|A) \triangleq P(X = x\,|A) = \frac{P(X = x, A)}{P(A)}, \quad \text{all } x \in \mathbb{R} \tag{9.33}$$

*Remarks:*

- The function $p(x\,|A)$ is a valid PMF. In particular (see Theorem 6.1), we have: $p(x\,|A) \geq 0$, $p(x\,|A) = 0$ for all $x \notin \mathcal{R}_X$ and

$$\sum_{x \in \mathcal{R}_X} p(x\,|A) = \sum_{\text{all } i} p(x_i\,|A) = 1 \tag{9.34}$$

- Furthermore, for any subset $B \subseteq \mathbb{R}$, we have

$$P(X \in B\,|A) = \sum_{x_i \in B} p(x_i\,|A) \tag{9.35}$$

*Definition:* Let $X$ be a continuous RV. The conditional PDF of $X$ given $A$ is defined as

$$f(x\,|A) \triangleq \frac{dF(x\,|A)}{dx}, \quad x \in \mathbb{R} \tag{9.36}$$

*Remarks:*

- The function $f(x|A)$ is a valid PDF (see Theorem 7.2). In particular, we have $f(x\,|A) \geq 0$ for all $x \in \mathbb{R}$ and

$$\int_{-\infty}^{\infty} f(x\,|A)\,dx = 1 \tag{9.37}$$

- For any subset $B \subseteq \mathbb{R}$:

$$P(X \in B\,|A) = \int_B f(x\,|A)\,dx. \tag{9.38}$$

- As a special case of the above, we note:

$$F(x\,|A) = \int_{-\infty}^{x} f(t\,|A)\,dt \tag{9.39}$$

### 9.4.2 Event $A$ related to discrete RV $Y$

*Introduction:* Let $Y$ be a discrete RV with possible values $\mathcal{R}_Y = \{y_1, y_2, ...\}$ and marginal PMF $p_Y(y)$. The event $A$ under consideration here is

$$A = \{s \in S : Y(s) = y\} = \{Y = y\} \tag{9.40}$$

for some $y$ such that $p_Y(y) > 0$.

*Definition:* Let $y \in \mathbb{R}$ be such that $p_Y(y) > 0$. The conditional CDF of $X$ given $Y = y$ is defined as

$$F_{X|Y}(x\,|y) \triangleq P(X \leq x\,|Y = y) = \frac{P(X \leq x, Y = y)}{P(Y = y)}, \quad x \in \mathbb{R} \tag{9.41}$$

*Remarks:*

- In theory, knowledge of $F_{X|Y}(x\,|y)$ is sufficient to compute any conditional probability of the type $P(X \in B|Y = y)$, for any subset $B \in \mathbb{R}$.
- In practice, depending whether $X$ is discrete or cont., we find it more convenient to work with closely related functions, as explained next.

*Definition:* Suppose $X$ is a discrete RV with set of possible values $\mathcal{R}_X = \{x_1, x_2, ...\}$. The conditional PMF of $X$ given $Y = y$ is defined as

$$p_{X|Y}(x\,|y) \triangleq P(X = x\,|Y = y), \quad x \in \mathbb{R} \tag{9.42}$$

*Remarks:*

- Invoking the definition of conditional probability, we have

$$p_{X|Y}(x\,|y) = \frac{P(X = x, Y = y)}{P(y = Y)} = \frac{p(x, y)}{p_Y(y)} \tag{9.43}$$

  where $p(x, y)$ is the joint PMF of $X$ and $Y$, as defined in (9.5).

- Since $p_{X|Y}(x\,|y)$ in (9.42) is a special case of $p(x\,|A)$ (9.33), it is also a valid PMF. In particular, it satisfies properties similar to those in (9.34)-(9.35) with obvious modifications in the notation.

*Example 9.5:*

▶ Let $X$ and $Y$ be defined as in Example 9.2. Find the conditional PMF of $X$ given $Y = j$ where $j \in \{1, \ldots, 6\}$.

*Solution:* From Example 9.2, we recall that

$$p(i, j) = \begin{cases} (1 + \epsilon)/36, & i = j \\ \alpha/36, & i \neq j \end{cases}$$

where $\alpha = 1 - \frac{\epsilon}{5}$ and

$$p_Y(j) = 1/6, \text{ all } j$$

The desired conditional probability is obtained as

$$p_{X|Y}(i|j) = \frac{p(i, j)}{p_Y(j)} = \begin{cases} (1 + \epsilon)/6, & i = j \\ \alpha/6, & i \neq j \end{cases}$$

◀

*Definition:* Suppose $X$ is a continuous RV. The conditional PDF of $X$ given $Y = y$ is defined as

$$f_{X|Y}(x|y) \triangleq \frac{\partial F_{X|Y}(x|y)}{\partial x} \tag{9.44}$$

*Remarks:*

- $f_{X|Y}(x\,|y)$ is a special case of $f(x\,|A)$ (9.36) and as such, it is a valid PDF.
- It satisfies properties similar to those in (9.37)-(9.39) with obvious modifications in notation, e.g.:

$$P(X \in B\,|Y = y) = \int_B f_{X|Y}(x\,|y)\,dx. \tag{9.45}$$

*Example 9.6: Binary communications over noisy channel.*

▶ Let discrete RV $Y \in \{-1, +1\}$ denote the amplitude of a transmitted binary pulse at the input of a digital communication link. Assume that

$$p_Y(y) = P(Y = y) = 1/2, \quad y \in \{-1, 1\}$$

Let $X$ denote the received voltage at the output of the link. Under the so-called additive Gaussian noise assumption, we may assume that conditional on $Y = y$, RV $X$ is $N(y, \sigma^2)$. Given a positive pulse was transmitted, find the probability that the receiver makes an erroneous decision, that is find $P(X \leq 0\,|Y = 1)$. ◀

### 9.4.3 Event $A$ related to continuous RV $Y$

*Introduction:*

- Let $X$ and $Y$ be jointly continuous with PDF $f(x, y)$ and consider the event $A = \{Y = y\}$, for some $y$ such that $f_Y(y) > 0$.
- How can we characterize the conditional probability distribution of RV $X$, given that event $A$, i.e. $Y = y$, has been observed?
- Our previous definition of conditional CDF is not applicable here. Indeed

$$F_{X|Y}(x\,|y) = P(X \leq x\,|Y = y) = \frac{P(X \leq x, Y = y)}{P(Y = y)} = \frac{0}{0} \quad (?)$$

- To accommodate this situation, the following extended definition of conditional CDF, based on the concept of limit, is commonly used.

*Definition:* Let $X$ and $Y$ be jointly continuous. The conditional CDF of $X$, given $Y = y$, is defined as

$$F_{X|Y}(x\,|y) \triangleq \lim_{\epsilon \to 0^+} P(X \leq x\,|y - \epsilon < Y \leq y + \epsilon). \tag{9.46}$$

*Definition:* Let $X$ and $Y$ be jointly continuous. The conditional PDF of $X$, given $Y = y$, is defined as

$$f_{X|Y}(x\,|y) \triangleq \frac{\partial F_{X|Y}(x\,|y)}{\partial x}. \tag{9.47}$$

*Remarks:*

- The function $f_{X|Y}(x\,|y)$ is a valid PDF; it satisfies properties similar to (9.37)-(9.39) with obvious modifications. In particular:

$$f_{X|Y}(x\,|y) \geq 0, \quad x \in \mathbb{R} \tag{9.48}$$

$$\int_{-\infty}^{\infty} f_{X|Y}(x\,|y)\,dx = 1 \tag{9.49}$$

$$F_{X|Y}(x\,|y) = \int_{-\infty}^{x} f_{X|Y}(t\,|y)\,dt \tag{9.50}$$

- In practice, the conditional PDF $f_{X|Y}(x\,|y)$ is used instead of the conditional CDF in the computation of probabilities:

$$P(X \in B|Y = y) = \int_B f_{X|Y}(x\,|y)\,dx \tag{9.51}$$

*Theorem 9.5:* Provided $f_Y(y) > 0$, the conditional PDF of $X$ given $Y = y$ can be expressed as

$$f_{X|Y}(x\,|y) = \frac{f(x,y)}{f_Y(y)} \tag{9.52}$$

*Proof (optional material):*

$$\begin{aligned}
F_{X|Y}(x\,|y) &= \lim_{\epsilon\to 0^+} P(X \le x\,|y-\epsilon < Y \le y+\epsilon) \\
&= \lim_{\epsilon\to 0^+} \frac{P(X \le x, y-\epsilon < Y \le y+\epsilon)}{P(y-\epsilon < Y \le y+\epsilon)} \\
&= \lim_{\epsilon\to 0^+} \frac{\int_{-\infty}^{x} dt \int_{y-\epsilon}^{y+\epsilon} du f(t,u)}{\int_{y-\epsilon}^{y+\epsilon} du f_Y(u)} \\
&= \int_{-\infty}^{x} dt \lim_{\epsilon\to 0^+} \left[\frac{\int_{y-\epsilon}^{y+\epsilon} du f(t,u)}{\int_{y-\epsilon}^{y+\epsilon} du f_Y(u)}\right] \\
&= \int_{-\infty}^{x} dt \lim_{\epsilon\to 0^+} \left[\frac{2\epsilon f(t,y)}{2\epsilon f_Y(y)} + O(\epsilon)\right] \\
&= \int_{-\infty}^{x} dt \frac{f(t,y)}{f_Y(y)} \qquad (9.53)
\end{aligned}$$

Taking the partial derivative with respect to $x$, we finally obtain

$$f_{X|Y}(x\,|y) = \frac{\partial F_{X|Y}(x\,|y)}{\partial x} = \frac{f(x,y)}{f_Y(y)} \quad \square$$

*Example 9.7:*

- A rope of length $L$ is cut into three pieces in the following way:
  - The first piece of length $X$ is obtained by cutting the rope at random.
  - The second piece of length $Y$ is obtained by cutting the remaining segment of length $L-X$ at random
  - The third piece is obtained as the remaining segment of length $L-X-Y$.

  (a) Find $f_{Y|X}(y|x)$, the conditional PDF of $Y$ given $X=x$ $(0<x<L)$.

  (b) Find $f(x,y)$, the Joint PDF of $X$ and $Y$, and illustrate the region of the plane where it takes on non-zero values.

  (c) What is the probability that both $X$ and $Y$ be less than $L/2$?

◀

## 9.5 Independent RVs

*Definition:* We say that RVs $X$ and $Y$ are independent if the events $\{X \in A\}$ and $\{Y \in B\}$ are independent for any real number subsets $A$ and $B$, that is:

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \tag{9.54}$$

for any $A \subseteq \mathbb{R}$ and $B \subseteq \mathbb{R}$.

*Link with joint CDF:*

- Let $F(x, y)$ denote the joint CDF of RVs $X$ and $Y$. If $X$ and $Y$ are independent, we have

$$\begin{aligned} F(x,y) &= P(X \leq x, Y \leq y) \\ &= P(X \leq x)P(Y \leq y) \\ &= F_X(x)F_Y(y) \end{aligned} \tag{9.55}$$

- Conversely, it can be shown that if $F(x, y) = F_X(x)F_Y(y)$ for all $(x, y) \in \mathbb{R}^2$, then $X$ and $Y$ are independent.

*Theorem 9.6:* $X$ and $Y$ are independent if and only if

$$F(x, y) = F_X(x)F_Y(y), \quad \text{all } (x, y) \in \mathbb{R}^2 \tag{9.56}$$

### 9.5.1 Discrete case

*Theorem 9.7:* Let $X$ and $Y$ be discrete RVs with joint PMF $p(x,y)$. $X$ and $Y$ are independent if and only if

$$p(x,y) = p_X(x)\, p_Y(y), \qquad \text{all } (x,y) \in \mathbb{R}^2 \tag{9.57}$$

*Proof:* Suppose that $X$ and $Y$ are independent. Then,

$$\begin{aligned} p(x,y) &= P(X=x, Y=y) \\ &= P(X=x)P(Y=y) \\ &= p_X(x)p_Y(y) \end{aligned}$$

Conversely, suppose that (9.57) is satisfied. Let $\mathcal{R}_X = \{x_1, x_2, ...\}$ and $\mathcal{R}_Y = \{y_1, y_2, ...\}$ denote the sets of possible values of $X$ and $Y$, respectively. Then, for any real number subsets $A$ and $B$, we have

$$\begin{aligned} P(X \in A, Y \in B) &= \sum_{x_i \in A} \sum_{y_j \in B} p(x_i, y_j) \\ &= \sum_{x_i \in A} p_X(x_i) \sum_{y_j \in B} p_Y(y_j) \\ &= P(X \in A)P(Y \in B) \quad \square \end{aligned}$$

*Example 9.8:*

▶ Consider 20 independent flips of a fair coin. What is the probability of 6 heads in the first 10 flips and 4 heads in the next 10 flips? ◀

### 9.5.2 Continuous case

*Theorem 9.8:* Let $X$ and $Y$ be continuous RVs with joint PDF $f(x,y)$. $X$ and $Y$ are independent if and only if

$$f(x,y) = f_X(x) f_Y(y), \quad \text{all } (x,y) \in \mathbb{R}^2 \tag{9.58}$$

*Example 9.9:*

► Suppose $X$ and $Y$ are independent RVs, each being exponentially distributed with parameter $\lambda = 1$. Find $P(Y > X + 1)$ ? ◄

### 9.5.3 Miscellaneous results

*Theorem 9.9:* If RVs $X$ and $Y$ are independent, so are $U = g(X)$ and $V = h(Y)$, for any functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$.

*Application:* Suppose that $X$ and $Y$ are independent RVs. Then so are RVs $\sin(X^2)$ and $e^{Y-1}$

*Theorem 9.10:* If $X$ and $Y$ are independent, then $F_{X|Y}(x\,|y) = F_X(x)$. Furthermore, if $X$ and $Y$ are jointly discrete, then $p_{X|Y}(x\,|y) = p_X(x)$, while if they are jointly continuous, then $f_{X|Y}(x\,|y) = f_X(x)$.

*Theorem 9.11:* Let $X$ and $Y$ be jointly normal, as defined in (9.28)-(9.29). Then $X$ and $Y$ are independent if and only if $\rho = 0$.

*Prof:* If $\rho = 0$ in (9.28)-(9.29), we immediately obtain

$$\begin{aligned} f(x,y) &= \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2} \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{1}{2\sigma_Y^2}(y-\mu_Y)^2} \\ &= f_X(x)\, f_Y(y) \end{aligned} \tag{9.59}$$

where the result of Theorem 9.4 has been used. Conversely, it can be shown that if $f(x,y)$ (9.28) is equal to the product of $f_X(x)$ and $f_Y(y)$ in (9.31), then we must have $\rho = 0$ $\square$.

## 9.6 Transformation of joint RVs

*Introduction:*

- Let $X$ and $Y$ be jointly continuous RVs with known PDF $f(x, y)$.
- In applications, we are often interested in evaluating the distribution of one or more RVs defined as a function of $X$ and $Y$, as in $h(X, Y)$.
- Here, we distinguish two relevant cases:
  - $h : \mathbb{R}^2 \to \mathbb{R}$
  - $h : \mathbb{R}^2 \to \mathbb{R}^2$
- In each case, we present a technique that can be used to determine the PDF of the transformed variables.

### 9.6.1 Transformation from $\mathbb{R}^2 \to \mathbb{R}$

*Problem formulation:*

- Let $Z = h(X, Y)$, where $h : \mathbb{R}^2 \to \mathbb{R}$.
- We seek the PDF of RV $Z$, say $g(z)$.

*Method of distribution:*

- For each $z \in \mathbb{R}$, find domain $D_z \subseteq \mathbb{R}^2$ such that

$$Z \leq z \Longleftrightarrow (X, Y) \in D_z \tag{9.60}$$

- Express the CDF of $Z$ as:

$$\begin{aligned} G(z) &= P(Z \leq z) \\ &= P((X,Y) \in D_z) = \iint_{D_z} f(x,y)\,dx\,dy \end{aligned} \tag{9.61}$$

- Find the PDF by taking the derivative of $G(z)$:

$$g(z) = \frac{dG(z)}{dz} \tag{9.62}$$

*Example 9.10:*

▶ Let $X$ and $Y$ be uniformly distributed over the square $(0,1)^2 \subseteq \mathbb{R}^2$. Find the PDF of $Z = X + Y$. ◀

*Theorem 9.12* Let $X$ and $Y$ be independent RVs with marginal PDFs $f_X(x)$ and $f_Y(y)$, respectively. The PDF of $Z = X + Y$ is given by

$$g(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx \tag{9.63}$$

*Remarks:*

- That is, the PDF of $Z$ is obtained as the convolution of the marginal PDFs of $X$ and $Y$.
- The proof is left as an exercise.
- Please note that the previous example is a special case of this theorem.

### 9.6.2 Transformation from $\mathbb{R}^2 \to \mathbb{R}^2$

*Introduction:*

- We consider the transformation $(U, V) = h(X, Y)$, where $h : \mathbb{R}^2 \to \mathbb{R}^2$.
- We seek the joint PDF of RV $U$ and $V$, say $g(u, v)$.
- The proposed approach is based on the following theorem, which provides a generalization of the method of transformation in Section 7.2.2

*Theorem 8.9:* For every $(u, v) \in \mathbb{R}^2$, let $x_i \equiv x_i(u, v)$ and $y_i \equiv y_i(u, v)$ $(i = 1, 2, ...)$ denote the distinct roots of the equation $(u, v) = h(x, y)$. The joint PDF of $U$ and $V$ may be expressed as

$$g(u, v) = \sum_i f(x_i, y_i)|J_i| \tag{9.64}$$

$$J_i = \det \begin{bmatrix} \frac{\partial x_i}{\partial u} & \frac{\partial x_i}{\partial v} \\ \frac{\partial y_i}{\partial u} & \frac{\partial y_i}{\partial v} \end{bmatrix} = \frac{\partial x_i}{\partial u}\frac{\partial y_i}{\partial v} - \frac{\partial y_i}{\partial u}\frac{\partial x_i}{\partial v} \tag{9.65}$$

*Remarks:*

- If the equation $(u, v) = h(x, y)$ has no root, then set $g(u, v) = 0$.
- In (8.52), $x_i$ and $y_i$ really stand for $x_i(u, v)$ and $y_i(u, v)$, respectively.
- The determinant $J_i$ in (405) is the so-called *Jacobian* of the inverse transformation $(u, v) \to (x_i, y_i)$.

*Example 9.11:*

► Assume $X$ and $Y$ are continuous with joint PDF $f(x, y)$. Let $U = X + Y$ and $V = X - Y$.
(a) Find the joint PDF $g(u, v)$ of $U$ and $V$.
(b) In the special case when $X$ and $Y$ are independent, find the marginal PDF of $U$, say $g_U(u)$. ◄

# Chapter 10

# Bivariate expectations

*Introduction:*

- In Chapters 6, 7 and 8, our discussions of the expectation operator have been limited to the case of a single RV taken in isolation, as in $E(X)$, and possible functions thereof, as in $E(g(X))$.
- In this Chapter, we extend the notion of expectation to the bivariate framework, where two RVs, say $X$ and $Y$, are jointly distributed.
- More generally, we shall consider expectations of the type $E(g(X,Y))$ where $X$ and $Y$ are jointly distributed.

## 10.1 Basic results

*Scope of our study:*

- Let $X$ and $Y$ be two RVs defined over a common sample space.
- We shall focus our attention onto two special cases of interest:
  - RVs $X$ and $Y$ are discrete with sets of possible values $\mathcal{R}_X = \{x_1, x_2, \ldots\}$ and $\mathcal{R}_Y = \{y_1, y_2, \ldots\}$, respectively, and joint PMF $p(x, y)$.
  - RVs $X$ are $Y$ are continuous with joint PDF $f(x, y)$
- We are mainly interested in computing expectations of the type $E(Z)$, $Z = h(X, Y)$ where $h : \mathbb{R}^2 \to \mathbb{R}$.
- We begin we a review of the (unified) definition of expectation, as given in Section 8.2.

*Definition:* Let $Z$ be an arbitrary RV. The expected value of $Z$ is defined as

$$E(Z) = \int_{\infty}^{\infty} z f_Z(z)\, dz \tag{10.1}$$

where $f_Z(z)$ denotes the (generalized) PDF of $Z$.

*Remarks:*

- This definition is identical to that in (8.8) and is applicable whether $Z$ is a discrete, continuous or mixed RV.

- In the special case where $Z$ is discrete with possible values $\{z_i\}$ and PMF $p_Z(z)$, (10.1) reduces to

$$E(Z) = \sum_i z_i \, p_Z(z_i)$$

- Now let $Z = h(X, Y)$, with $X$ and $Y$ as previously defined. Direct application of (10.1) to compute $E(Z) = E(h(X, Y))$ requires the knowledge of the PDF $f_Z(z)$, or the PMF $p_Z(z)$ in the discrete case.

- Computing $f_Z(z)$ may be very difficult in practice. Fortunately, as the following theorem states, it is not actually necessary to know $f_Z(z)$ explicitly to compute $E(Z)$.

*Theorem 10.1:* Let $Z = h(X, Y)$ where $h : \mathbb{R}^2 \to \mathbb{R}$. The expected value of $Z$ may be expressed in one of the following forms.

(a) If $X$ and $Y$ are jointly discrete:

$$E(Z) = E(h(X, Y)) = \sum_i \sum_j h(x_i, y_j) p(x_i, y_j) \tag{10.2}$$

(b) If $X$ and $Y$ are jointly continuous:

$$E(Z) = E(h(X, Y)) = \int_\infty^\infty \int_\infty^\infty h(x, y) f(x, y) \, dx \, dy \tag{10.3}$$

*Remarks:*

- The proof of this theorem is beyond the scope of the course.
- Below, we verify the validity of the theorem in a very simple case. Consider the continuous case and let $h(X,Y) = X$. Then according to Theorem 10.1:

$$\begin{aligned} E(X) &= \int_{\infty}^{\infty}\int_{\infty}^{\infty} x\, f(x,y)\, dx\, dy \\ &= \int_{\infty}^{\infty} x \left( \int_{-\infty}^{\infty} f(x,y)\, dy \right) dx \\ &= \int_{\infty}^{\infty} x f_X(x)\, dx \end{aligned} \tag{10.4}$$

  which corresponds precisely to the definition of $E(X)$.

*Example 10.1:*

▶ Let RVs $X$ and $Y$ be jointly uniform over the region $D = \{(x,y) : 0 < x < y < 1\}$. Find $E(X), E(Y), E(X^2), E(Y^2)$ and $E(XY)$.

*Solution:* The region $D$ is illustrated below:

Note that the area of $D$ is $1/2$. Since $X$ and $Y$ are jointly uniform over that region, their joint PDF is given by

$$f(x,y) = \begin{cases} 2, & (x,y) \in D \\ 0, & \text{otherwise} \end{cases}$$

Using $f(x,y)$, the desired expectations can be easily obtained as follows:

$$\begin{aligned} E(X) &= \iint_{\mathbb{R}^2} x\, f(x,y)\, dx\, dy = 2 \iint_D x\, dx\, dy \\ &= 2\int_0^1 dx\, x \int_x^1 dy = 2\int_0^1 dx\, x\,(1-x^2) \\ &= 2\left(\frac{x^2}{2} - \frac{x^3}{3}\right)\Big|_0^1 = \frac{1}{3} \end{aligned}$$

In the same way, we find:

$$\begin{aligned} E(Y) &= 2\iint_D y\, dx\, dy = 2\int_0^1 dx \int_x^1 y\, dy \\ &= 2\int_0^1 dx\, (\frac{1}{2} - \frac{x^2}{2}) = 2\left(\frac{x^2}{2} - \frac{x^3}{6}\right)\Big|_0^1 = \frac{2}{3} \end{aligned}$$

$$\begin{aligned} E(X^2) &= 2\int_0^1 dx\, x^2 \int_x^1 dy = 2\int_0^1 dx\, (x^2 - x^3) \\ &= 2\left(\frac{x^3}{3} - \frac{x^4}{4}\right)\Big|_0^1 = \frac{1}{6} \end{aligned}$$

$$\begin{aligned} E(Y^2) &= 2\int_0^1 dx \int_x^1 dy\, y^2 = \frac{2}{3}\int_0^1 dx\, (1-x^3) \\ &= \frac{2}{3}\left(x - \frac{x^4}{4}\right)\Big|_0^1 = \frac{1}{2} \end{aligned}$$

$$\begin{aligned} E(XY) &= 2\int_0^1 dx\, x \int_x^1 dy\, y = \int_0^1 dx\, x(1-x^2) \\ &= \left(\frac{x^2}{2} - \frac{x^4}{4}\right)\Big|_0^1 = \frac{1}{4} \end{aligned}$$

◀

*Theorem 10.2:*

$$E(\sum_k \alpha_k h_k(X,Y)) = \sum_k \alpha_k E(h_k(X,Y)) \tag{10.5}$$

*Proof (continuous case):*

$$\begin{aligned} E(\sum_k \alpha_k h_k(X,Y)) &= \int_\infty^\infty \int_\infty^\infty (\sum_k \alpha_k h_k(x,y)) f(x,y)\, dx\, dy \\ &= \sum_k \alpha_k \int_\infty^\infty \int_\infty^\infty h_k(x,y) f(x,y)\, dx\, dy \\ &= \sum_k \alpha_k E(h_k(X,Y)) \quad \square \end{aligned} \tag{10.6}$$

*Remarks:*

- $E(.)$ acts linearly on its arguments:
- This theorem is useful when computing the expectation of complex random expressions. For example:

  $$E(3X^2 + 2\sin(XY)) = 3E(X^2) + 2E(\sin(XY))$$

  Note however that in general, $E(\sin(XY)) \neq \sin(E(XY))$.

*Theorem 10.3:* Suppose RVs $X$ and $Y$ are independent. Then

$$E(g(X)h(Y)) = E(h(X))E(g(Y)) \tag{10.7}$$

*Proof (discrete case):* Suppose $X$ and $Y$ are independent. Then, their joint PMF is expressible as $p(x,y) = p_X(x)p_Y(y)$ and

$$\begin{aligned} E(g(X)h(Y)) &= \sum_i \sum_j g(x_i)h(y_j)\, p(x_i, y_j) \\ &= (\sum_i g(x_i)p_X(x_i))\,(\sum_j h(y_j)p_Y(y_j)) \\ &= E(g(X))E(h(Y)) \quad \square \end{aligned} \tag{10.8}$$

*Remarks:*

- As a special case of Theorem 10.3, if RVs $X$ and $Y$ are independent, then $E(XY) = E(X)E(Y)$.
- Equivalently, $E(XY) \neq E(X)E(Y)$ implies that $X$ and $Y$ are not independent. However, $E(XY) = E(X)E(Y)$ does not imply that $X$ and $Y$ are independent in general.

## 10.2 Covariance and correlation

*Introduction:*

- Recall the definition of the variance for a single RV $X$, i.e.:

$$Var(X) = E[(X - \mu_X)^2)] \tag{10.9}$$

- Here, we study a generalization of the concept of variance, called the covariance, which is applicable to a pair of jointly distributed RVs.

*Definition:* Let $X$ and $Y$ be jointly distributed with mean $\mu_X$ and $\mu_Y$, respectively. The covariance of $X$ and $Y$ is defined as

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{10.10}$$

*Remarks:*

- We note that in the special case $Y = X$, $Cov(X, Y) = Var(X)$.
- In order to develop an intuitive feel for the concept of variance, we need to further analyze its properties.
- This analysis will make use of the following property of bivariate expectation, stated as a lemma without proof.

*Lemma 10.4 (Cauchy-Schwarz inequality):*

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)} \tag{10.11}$$

with equality if and only if $Y = \alpha X$ for some constant $\alpha$.

*Theorem 10.5:* Basic properties of the covariance

(a) $Cov(X, X) = Var(X) = \sigma_X^2$

(b) $Cov(X, Y) = Cov(Y, X)$

(c) $Cov(aX + b, cY + d) = ac\,Cov(X, Y)$

(d) $|Cov(X, Y)| \leq \sigma_X \sigma_Y$

(e) $Cov(X, Y) = E(XY) - E(X)E(Y)$

(f) If $X$ and $Y$ are independent, then $Cov(X, Y) = 0$

*Proof:* Properties (a), (b) and (c) follow trivially from the definition of the covariance. To prove (d), define

$$X' = X - \mu_X \quad \text{and} \quad Y' = Y - \mu_Y$$

and note that

$$Cov(X, Y) = E(X'Y'), \quad E(X'^2) = \sigma_X^2 \quad \text{and} \quad E(Y'^2) = \sigma_Y^2.$$

Then, making use of Lemma 10.4, we have

$$|Cov(X, Y)| = |E(X'Y')| \leq \sqrt{E(X'^2)E(Y'^2)} = \sigma_X \sigma_Y$$

Property (e) can be proved as follows:

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E(XY-\mu_X Y-\mu_Y X+\mu_X\mu_Y) \\ &= E(XY)-\mu_X E(Y)-\mu_Y E(X)+\mu_X\mu_Y \\ &= E(XY)-\mu_X\mu_Y \end{aligned}$$

Finally, for (f), we note from Theorem 10.3 that if $X$ and $Y$ are independent:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(X-\mu_X)E(Y-\mu_Y) = 0 \quad \square$$

*Example 10.2:*

► Consider the joint RVs $X$ and $Y$ as defined in Example 10.1. Find $Cov(X,Y)$.

*Solution:* We have

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{4} - \frac{2}{9} = \frac{1}{36}$$

◄

*Example 10.3:*

► Let $X$ and $Y$ be jointly distributed. Express $Var(X+Y)$ in terms of $\sigma_X^2, \sigma_Y^2$ and $Cov(X,Y)$.

$$\begin{aligned} Var(X+Y) &= E((X+Y)^2) - (E(X+Y))^2 \\ &= E(X^2+Y^2+2XY) - (\mu_X+\mu_Y)^2 \\ &= E(X^2)+E(Y^2)+2E(XY) \\ &\quad -\mu_X^2 - \mu_Y^2 - 2\mu_X\mu_Y \\ &= Var(X)+Var(Y)+2Cov(X,Y) \end{aligned} \tag{10.12}$$

Observe that when $Cov(X,Y)=0$, then $Var(X+Y) = Var(X)+Var(Y)$ ◄

*Definition:* The correlation coefficient of RVs $X$ and $Y$ is defined as

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{10.13}$$

*Remark:* The main difference between $\rho(X,Y)$ and $Cov(X,Y)$ is that the former has been normalized and is a dimensionless quantity. In effect (see Theorem below), we have that $-1 \leq \rho(X,Y) \leq 1$

*Example 10.4:*

► Find the correlation coefficient of the joint RVs $X$ and $Y$ in Example 10.1.

*Solution:* From Example 10.1, recall that

$$E(X) = \mu_X = \frac{1}{3}, \quad E(Y) = \mu_Y = \frac{2}{3}, \quad E(X^2) = \frac{1}{6}, \quad E(Y^2) = \frac{1}{2}$$

This information is first used to compute $Var(X)$ and $Var(Y)$:

$$\sigma_X^2 = Var(X) = E(X^2) - \mu_X^2 = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}$$

$$\sigma_Y^2 = Var(Y) = E(Y^2) - \mu_Y^2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}$$

Finally, we obtain:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{1/36}{1/18} = \frac{1}{2}$$

◄

*Theorem 10.6:* Basic properties of the correlation coefficient:

(a) $\rho(X, X) = 1$

(b) $\rho(X, Y) = \rho(Y, X)$

(c) $\rho(aX + b, cY + d) = \epsilon\, \rho(X, Y)$, where $\epsilon = \text{sign}(ac)$.

(d) $|\rho(X, Y)| \leq 1$

(e) $\rho(X, Y) = 1 \Longleftrightarrow Y = aX + b$ for some $a > 0$
$\rho(X, Y) = -1 \Longleftrightarrow Y = aX + b$ for some $a < 0$

(f) If $X$ and $Y$ are independent, then $\rho(X, Y) = 0$

*Discussion:*

- The proofs of these properties parallel those under Theorem 10.5 and are left as exercise for the students.
- The correlation coefficient $\rho(X, Y)$ provides a measure of the degree (and sign) of linear association between RVs $X$ and $Y$.
- If $\rho(X, Y) = 1$, then $Y = aX + b$ for some real numbers $a > 0$ and $b$. That is, let $L = \{(x, y) : y = ax + b\}$. Then $P((X, Y) \in L) = 1$.
- If $0 < \rho(X, Y) < 1$, we have an intermediate situation: the contour curves of the joint pdf of $X$ and $Y$ are more or less concentrated along some line $L = \{(x, y) : y = ax + b\}$ with positive slope $a > 0$:

- A similar interpretation applies for negative values of $\rho(X,Y)$, but the slope of line $L$ is now negative:

- Standard terminology:
  - if $\rho(X,Y) > 0$, we say that $X$ and $Y$ are positively correlated
  - if $\rho(X,Y) < 0$, we say that $X$ and $Y$ are negatively correlated
  - if $\rho(X,Y) = 0$, we say that $X$ and $Y$ are uncorrelated
- A final note of caution: if RVs $X$ and $Y$ are independent, then $\rho(X,Y) = 0$ and they are uncorrelated. However, if RVs $X$ and $Y$ are uncorrelated, they are note necessarily independent.

*Theorem 10.7:* Let $X$ and $Y$ be jointly normal with parameters $\mu_x$, $\mu_y$, $\sigma_X$, $\sigma_Y$ and $\rho$, i.e. with joint PDF as defined in (9.28)-(9.29). The parameter $\rho$ in these equations is precisely the correlation coefficient of $X$ and $Y$:

$$\rho(X,Y) \triangleq \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \rho \tag{10.14}$$

*Proof:* Calculus manipulations left as an optional exercise.

*Corollary:* Let $X$ and $Y$ be jointly normal: Then $X$ and $Y$ are independent if and only if $\rho(X,Y) = 0$.

*Remarks:*

- This is an immediate consequence of Theorems 10.7 and 9.11.
- Recall: $X$ and $Y$ independent $\Rightarrow \rho(X,Y) = 0$
- In general, $\rho(X,Y) = 0$ does not imply that $X$ and $Y$ are independent.
- However, if $X$ and $Y$ are jointly normal, then $\rho(X,Y) = 0$ implies independence.

*Example 10.5:*

► Suppose $X$ and $Y$ are independent, normally distributed RVs with mean $\mu_X = \mu_Y = 0$ and variance $\sigma_X^2$ and $\sigma_Y^2$. Let $U$ and $V$ be defined via the transformation

$$\begin{aligned} U &= \frac{1}{\sqrt{2}}(X - Y) \\ V &= \frac{1}{\sqrt{2}}(X + Y) \end{aligned} \tag{10.15}$$

Find $\rho(X,Y)$ and $\rho(U,V)$. ◄

## 10.3 Conditional expectations

*Conditional distributions (Recap):*

- Let $X$ and $Y$ be two RVs defined on the same sample space:
- If $X$ and $Y$ are discrete with joint PMF $p(x, y)$, then

$$p_{X|Y}(x|y) = \frac{p(x,y)}{p_Y(y)} \tag{10.16}$$

- If $X$ and $Y$ are continuous with joint PDF $f(x, y)$, then

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \tag{10.17}$$

*Definition:* The conditional expectation of $X$ given $Y = y$ is defined as

$$E(X|Y=y) = \begin{cases} \sum_i x_i \, p_{X|Y}(x_i|y) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} x \, f_{X|Y}(x|y)\, dx & \text{in the continuous case} \end{cases} \tag{10.18}$$

*Remarks:*

- Conceptually, $E(\,.\,|Y = y)$ is similar to the conventional expectation $E(\,.\,)$, except that it is based on conditional PMF or PDF.
- All the properties of $E(.)$ extend to $E(\,.\,|Y = y)$

*Theorem 10.8:*

$$E(X) = \begin{cases} \sum_i E(X|Y=y_i)\, p_Y(y_i) & \text{discrete case} \\ \int_{-\infty}^{\infty} E(X|Y=y)\, f_Y(y)\, dy & \text{continuous case} \end{cases} \tag{10.19}$$

*Proof (continuous case):* From Theorem 10.1, equation (10.3), we have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)\, dx\, dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\, f_Y(y)\, dx\, dy \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\, dx\right) f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} E(X|Y=y)\, f_Y(y)\, dy \quad \square \end{aligned} \tag{10.20}$$

*Remark:* Theorem 10.8 is very useful when:

- the direct evaluation of $E(X)$ is not trivial
- but $E(X|Y=y)$ may be computed easily

*Special notation:*

- Let $h(y) \triangleq E(X|Y=y)$, so that Theorem 10.8 may be expressed as

$$E(X) = \int h(y) f_Y(y)\, dy = E(h(Y)) \tag{10.21}$$

- Now, introducing the notation $E(X|Y) \triangleq h(Y)$, Theorem 10.8 can be written compactly as

$$E(X) = E(E(X|Y)) \tag{10.22}$$

- This notation may appear confusing at first, but it is often used.

*Example 10.6:*

▶ The number of people who pass by a store during lunch time (say form 12:00 to 13:00) is a Poisson RV with parameter $\lambda = 100$. Assume that each person may enter the store, independently of the other people, with a given probability $p = .15$. What is the expected number of people who enter the store during lunch time?

*Solution:* Define the following RVs:

$$\begin{aligned} X &= \text{number of people entering the store} \\ Y &= \text{number of people passing by} \end{aligned}$$

RV $Y$ is Poisson with parameter $\lambda = 100$:

$$p_Y(y) = P(Y=y) = \begin{cases} \frac{\lambda^y}{y!} e^{-\lambda}, & y = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$$

Given $Y = y$, RV $X$ is $B(y, p)$ and therefore:

$$E(X|Y=y) = yp$$

Finally, applying Theorem 10.8, we obtain:

$$\begin{aligned} E(X) &= \sum_{y=0}^{\infty} E(X|Y=y) p_Y(y) = \sum_{y=0}^{\infty} yp \frac{\lambda^y}{y!} e^{-\lambda} \\ &= p\, e^{-\lambda} \sum_{y=1}^{\infty} \frac{\lambda^y}{(y-1)!} = p\, \lambda\, e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\ &= p\, \lambda = 0.15 \times 100 = 15 \end{aligned}$$

◀

# Chapter 11

# Multivariate distributions

*Introduction:*

- In engineering applications of probability, we often have to deal with several RVs (i.e. more than 2). Examples of this include the design and analysis of: digital receivers, speech recognition systems, routing algorithms for packet networks, etc.
- In Chapters 9 and 10, we developed probability models and techniques for the solution of problems involving two random variables jointly defined over a common sample space.
- In this chapter, we extend these concepts to the case of multiple (say $n \geq 2$) random variables. The developments are conceptually simple but the notation is at times a bit tedious.
- We also discuss briefly the multivariate normal distributions, which find several important applications in science and engineering.

## 11.1 Probability functions

*Joint CDF:*

- Let $X_1, X_2, ..., X_n$ be $n$ RVs defined on the same sample space.
- The joint CDF of $X_1, \ldots, X_n$ is defined as

$$F(x_1, x_2, ..., x_n) \triangleq P(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n) \tag{11.1}$$

- Some of the most important properties of the joint CDF:
  (a) $F(x_1, x_2, ..., x_n)$ is non-decreasing in each of its arguments.
  (b) $F(x_1, x_2, ..., x_n)$ is right-continuous in each of its arguments.
  (c) For any particular $i$, (all other coordinates being fixed)

$$\lim_{x_i \to -\infty} F(x_1, x_2, ..., x_n) = 0 \tag{11.2}$$

  (d) In the limit $x_i \to \infty$ for all $i$, $F(x_1, x_2, ..., x_n) \to 1$, i.e.:

$$F(\infty, ..., \infty) = 1 \tag{11.3}$$

- The marginal CDF of $X_i$ is obtained by letting $x_j \to \infty$ for all $j \neq i$:

$$F_{X_i}(x_i) = \lim_{\substack{x_j \to \infty \\ \text{all } j,\ j \neq i}} F(x_1, x_2, ..., x_n) \tag{11.4}$$

*Joint PMF:*

- Suppose that RVs $X_i$ $(i = 1, ..., n)$ are discrete with set of possible values $\mathcal{R}_i$, respectively.
- The joint PMF of $X_1, \ldots, X_n$ is defined as

$$p(x_1, x_2, ..., x_n) \triangleq P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) \tag{11.5}$$

- Some of the most important properties of the joint PMF:
  (a) $p(x_1, x_2, ..., x_n) \geq 0$.
  (b) If for some $i$, $x_i \notin \mathcal{R}_i$, then $p(x_1, x_2, ..., x_n) = 0$
  (c) Normalization:

$$\sum_{x_1 \in \mathcal{R}_1} \cdots \sum_{x_n \in \mathcal{R}_n} p(x_1, x_2, ..., x_n) = 1 \tag{11.6}$$

- The marginal PMF of $X_i$ is obtained by summing over all possible values of $x_j$, for all $j \neq i$:

$$p_{X_i}(x_i) = \sum_{\substack{x_j \in \mathcal{R}_j \\ \text{all } j,\ j \neq i}} p(x_1, x_2, ..., x_n) \tag{11.7}$$

*Joint PDF:*

- We say that $X_i$ are jointly continuous RVs if there exists an integrable function $f : \mathbb{R}^n \to [0, \infty)$, called the joint PDF, such that for any region $D \subseteq \mathbb{R}^n$:

$$P((X_1, \ldots, X_n) \in D) = \int \ldots \int_D f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \tag{11.8}$$

- The following properties follow from the above definition:

  (a) $f(x_1, \ldots, x_n) \geq 0$

  (b) Normalization:

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = 1 \tag{11.9}$$

- Relationships between the joint PDF and joint CDF:

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n} \tag{11.10}$$

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} dt_1 \ldots \int_{-\infty}^{x_n} dt_n\, f(t_1, \ldots, t_n) \tag{11.11}$$

- The marginal PDF of $X_i$ is obtained by integrating over all $x_j$, $j \neq i$:

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) \underbrace{dx_1 \ldots dx_n}_{\text{omit } dx_i} \tag{11.12}$$

  where the integration is $(n-1)$-fold.

*Example 11.1:*

► Random variables $X$, $Y$ and $Z$ are uniformly distributed over the sphere $D = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq 1\}$. That is

$$f(x, y, z) = \begin{cases} k, & (x, y, z) \in D \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the constant $k$.
(b) Find the probability that $P(Z > 0)$.
(c) More generally, find $P(aX + bY + cZ > 0)$ for any real numbers $a$, $b$ and $c$.

*Solution:* (a) Using the normalization condition (11.9):

$$\iiint_{\mathbb{R}^3} f(x, y, z)\, dx\, dy\, dy = k \iiint_D dx\, dy\, dy = k\, \text{Vol}(D) = 1$$

which implies

$$k = \frac{1}{\text{Vol}(D)} = \frac{3}{4\pi}$$

(b) Define $E = \{(x, y, z) \in \mathbb{R}^3 : z > 0\}$. We seek

$$\begin{aligned} P(Z > 0) &= \iiint_E f(x, y, z)\, dx\, dy\, dz \\ &= k \iiint_{E \cap D} dx\, dy\, dz \\ &= \frac{\text{Vol}(E \cap D)}{\text{Vol}(D)} = \frac{1}{2} \end{aligned}$$

(c) The answer is also 1/2... ◄

## 11.2 Conditional distributions and independence

*Conditional distributions:*

- The material of Section 9.4 on conditional distributions can also be extended to the multivariate case.
- Without loss of generality, suppose we are interested in the conditional distribution of RVs $X_1, \ldots, X_k$, given the knowledge of the remaining $n-k$ variables, i.e. $X_{k+1}, \ldots, X_n$.
- In the discrete case, we define the conditional PMF as follows:

$$p_{X_1 \ldots X_k | X_{k+1} \ldots X_n}(x_1, ..., x_k \,|x_{k+1}, ..., x_n) = \frac{p(x_1, ..., x_n)}{p_{X_{k+1} \ldots X_n}(x_{k+1}, ..., x_n)} \quad (11.13)$$

  where the denominator is assumed to be non-zero. For given values of $x_{k+1}, ..., x_n$, the above conditional PMF is a valid PMF in $k$ dimensions.
- In the continuous case, we define the conditional PDF as follows:

$$f_{X_1 \ldots X_k | X_{k+1} \ldots X_n}(x_1, ..., x_k \,|x_{k+1}, ..., x_n) = \frac{f(x_1, ..., x_n)}{f_{X_{k+1} \ldots X_n}(x_{k+1}, ..., x_n)} \quad (11.14)$$

  For given values of $x_{k+1}, ..., x_n$, (11.14) is a valid PDF in $k$ dimensions. For example, in the case of 3 RVs, say $X$, $Y$ and $Z$, we have

$$f_{X|YZ}(x|y,z) = \frac{f(x,y,z)}{f_{YZ}(y,z)} \geq 0,$$

  where we assume $f_{YZ}(y,z) > 0$, and

$$\int_{-\infty}^{\infty} f_{X|YZ}(x|y,z)\, dx = 1.$$

*Independence:*

- We say that RVs $X_1, \ldots, X_n$ are independent iff for any real number subsets $A_i \subseteq \mathbb{R}$ $(i = 1, \ldots, n)$, the events $\{X_1 \in A_1\}$,...,$\{X_n \in A_n\}$ are mutually independent. This implies:

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \ldots P(X_n \in A_n) \qquad (11.15)$$

- It can be shown that discrete RVs $X_1, \ldots, X_n$ are independent iff

$$p(x_1, \ldots, x_n) = p_{X_1}(x_1) \ldots p_{X_n}(x_n) \qquad (11.16)$$

- Similarly, continuous RVs $X_1, \ldots, X_n$ are independent iff

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n) \qquad (11.17)$$

*Example 11.2:*

► You buy $n$ identical memory chips. Let $X_i \geq 0$ denote the lifetime of the $i$th chip. Assuming that the RVs $X_i$ are independent and identically distributed, find the probability that the chip #1 outlasts all the others.

*Solution:* Define the event

$$A = \{ \text{ chip \#1 outlasts all the others } \}$$

Intuitively, since the chips are identical, we should have $P(A) = 1/n$. Let us verify that this is indeed the case. Assuming that the RVs $X_i$ are independent and identically distributed (i.i.d.), we have

$$f(x_1, \ldots, x_n) = f_X(x_1) \ldots f_X(x_n)$$

where $f_X(.)$ denotes the common marginal PDF of the individual RVs. Note here that $f_X(x) = 0$ for $x < 0$. We seek

$$\begin{aligned} P(A) &= P(X_1 \geq X_2, X_1 \geq X_3, \ldots, X_1 \geq X_n) \\ &= \int_0^\infty dx_1 \int_0^{x_1} dx_2 \ldots \int_0^{x_1} dx_n \, f(x_1, \ldots, x_n) \\ &= \int_0^\infty dx_1 f_X(x_1) \left( \int_0^{x_1} dy \, f_X(y) \right)^{n-1} \\ &= \int_0^\infty dx_1 \, f_X(x_1) \, (F_X(x_1))^{n-1} \end{aligned}$$

where $F_X(x)$ is the CDF associated to $f_X(x)$. To evaluate the integral, we make the following change of variables:

$$u = F_X(x_1), \quad du = f_X(x_1) \, dx_1$$

The new limits of integration become $F_X(0) = P(X \leq 0) = 0$ and $F_X(\infty) = 1$. Therefore:

$$P(A) = \int_0^1 u^{n-1} du = \frac{u^n}{n} \Big|_0^1 = \frac{1}{n}$$

Interestingly, we could solve this problem without knowing the explicit form of $f_X(x)$, the common PDF of the lifetime RVs $X_i$. ◄

## 11.3 Transformation of multiple RVs

*Transformation theorem:*

- The transformation theorem admits a direct extension to $\mathbb{R}^n$.
- Let RVs $U_1, \ldots, U_n$ be defined in terms of $X_1, \ldots, X_n$ via

  $$(U_1, \ldots, U_n) = h(X_1, \ldots, X_n)$$

  where $h : \mathbb{R}^n \to \mathbb{R}^n$.
- For any $(u_1, \ldots, u_n) \in \mathbb{R}^n$, let $(x_{1i}, \ldots, x_{ni})$ denote the $i$th distinct root of the equation $(u_1, \ldots, u_n) = h(x_1, \ldots, x_n)$.
- The joint PDF of $U_1, \ldots, U_n$ is given by

  $$g(u_1, \ldots, u_n) = \sum_i f(x_{1i}, \ldots, x_{ni}) |J_i| \tag{11.18}$$

  where

  $$J_i = \det \begin{bmatrix} \frac{\partial x_{1i}}{\partial u_1} & \cdots & \frac{\partial x_{1i}}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_{ni}}{\partial u_1} & \cdots & \frac{\partial x_{ni}}{\partial u_n} \end{bmatrix} \tag{11.19}$$

- In (11.18), the sum is over all the roots $(x_{1i}, \ldots, x_{ni})$. If for a given point $(u_1, \ldots, u_n) \in \mathbb{R}^n$ there is no such root, than $g(u_1, \ldots, u_n) = 0$.

*Example 11.3:*

▶ Let $X$, $Y$ and $Z$ be independent and identically (iid) distributed RVs with common $N(0, 1)$ PDF. Find the joint PDF of corresponding spherical coordinates:

$$\begin{aligned} R &= \sqrt{X^2 + Y^2 + Z^2} \\ \Phi &= \angle(X, Y) \\ \Theta &= \angle(\sqrt{X^2 + Y^2}, Z) \end{aligned}$$

◀

## 11.4 Multivariate expectations

*Introduction:*

- Recall the definition of the expectation of a single random variable $Y$:

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)\, dy$$

  where $f_Y(y)$ denotes the PDF of $Y$. In the special case where $Y$ is discrete RV with set of possible values $\mathcal{R}_Y$, this reduces to

$$E(Y) = \sum_{y \in \mathcal{R}_Y} y\, p_Y(y)$$

  where $p_Y(y)$ denotes the DPF of $Y$.

- Let RVs $X_1, \ldots, X_n$ be defined over the same sample space. In this Section, we are interested in evaluating $E(Y)$ when $Y$ is a of function $X_1, \ldots, X_n$, say

$$Y = h(X_1, \ldots, X_n)$$

- As before, we focus on two special cases of interest:
  - RVs $X_i$ $(i = 1, \ldots, n)$ are discrete with sets of possible values $\mathcal{R}_{X_i}$, respectively, and joint PMF $p(x_1, \ldots, x_n)$.
  - RVs $X_i$ are continuous with joint PDF $f(x_1, \ldots, x_n)$.

*Theorem 11.1:* Let $Y = h(X_1, \ldots, X_n)$. Then $E(Y) = E(h(X_1, \ldots, X_n))$ may be expressed in one of the following forms:

(a) In the discrete case,

$$E(Y) = \sum_{x_1} \ldots \sum_{x_n} h(x_1, \ldots, x_n) p(x_1, \ldots, x_n) \tag{11.20}$$

(b) In the continuous case,

$$E(Y) = \int_{\infty}^{\infty} \ldots \int_{\infty}^{\infty} h(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \ldots dx_n \tag{11.21}$$

*Remark:* According to this Theorem, it is not necessary to know $f_Y(y)$ (or $p_Y(y)$) explicitly to compute the expected value of $Y = h(X_1, \ldots, X_n)$. Knowledge of the joint PDF $f(x_1, \ldots, x_n)$ (or joint PMF) is sufficient. The proof of the theorem is beyond the scope of this course.

*Corollary:*

(a) $E(\sum_{i=1}^{n} \alpha_i h_i(X_1, ..., X_n)) = \sum_{i=1}^{n} \alpha_i E(h_i(X_1, ..., X_n))$

(b) $E(\sum_{i=1}^{n} \alpha_i X_i) = \sum_{i=1}^{n} \alpha_i E(X_i)$

(c) $E(X_1 + X_2 + ... + X_n) = E(X_1) + ... + E(X_n)$

*Proof:* In the continuous case, (a) is obtained via the application of (11.21) as follows:

$$\begin{aligned} E(\sum_{i=1}^{n} \alpha_i h_i(X_1, ..., X_n)) &= \int_{\infty}^{\infty} ... \int_{\infty}^{\infty} (\sum_{i=1}^{n} \alpha_i h_i(x_1, ..., x_n)) f(x_1, ..., x_n) dx_1 ... dx_n \\ &= \sum_{i=1}^{n} \alpha_i \int_{\infty}^{\infty} ... \int_{\infty}^{\infty} h_i(x_1, ..., x_n) f(x_1, ..., x_n) dx_1 ... dx_n \\ &= \sum_{i=1}^{n} \alpha_i E(h_i(x_1, ..., x_n)) \end{aligned}$$

(b) is obtained as a special case of (a) with $h_i(X_1, ..., X_n) = X_i$, while (c) is a special case of (b) with $\alpha_i = 1$ □.

*Example 11.4:*

▶ A fair die is rolled 10 times. Let $Y$ denote the sum of the resulting outcomes. Find $E(Y)$.

*Solution:* Let RV $X_i$, with set of possible values $\mathcal{R}_i = \{1, \ldots, 6\}$, denote the outcome of the $i$th roll ($i = 1, ..., 10$). Since the die is fair,

$$E(X_i) = \sum_{k=1}^{6} k \frac{1}{6} = \frac{1}{6} \cdot \frac{6 \cdot 7}{2} = 3.5$$

Here, we have

$$Y = \sum_{i=1}^{10} X_i$$

Therefore, according to the Corollary,

$$E(Y) = \sum_{i=1}^{10} E(X_1) = 10 \cdot 3.5 = 35$$

◀

*Example 11.5:*

► A complex parallel computing system is made up of $n$ circuit boards connected by a fast bus. Each board contains two identical CPU chips that must both be non-defective for the circuit board to operate properly. What is the expected number of operational circuit boards after $m$ CPU chips have gone defective? ◄

*Theorem 11.2:* Suppose RVs $X_1, \ldots, X_n$ are independent. Let $h_i : \mathbb{R} \to \mathbb{R}$ be arbitrary functions of the real variable. Then

$$E(h_1(X_1) \ldots h_n(X_n)) = E(h_1(X_1)) \ldots E(h_n(X_n)) \tag{11.22}$$

*Corollary:* If $X_1, \ldots, X_n$ are independent, we have

$$E(X_1 X_2 \ldots X_n) = E(X_1)E(X_2) \ldots E(X_n) \tag{11.23}$$

## 11.5 Variance and covariance

*Introduction:*

- Previously given definitions of the variance of a single RV and the covariance of two jointly distributed RVs still apply in the multivariate framework.
- In particular, if $X_1, \ldots, X_n$ are jointly distributed RVs with respective means $\mu_1, \ldots, \mu_n$, then

$$Var(X_i) = E[(X_i - \mu_i)^2]$$

$$Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$$

- All the previously derived properties of the variance and covariance remain valid in the multivariate context. In particular:
  - $Cov(X_i, X_j) = E(X_i X_j) - \mu_i \mu_j$
  - $X_i$ and $X_j$ independent implies $Cov(X_i, X_j) = 0$
- Generally, and in the same way as in Chapter 10, we say that RVs $X_i$ and $X_j$ $(i \neq j)$ are uncorrelated if

$$Cov(X_i, X_j) = 0$$

  Thus independence implies *uncorrelatedness* but the converse is not true in general.

- For future reference, we also note the following result: If RVs $X_1, X_2, ..., X_n$ are uncorrelated, then

$$Cov(X_i, X_j) = \begin{cases} Var(X_i), & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{11.24}$$

*Theorem 11.3:*

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} Cov(X_i, X_j) \tag{11.25}$$

*Remarks:*

- From (11.24), we note that if the RVs $X_i$ $(i = 1, \ldots n)$ are uncorrelated, that is if $Cov(X_i, X_j) = 0$ for all $i \neq j$, then

$$Var(X_1 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n) \tag{11.26}$$

- Clearly, this is the case when RVs $X_1, X_2, ..., X_n$ are independent.

*Example 11.6:*

► A fair die is rolled 10 times. Let $Y$ denote the sum of the resulting outcomes. Find $Var(Y)$.

*Solution:* Let RV $X_i$, with possible values $\{1, \ldots, 6\}$, denote the outcome of the $i$th roll ($i = 1, ..., 10$). Since the die is fair,

$$\begin{aligned} E(X_i) &= \sum_{k=1}^{6} k \frac{1}{6} = 3.5 \\ E(X_i^2) &= \sum_{k=1}^{6} k^2 \frac{1}{6} = \frac{87}{6} = 14.5 \\ Var(X_i) &= E(X_i^2) - (E(X_i))^2 = 2.25 \end{aligned}$$

Here, $Y = \sum_{i=1}^{10} X_i$. Since the RVs $X_i$ are independent, (11.26) can be applied and we obtain

$$Var(Y) = \sum_{i=1}^{10} Var(X_i) = 22.5$$

◄

*Example 11.7:*

► Let $X_1, \ldots, X_n$ be independent RVs with common mean $\mu_X$ and variance $\sigma_X^2$. Find the mean and variance of their *sample mean*, defined as

$$Y \triangleq \frac{1}{n}(X_1 + \cdots + X_n) \tag{11.27}$$

◄

# Chapter 12

# Limit Theorems

Suppose we flip a fair coin a large number of times, say $n$. Let $\eta(H, n)$ denote the number of trials, out of $n$, in which heads is observed. Intuitively, we know that for $n$ large, the relative frequency

$$\frac{\eta(H, n)}{n} \approx \frac{1}{2}$$

Let the outcome of the $ith$ flip be represented by the RV

$$X_i = \begin{cases} 1, & \text{if heads} \\ 0, & \text{if tails} \end{cases}$$

The RVs $X_1, \ldots, X_n$ are independent with common mean $\mu = E(X_i) = 1/2$. Also note that $X_1 + \cdots + X_n = \eta(H, n)$. Thus, (12.1) may be expressed in the equivalent form

$$\frac{1}{n} \sum_{i=1}^{n} X_i \approx \mu.$$

provided $n$ is large.

More generally, consider a collection of $n$ independent RVs, say $X_i$ ($i = 1, 2, ..., n$), with common mean $\mu = E(X_i)$ and variance $\sigma^2$. It has been observed in many practical situations that the so-called *sample mean*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \text{as} \quad n \to \infty \tag{12.1}$$

More strikingly, as $n$ gets larger, the distribution of the sample mean $\bar{X}_n$ is seen to approach that of a normal RV with mean $\mu$ and variance $\sigma^2/n$.

This type of regularity in the observed behavior of the sample mean and other related averages, as the number of repeated trials becomes increasingly large, provides the main motivation and justification for the development and application of modern probability theory. For example, it plays a central role in the development of statistical techniques of data analysis.

In this Chapter, we investigate the asymptotic behavior of the sample means and related averages, including the relative frequency. We show that the observed regularity of these quantities can be explained as a natural consequence of the concepts of independence and/or uncorrelatedness within the axiomatic framework of probability.

The main results of the Chapter take the form of so-called *limit theorems* that describe the behavior of these averages as the number of trials, say $n$, gets increasingly large. The theorems also provide a precise meaning for the type of convergence exhibited by these averages.

## 12.1 Some basic inequalities

*Theorem 12.1 (Markov's inequality):* Let $X$ be a non-negative RV; that is, the PDF of $X$, $f(x)$, satisfies $f(x) = 0$ for $x < 0$. Then, for any $t > 0$,

$$P(X \geq t) \leq \frac{E(X)}{t} \tag{12.2}$$

*Proof:* Because of the assumed property of $f(x)$, we have:

$$\begin{aligned} E(X) &= \int_0^\infty x\, f(x)\, dx \\ &\geq \int_t^\infty x\, f(x)\, dx \\ &\geq t \int_t^\infty f(x)\, dx = t\, P(X \geq t) \quad \square \end{aligned}$$

*Remarks:* Markov's inequality provides a rough bound on $P(X \geq t)$. Clearly, it is useful only for values of $t$ larger than $E(X)$.

*Example 12.1:*

► Consider the transmission of several 10Mbytes files over a noisy channel. Suppose that the average number of erroneous bits per transmitted file at the receiver output is $10^3$. What can be said about the probability of having $\geq 10^4$ erroneous bits during the transmission of one of these files?

*Solution:* Let $X$ denote number of erroneous bits in a given file transmission. We know that $E(X) = 10^3$. We want $P(X \geq 10^4)$, but we don't know the PDF of $X$. We can use Markov's inequality to obtain an upper bound on the desired probability:

$$P(X \geq 10^4) \leq \frac{E(X)}{10^4} = 10^{-1} \tag{12.3}$$

◄

*Theorem 12.2 (Chebyshev's inequality):* Let $X$ be a RV with expected value $\mu$ and variance $\sigma^2$. Then for any real number $t > 0$,

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \tag{12.4}$$

*Proof:* Introduce RV $Z = \frac{X-\mu}{\sigma}$; that is, $Z$ is the standardized X as defined in (8.17). Clearly, $|X - \mu| \geq t$ if and only if $Z^2 \geq t^2/\sigma^2$. Applying Markov's inequality to non-negative RV $Z^2$, we then have

$$P(|X - \mu| \geq t) = P(Z^2 \geq \frac{t^2}{\sigma^2}) \leq \frac{E(Z^2)}{t^2/\sigma^2} = \frac{\sigma^2}{t^2} \tag{12.5}$$

where the last equality follows because $E(Z^2) = 1$. □

*Remarks:*

- The LHS in (12.4) represents the probability that $X$ deviates from its mean $\mu$ by $t$ or more. According to (12.4), this probability is upper bounded by $\sigma^2/t^2$, which decays to zero as $t$ gets larger.
- Alternatively, setting $t = k\sigma$ in (12.4), we obtain

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{12.6}$$

*Example 12.2:*

▶ Let $X \sim N(0, 1)$ (i.e. standard normal). Using the table of the standard normal CFD, compute $P(|X| \geq t)$ for $t = 1, 2, 3$ and compare the results to Chebyshev inequality. ◀

## 12.2 Law of large numbers

*Definition:* We define the sample mean of RVs $X_1, \ldots, X_n$ as

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \tag{12.7}$$

*Theorem 12.3:* Suppose $X_i$ $(i = 1, \ldots, n)$ are uncorrelated with common mean $\mu$ and variance $\sigma^2$. The mean and variance of $\bar{X}_n$ are then given by

$$E(\bar{X}_n) = \mu, \quad Var(\bar{X}_n) = \frac{\sigma^2}{n} \tag{12.8}$$

*Proof:* For the expected value, we have,

$$\begin{aligned} E(\bar{X}_n) &= \frac{1}{n} E(X_1 + \cdots + X_n) \\ &= \frac{1}{n}[E(X_1) + \cdots + E(X_n)] \\ &= \frac{1}{n}\, n\, \mu = \mu \end{aligned}$$

Since the RVs $X_i$ are uncorrelated, we have $Cov(X_i, X_j) = 0$ for $i \neq j$. Then, the variance may be computed as follows:

$$\begin{aligned} Var(\bar{X}_n) &= \frac{1}{n^2}\, Var(X_1 + \cdots + X_n) \\ &= \frac{1}{n^2}\, [Var(X_1) + \cdots + Var(X_n)] \\ &= \frac{1}{n^2}\, n\, \sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

where the second line follows from (11.25). □

*Remarks:*

- The sample mean, as defined in (12.7), is equivalent to the arithmetic average of the RVs $X_i$.
- Think of the RVs $X_i$ as independent measurements of a physical quantity, with mean $\mu$ representing the true (usually unknown) value of this quantity, and $\sigma^2$ representing the variance of the measurement error.
- We note from (12.8) that the expected value of the sample mean is equal to the true mean. Accordingly, we say that the sample mean $\bar{X}_n$ is an unbiased estimator of $\mu$.
- We also note from (12.8) that increasing the number $n$ of independent measurements reduces the variance of the sample mean.
- The above desirable properties are consistent with the intuitive notion of repeating and averaging over several measurements to reduce, or smooth out the effects of the measurement errors.

*Theorem 12.4:* Let $X_i$ $(i = 1, 2, ..)$ be a sequence of uncorrelated RVs with common mean $\mu = E(X_i)$ and variance $\sigma^2 = Var(X_i) < \infty$. For any $\epsilon > 0$,

$$\lim_{n\to\infty} P(\,|\,\bar{X}_n - \mu\,|\,<\,\epsilon\,) = 1 \tag{12.9}$$

where $\bar{X}_n$ denote the sample mean (12.7).

*Proof:* From (12.8), the mean and variance of $\bar{X}_n$ are given by $\mu$ and $\sigma^2/n$, respectively. Thus, applying Chebyshev's inequality (12.4) to $\bar{X}_n$, we have that for any $\epsilon > 0$,

$$0 \leq P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \tag{12.10}$$

Finally, taking the limit on both sides as $n \to \infty$, we obtain:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$$

which is equivalent to (12.9). □

*Discussion:*

- Theorem 12.4 is called the weak law of large numbers (WLLN).
- The WLLN (12.9) admits the following interpretation: for large $n$, it is very likely that the sample mean $\bar{X}_n$ is close to $\mu$, that is,

  $$P(\mu - \epsilon \leq \bar{X}_n \leq \mu + \epsilon) \to 1 \quad \text{as} \quad n \to \infty \tag{12.11}$$

  regardless of how small $\epsilon$ is.
- The WLLN provides a theoretical basis for using the sample mean $\bar{X}_n$ as an estimate of $\mu$ in statistics. It can also be used to justify the relative frequency interpretation of probability (see next Section).
- Several more sophisticated variations of this result do exist. An example is the so-called strong law of large number (SLLN):

  $$P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1 \tag{12.12}$$

  which is applicable whenever the RVs $X_i$ are independent.

## 12.3 Relative frequency interpretation of probability

*Relative frequency:*

- Consider $n$ independent repetitions of the same random experiment (e.g. rolling a die $n$ times).
- Let $A$ denote an event that may or nor occur at each repetition (e.g. the die shows 5 or 6)
- Let $\eta(A, n)$ denote number of occurrences of event $A$ in $n$ repetitions. The ratio
$$\frac{\eta(A, n)}{n} \tag{12.13}$$
is called the relative frequency of event $A$.

*Discussion:*

- Historically, it has been observed that
$$\frac{n(A)}{n} \rightarrow \text{ constant} \quad \text{as} \quad n \rightarrow \infty \tag{12.14}$$
- This has motivated earlier definitions of the probability of $A$, namely:
$$P(A) = \lim_{n\rightarrow\infty} \frac{n(A)}{n} \tag{12.15}$$
- Below, we use the WLLN (Theorem 12.4) to reconcile this earlier definition of probability with the modern axiomatic definition.

*Theorem 12.5:* Consider $n$ independent repetitions of a random experiment in which event $A$ has been identified. For any $\epsilon > 0$, we have

$$\lim_{n\to\infty} P(\,|\,\frac{\eta(A,n)}{n} - P(A)\,|\,<\epsilon\,) = 1 \tag{12.16}$$

*Proof:* Define RVs $X_i$ $(i = 1, 2, ...)$ as follows:

$$X_i = \begin{cases} 1, & \text{if } A \text{ occurs at } i\text{th repetition} \\ 0, & \text{if not} \end{cases}$$

These RVs are independent (thus uncorrelated) with common mean

$$\mu = E(X_i) = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A)$$

and finite variance. Observe that

$$\bar{X}_n = \frac{1}{n}(X_1 + ... + X_n) = \frac{\eta(A,n)}{n}$$

Applying the WLLN (Theorem 12.4), we have:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = \lim_{n\to\infty} P(|\frac{\eta(A,n)}{n} - P(A)| < \epsilon) = 1 \quad \square$$

*Remarks:* Since $\epsilon > 0$ can be taken as small as we want, it is very likely that the relative frequency $\eta(A,n)/n$ is close to $P(A)$ for $n$ sufficiently large.

## 12.4 Central limit theorem

*Lemma 12.6:* Suppose RVs $X_1, X_2, \ldots, X_n$ are independent with characteristic functions $\psi_1(\omega), \psi_2(\omega), \ldots, \psi_n(\omega)$, respectively. The characteristic function of $Y = X_1 + X_2 + \cdots + X_n$ is given by the product

$$\psi_Y(\omega) = \psi_1(\omega)\psi_2(\omega)\ldots\psi_n(\omega) \tag{12.17}$$

*Proof:* From the definition (8.19) of the CF, we have:

$$\begin{aligned} \psi_Y(\omega) &= E(e^{-j\omega Y}) \\ &= E(e^{-j\omega(X_1+X_2+\cdots+X_n)}) \\ &= E(e^{-j\omega X_1}e^{-j\omega X_2}\ldots e^{-j\omega X_n}) \end{aligned} \tag{12.18}$$

Making use of Theorem 11.2, we obtain

$$\begin{aligned} \psi_Y(\omega) &= E(e^{-j\omega X_1})E(e^{-j\omega X_2})\ldots E(e^{-j\omega X_n}) \\ &= \psi_1(\omega)\psi_2(\omega)\ldots\psi_n(\omega) \quad \square \end{aligned} \tag{12.19}$$

*Interpretation:*

- Recall that the PDF of $Y$, say $f_Y(y)$, can be obtained as the inverse Fourier transform of $\psi_Y(\omega)$. Similarly, the PDF of $X_i$ $(i = 1, \ldots, n)$, say $f_i(x)$, can be obtained as the inverse Fourier transform of $\psi_i(\omega)$.
- Therefore, applying the inverse Fourier transform operator on both sides of (12.17), we obtain

  $$f_Y = f_1 * f_2 * \cdots * f_n \tag{12.20}$$

  where $*$ denotes the convolution.

*Example 12.3:*

► Suppose RVs $X_1, X_2, \ldots, X_n$ are independent and identically distributed with common marginal PDF

$$f(x) = \begin{cases} 1, & |x| < 1/2 \\ 0, & \text{otherwise.} \end{cases}$$

Sketch the PDF of the sum $Y = X_1 + X_2 + \cdots + X_n$ for $n = 1, 2, 4$.

*Solution:* ◄

*Theorem 12.7:* Let $X_i$ $(i = 1, 2, ...)$ be a sequence of independent, identically distributed RVs with mean $\mu = E(X_i)$ and variance $\sigma^2 = Var(X_i) < \infty$. Define

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \tag{12.21}$$

In the limit $n \to \infty$, the distribution of $Z_n$ tends to the standard normal:

$$\lim_{n\to\infty} P(Z_n \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt \tag{12.22}$$

That is, $Z_n \to N(0, 1)$.

*Remarks:*

- Theorem 12.7 is called the central limit theorem. Its proof makes use of Lemma 12.6 but the technical details are beyond the scope of this course.
- We often express (12.22) in the more compact form $Z_n \to N(0, 1)$.
- Many phenomena occuring in nature or in man-made systems are the results of averaging a large number of independent contributions:
  - Thermal noise in radio systems.
  - Measurement/observation errors.
- In such cases, Theorem 10.6 motivates the use of the normal density.

*Example 12.4: Normal approximation to the Binomial*

► Let $Y$ be a binomial RV with parameters $N$ and $p$, that is $Y \sim B(n,p)$. Recall that such a binomial RV can be expressed as a sum of independent Bernouilli RVs $X_i$, that is

$$Y = X_1 + \cdots + X_n$$

where

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases}$$

with mean $\mu = E(X_i) = p$ and variance $\sigma^2 = Var(X_i) = pq$. Define

$$\begin{aligned} Z_n &= \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \\ &= \frac{\frac{1}{n}Y - p}{\sqrt{pq}/\sqrt{n}} \\ &= \frac{Y - np}{\sqrt{npq}} \end{aligned}$$

According to the central limit, we expect that for $n$ sufficiently large:

$$P(Z_n \leq z) = P(\frac{Y-np}{\sqrt{npq}} \leq z) \approx \Phi(z) \tag{12.23}$$

where $\Phi(z)$ denote the CDF of the standard normal. Equivalently, we may express (12.23) in the form

$$P(Y \leq y) \approx \Phi(\frac{y-np}{\sqrt{npq}}) \tag{12.24}$$

(12.24) is often referred to as the DeMoivre-Laplace approximation. ◄

# Chapter 13

# Introduction to Random Processes

In several areas of sciences and engineering, we encounter signals (i.e. function of time) which exhibit a random behavior. That is, no deterministic model can be used to predict the time evolution of these signals in advance of their observation.

A typical example is provided by the digital information signals used in modern telecommunications systems. Another example is the evolution of the value of a company's share on the stock market.

While such signals cannot be predicted exactly in advance of their observation, they usually exhibit regularity in their behavior that can often be exploited in the design of an engineering system or in the prediction of trends.

In this chapter, we present an introduction to such random signals, also known as stochastic processes. We cover the following topics:

- Basic definition and general concepts;
- Stationary processes and filtering thereof;
- Poisson points processes.

## 13.1 Terminology

*Definition:* Consider a probability space $(\mathcal{S}, \mathcal{F}, P)$. A random process is a family of random variables, say $\{X(t), t \in T\}$, defined on $\mathcal{S}$ and indexed by a parameter $t$ taken from a set $T \subseteq \mathbb{R}$.

*Remarks:*

- Recal that a random variable is a mapping from $S$ into $\mathbb{R}$. Thus, a random process may be viewed as a function of two arguments:

$$(t, s) \in T \times S \rightarrow X(t, s) \in \mathbb{R} \tag{13.1}$$

- For a fixed value of $t = t_o$, $X(t_o)$ is simply a random variable as previously defined in Chapter 5:

$$s \in S \rightarrow X(t_o, s) \in \mathbb{R} \tag{13.2}$$

- For a fixed value of $s = s_o$, $X(t)$ defines a real-valued function of the variable $t$:

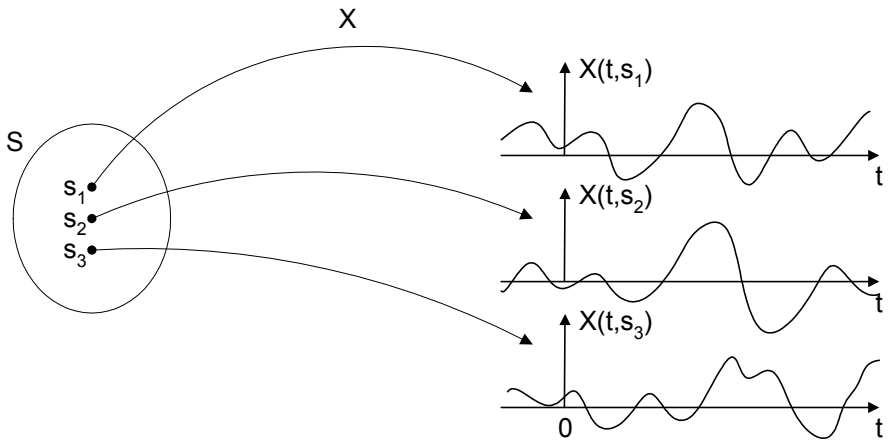
$$t \in T \rightarrow X(t, s_o) \tag{13.3}$$

  The latter is called a *sample function* or *realization* of the process.

- This leads to an alternative interpretation of a random process as a mapping from the sample space $\mathcal{S}$ into the set of all possible sample functions, $\{X(., s), : s \in \mathcal{S}\}$, also called *ensemble*.

- This is illustrated in the figure below:




- It is a common practice in probability to omit the argument $s$ from $X(t, s)$. We shall usually do so, unless it is desired to emphasize certain aspects of the theory.

*Continuous versus discrete-time processes:* In many applications of random processes, indexing parameter $t$ as the signification of a time argument. Accordingly, we distinguish two basic types of random processes:

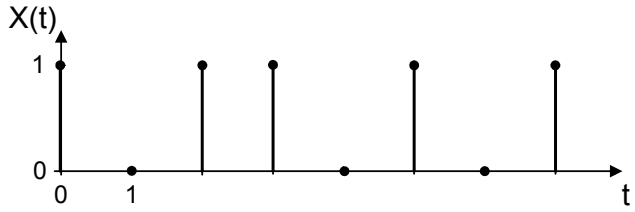
- Discrete-time: if the index set $T$ is finite or countably infinite.
- Continuous-time: if the index set $T$ is uncountably infinite.

*Continuous versus discrete-state processes:* The value taken by $X(t)$ at any given time $t$ is called the *state* of the process, and the set of all such possible values is called the *state space*. We classify random processes as follow:

- Discrete-state: if the state space is discrete.
- Continuous-state: if the state space is continuous.

*Example 13.1:*

► A simple example of a random process is provided by a sequence of binary digits (bits) at the input of a digital communications system. The value of each bit is represented by a binary random variable $X(t)$, where $t$ denotes a discrete-time index within the index set $T = \{0, 1, 2, \ldots\}$. The exact relationship between index $t$ and the physical time depends on the bit rate. At any given time $t \in T$, RV $X(t)$ may take on two possible values, say 0 or 1, with probabilities $P(X(t) = 1) = p$ and $P(X(t) = 0) = 1 - p = q$. The process $X_t$ is therefore a discrete-time, discrete-state process. An example of a realization of $X(t)$ is illustrated below (the corresponding bit sequence is $10110101\ldots$): ◄




*Example 13.2:*

► Consider the random process defined by

$$X(t) = A\cos(\omega t + \phi), \quad t \in \mathbb{R} \tag{13.4}$$

where $A$ is a random amplitude uniformly distributed within the range $[-5, +5]$ volts, $\phi$ is a random phase uniformly distributed in the range $[-\pi, \pi]$, $\omega$ denotes a physical angular frequency in Hertz, and the parameter $t$ denotes time in seconds.

This is an example of a continuous-time, continuous-state process. The state space is given by the interval $[-5, +5]$ volts and the index set is $T = \mathbb{R}$. Realizations of $X(t)$ are illustrated below:

◄

## 13.2 Characterization of random processes

In the solution of problems involving random processes, it is important to characterize the latter so that relevant probabilities and/or moments can be computed. Several types of characterization exist; they vary in their level of refinement.

*Definition:* The $n$th-order CDF of random process $X(t)$ is a function of $2n$ arguments defined as:

$$F_X(x_1, \ldots, x_n; t_1, \ldots, t_n) = P(X(t_1) \leq x_1, \ldots, X(t_n) \leq x_n) \tag{13.5}$$

where $x_i \in \mathbb{R}$ and $t_i \in T$, for $i = 1, \ldots, n$.

*Complete Characterization:* We say that we have a complete characterization of the process $X(t)$ if the $n$th order CDF (13.5) is known for all positive integers $n \in \mathbb{N}$.

*Remark:*

- In general, it is not possible to obtain such a complete characterization for an arbitrary process $X(t)$.
- Often, we must content ourselves with so-called partial characterizations, such as the second-moment characterization introduced below.

*Definition:* The mean, autocorrelation and autocovariance functions of process $X(t)$ are respectively defined as

$$\mu_X(t) \triangleq E[X(t)] \tag{13.6}$$
$$R_X(t,u) \triangleq E[X(t)X(u)] \tag{13.7}$$
$$K_X(t,u) \triangleq Cov(X(t),X(u)) \tag{13.8}$$

*Properties:* The following properties of the autocorrelation function follow from its definition:

$$R_X(t,u) = R_X(u,t) \tag{13.9}$$
$$R_X(t,t) = E[X(t)^2] \tag{13.10}$$

For the autocovariance, we have:

$$K_X(t,u) = K_X(u,t) \tag{13.11}$$
$$K_X(t,u) = R_X(t,u) - \mu_X(t)\mu_X(u) \tag{13.12}$$
$$K_X(t,t) = Var(X(t)) \triangleq \sigma_X^2(t) \tag{13.13}$$
$$|K_X(t,u)| \leq \sigma_X(t)\sigma_X(u) \tag{13.14}$$

*Second-moment characterization:* Knowledge of the mean function $\mu_X(t)$ and the autocorrelation function $R_X(t,u)$ for all possible values of $t$ and $u$ in the index set $T$ provides a second moment characterization of the process $X(t)$.

*Remarks:* In many applications, a second-moment characterization is adequate to answer most questions of practical interest. For certain types of process, like the Gaussian process, it is actually possible to derive a complete characterization from the 2nd-moment one. These facts motivate the use of the second-moment characterization.

*Example 13.3:*

▶ Consider the binary random process $X(t)$ in example 13.1. Find the mean function $\mu_X(t)$. Assuming that each bit in the process $X(t)$ is independently generated from the others, find the autocorrelation and autocovariance functions of $X(t)$.

*Solution:* Here, $X(t) = 1$ with probability $p$ and $X(t) = 0$ with probability $q = 1 - p$. The mean function is obtained as

$$\mu_X(t) = E[X(t)] = 1 \cdot p + 0 \cdot q = p$$

When evaluating the autocorrelation function, we distinguish 2 cases: If $t = u$, we have

$$R_X(t,t) = E[X(t)^2] = 1^2 \cdot p + 0^2 \cdot q = p$$

If $t \neq u$, $X(t)$ and $X(u)$ are independent, so that

$$R_X(t,u) = E[X(t)X(u)] = E[X(t)]E[X(u)] = p^2$$

For the autocovariance function, we have

$$\begin{aligned} K_X(t,u) &= R_X(t,u) - \mu_X(t)\mu_X(u) \\ &= \begin{cases} p\,q & \text{if } t = u \\ 0 & \text{if } t \neq u \end{cases} \end{aligned} \tag{13.15}$$

◀

*Example 13.4:*

▶ Consider the process

$$X(t) = A\cos(2\pi f t), \quad t \in \mathbb{R} \tag{13.16}$$

where $A$ is a normally distributed random amplitude with zero-mean and standard deviation 1 volt, $f$ denotes a fixed frequency in Hertz, and $t$ denotes time in seconds. Find $\mu_X(t)$, $R_X(t,u)$ and $K_X(t,u)$. ◀

## 13.3 Wide sense stationary processes

*Scope:* In this section, it is assumed that $X(t)$ is a continuous-time process with index set $T = \mathbb{R}$.

*Definition:* Process $X(t)$ is said to be *strict-sense stationary* (SSS) if for all integer $n$, its $n$th-order CDF is unaffected by a shift of the time origin. That is, for all $n \in \mathbb{N}$, and for all $x_i$ and $t_i$ in $\mathbb{R}$, we have

$$F_X(x_1, \ldots, x_n; t_1 + \tau, \ldots, t_n + \tau) = F_X(x_1, \ldots, x_n; t_1, \ldots, t_n) \qquad (13.17)$$

for any possible value of the time shift $\tau$.

*Definition:* Process $X(t)$ is said to be *wide-sense stationary* (WSS) if

(a) the mean function $\mu_X(t)$ is constant, that is: $\mu_X(t) \equiv \mu_X$

(b) the autocorrelation function $R_X(t, u)$ is only a function of $t - u$:

$$R_X(t, u) \equiv R_X(t - u) \qquad (13.18)$$

*Remarks:*

- We often refer to the difference $\tau = t - u$ as the lag and use the notation $R_X(\tau)$ for the autocorrelation function.
- SSS implies WSS but not vice versa, except for certain special types of processes. WSS is easier to deal with than SSS.

*Properties:* The following properties of the autocorrelation function can be demonstrated:

$$\begin{aligned} R_X(\tau) &= E[X(t)X(t-\tau)], \quad \text{for any } t \in \mathbb{R} && (13.19)\\ R_X(0) &= E[X(t)^2] \geq 0 && (13.20)\\ R_X(\tau) &= R_X(-\tau) && (13.21)\\ |R_X(\tau)| &\leq R_X(0) && (13.22) \end{aligned}$$

*Remark:* Think of $X(t)$ as a voltage signal applied across a 1 ohm resistor. Then $X(t)^2$ represents the instantaneous power dissipated through the resistor at time $t$. According to (13.20), the expected value of the instantaneous power is constant over time (due to WSS assumption) and equal to $R_X(0)$.

*Example 13.5:*

► Consider the process

$$X(t) = A\cos(\omega t) + B\sin(\omega t), \quad t \in \mathbb{R}$$

where $A$ and $B$ are independent random variables with zero-mean and common variance $\sigma^2$. Show that $X(t)$ is WSS.

*Solution:* ◄

*Definition:* The power spectral density (PSD) of process $X(t)$, denoted $S_X(\omega)$, is defined as the Fourier transform of its autocorrelation function $R_X(\tau)$:

$$S_X(\omega) = \int_{-\infty}^{\infty} R_X(\tau) e^{-j\omega\tau} d\tau, \quad \omega \in \mathbb{R} \tag{13.23}$$

*Remarks:*

- Clearly, if $S_X(\omega)$ is know, then $R_X(\tau)$ can be recovered by applying the inverse Fourier transform:

$$R_X(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(\omega) e^{j\omega\tau} d\omega, \quad \tau \in \mathbb{R} \tag{13.24}$$

- Equations (13.24)-(13.25) are often referred to as the Wiener-Khinchin relations.

*Properties:* $S_X(\omega)$ satisfies the following basic properties:

(a) $S_X(\omega)$ is real and non-negative, that is: $S_X(\omega) \geq 0$

(b) $S_X(\omega)$ is an even function of $\omega$: $S_X(\omega) = S_X(-\omega)$

(c) The average instantaneous power in $X(t)$ can be obtained as

$$R_X(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(\omega) d\omega \tag{13.25}$$

*Remarks:*

- The power spectral density $S_X(\omega)$ derives its name from (13.25), where it is seen that the total average instantaneous power can be obtained by integrating $S_X(\omega)$ over all frequencies.
- Accordingly, $S(\omega)$ has the units of power per Hertz and the product $S_X(\omega)d\omega/2\pi$ represents the amount of power in a frequency band of width $d\omega$.
- The PSD is particularly useful as it makes it possible to study WSS processes directly in the frequency domain.

*Example 13.6:*

▶ Consider the process

$$X(t) = A\cos(\omega_0 t) + B\sin(\omega_0 t)$$

where $A$ and $B$ are uncorrelated RVs with zero-mean and common variance $\sigma^2$, and $\omega_0$ is a fixed angular frequency. Find the PSD of $X(t)$.

*Solution:*

◀

*Definition:* We say that $W(t)$ is a white noise process if it is WSS with zero-mean, i.e. $\mu_W = 0$, and autocorrelation function

$$R_W(\tau) = N\delta(\tau) \tag{13.26}$$

where $N$ is a positive constant and $\delta(\tau)$ is the (continuous-time) unit impulse function.

*Remarks:*

- Taking the Fourier transform of (13.26), we obtain

  $$S_W(\omega) = N \tag{13.27}$$

  which reveals that a white noise process has a constant PSD for all frequency values (hence the name).

- In communications engineering and statistical physics, white noise is often used to model random signal fluctuations whose frequency content extends beyond that of the signal of interest (e.g. thermal noise).

*Example 13.7:*

▶ Find the PSD of the signal

$$Y(t) = X(t) + W(t)$$

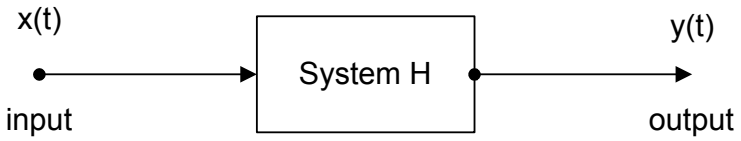
where $X(t)$ is defined as in example 13.6 and $W(t)$ is a white noise with constant PSD $N$. Assume that RVs $A$, $B$ and $W(t)$ (for any $t$) are mutually independent.
◀

## 13.4 Filtering of WSS processes

*Recap on LTI systems:*

- A system is broadly defined as a device or physical process that transforms a time-domain signal applied to its input, say $x(t)$, into a corresponding output signal say $y(t)$:




- A system $\mathcal{H}$ is represented mathematically as a mapping between a set of input signals and a set of output signals. Accordingly,

$$y(t) = \mathcal{H}\{x(t)\} \tag{13.28}$$

  For now, let us assume that the signals of interest are deterministic real-valued, signals defined for all $t \in \mathbb{R}$ (i.e. continuous-time).

- We say that system $\mathcal{H}$ is linear iff, for any numbers $a_1$ and $a_2$ and any input signals $x_1(t)$ and $x_2(t)$, we have

$$\mathcal{H}\{a_1 x_1(t) + a_2 x_2(t)\} = a_1 y_1(t) + a_2 y_2(t) \tag{13.29}$$

- We say that system $\mathcal{H}$ is time-invariant iff, for any $t_0 \in \mathbb{R}$, we have

$$\mathcal{H}\{x(t - t_0)\} = y(t - t_0) \tag{13.30}$$

- A system $\mathcal{H}$ is said to be LTI if it is both linear and time-invariant.

- If $\mathcal{H}$ is LTI, it is possible to show that (13.28) reduces to

$$y(t) = \int_{-\infty}^{\infty} h(u)x(t-u)du \triangleq h(t) * x(u), \tag{13.31}$$

  also known as a convolution integral.

- In (13.31), $h(t)$ is the so-called impulse response of the system, i.e.:

$$h(t) = \mathcal{H}\{\delta(t)\} \tag{13.32}$$

  The Fourier transform of $h(t)$, that is

$$H(\omega) = \int_{-\infty}^{\infty} h(t)e^{-j\omega t}dt, \tag{13.33}$$

  is known as the frequency response of the system.

- Here, we shall assume that the systems under consideration have absolutely integrable impulse responses (i.e. stable systems), so that $H(\omega)$ in (13.33) is well defined. Furthermore, to simplify the discussion, we assume that $h(t) \in \mathbb{R}$.

*Discussion:*

- In introductory courses on signals and systems, the class of inputs is usually restricted to deterministic signals.

- Here, we extend the above concepts and consider systems that operate on random signals. We focus on WSS signals, for which the concept of frequency is particularly meaningful.

*Problem formulation:*

- Let $X(t)$ denote a WSS process applied to the input of an LTI system with impulse respone $h(t)$. Let $Y(t)$ denote the corresponding output:

$$Y(t) = \int_{-\infty}^{\infty} h(u)X(t-u)du \tag{13.34}$$

  Note that because $X(t)$ is a random process, so is the output $Y(t)$.

- We are interested in determining the effects of filtering WSS process $X(t)$ with LTI filter $h(t)$. More precisely, we seek to develop a second-moment characterization for the output process $Y(t)$.

- In these developments, the key is to realize that the expectation and integration operators commute, that is $E[\int \ldots dt] = \int E[\ldots]dt$.

*Theorem 13.1:* The mean value of $Y(t)$ is constant and is given by

$$\mu_Y(t) = \mu_X H(0) \tag{13.35}$$

*Proof:* First note that since $X(t)$ is WSS, we have $E[X(t-u)] = \mu_X$. Then:

$$\begin{aligned}
\mu_Y(t) &= E[Y(t)] \\
&= E[\int_{-\infty}^{\infty} h(u)X(t-u)du] \\
&= \int_{-\infty}^{\infty} h(u)E[X(t-u)]du \\
&= \mu_X \int_{-\infty}^{\infty} h(u)du = \mu_X H(\omega)|_{\omega=0} \quad \square
\end{aligned}$$

*Theorem 13.2:* The autocorrelation function of $Y(t)$ is given by

$$R_Y(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u_1)h(u_2)R_X(t_1 - t_2 - u_1 + u_2)du_1du_2 \qquad (13.36)$$

*Proof:* Left as an exercise to the reader.

*Main conclusion:*

- Note from (13.25) that $\mu_Y(t)$ is a constant. Also note from (13.36) that $R_Y(t_1, t_2)$ is actually a function of $\tau = t_1 - t_2$.
- This shows that $Y(t)$ is WSS. In other words, if a WSS process is passed through an LTI system, the resulting output process is also WSS.

*Theorem 13.3:* The power spectral density of $Y(t)$ is given by

$$S_Y(\omega) = |H(\omega)|^2 S_X(\omega) \qquad (13.37)$$

*Proof:* Starting from (13.36), we have

$$\begin{aligned} R_Y(\tau) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u_1)h(u_2)R_X(\tau - u_1 + u_2)du_1du_2 \\ &= h(\tau) * \int_{-\infty}^{\infty} h(u_2)R_X(\tau + u_2)du_2 \\ &= h(\tau) * \int_{-\infty}^{\infty} h(-u_2)R_X(\tau - u_2)du_2 \\ &= h(\tau) * h(-\tau) * R_X(\tau) \end{aligned}$$

Finally, taking the Fourier transform on both sides and noting that here, $h(t) \in \mathbb{R}$, we obtain:

$$S_Y(\omega) = H(\omega)H^*(\omega)S_X(\omega) \quad \square$$

*Example 13.8:*

► The random process $X(t)$ in Example 13.6 is passed through and LTI system with square magnitude response

$$H(\omega) = \frac{1}{\alpha^2 + \omega^2}$$

Find the PSD and autocorrelation function of the output process $Y(t)$.

*Solution:* Applying (13.37), we have

$$\begin{aligned} S_Y(\omega) &= |H(\omega)|^2 S_X(\omega) \\ &= \frac{1}{\alpha^2 + \omega^2}\, \sigma^2 \pi[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)] \\ &= \frac{\sigma^2}{\alpha^2 + \omega_0^2}\, \pi[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)] \end{aligned}$$

Taking the inverse Fourier transform

$$R_Y(\omega) = \frac{\sigma^2}{\alpha^2 + \omega_0^2}\, \cos(\omega_0 \tau)$$

◄

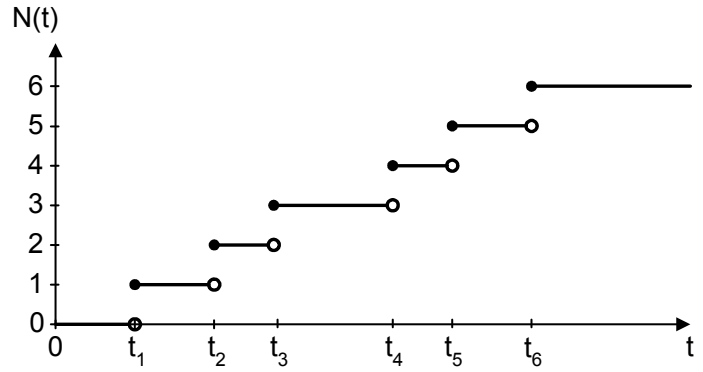
## 13.5 Poisson processes

*Counting process:*

- Consider a random experiment that takes place over the time interval $T = [0, \infty)$, and assume that at any given time $t > 0$, a certain event may or not occur.
- Let $N(t)$ denote the number of occurrences of this event over the time interval $(0, t]$. We refer to $N(t)$ as a *counting process*.

*Remarks:*

- By definition, $N(t)$ is a non-decreasing function of time $t$. A typical realization of a counting process $N(t)$ is illustrated below, where $t_i$, $(i = 1, 2, \ldots)$ denote the time of occurence of the $i$th event:




- Note that for each $t > 0$, $N(t)$ is a RV with set of possible values $\{0, 1, 2, ....\}$. In practice it is of interest to characterize the PMF of $N(t)$, i.e. $P(N(t) = n)$ for $n = 0, 1, 2, ....$

- Below, we develop such a characterization for a special type of process called a Poisson process.

*Definition:* A counting process $N(t)$ is called a Poisson process with rate $\lambda$ if the following three basic properties are satisfied:

(a) *Stationarity*: the PMF of the number of events in a given time interval depends only on the length of this interval and not its location.

(b) *Independent increments*: the number of events that occur in disjoint time intervals are independent.

(c) *Orderliness*: for $t$ small, $P(N(t) = 1) \approx \lambda t$ and $P(N(t) \geq 2) \approx 0$.

*Remarks:*

- These basic assumptions are often satisfied in practice.
- As a result, there are numerous examples of Poisson process in science and engineering, including:
  - number of alpha particles emitted by a radio-active substance
  - number of earthquakes in a certain geographical area of a country
  - number of requests for connections in a communication network
  - number of binary packets received at a swithching node of a communications network

  all measured from some time 0 up to $t$.

*Theorem 13.4:* Let $N(t)$ be a Poisson process with rate $\lambda > 0$, and suppose that $N(0) = 0$. For any value of $t > 0$, the PMF of $N(t)$ is given by

$$P(N(t) = n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \quad n = 0, 1, 2, ... \tag{13.38}$$

*Remarks:*

- A formal proof of this result is beyond the scope of this course.
- The theorem essentially says that for any given time $t > 0$, the RV $N(t)$ is Poisson with parameter $\lambda t$.
- Accordingly, the expected value of $N(t)$ is given by

  $$E[N(t)] = \lambda t \tag{13.39}$$

  that is, the expected value of the count increases linearly with time at the rate $\lambda$.

*Example 13.9:*

► Suppose that in a certain geographical area, earthquakes occur at a rate of 7 per year, in accordance with a Poisson process.

(a) What is the probability of no earthquakes in one year?

(b) What is the probability that in exactly three of the next 8 years, no earthquake will occur?

*Solution:* Let $N(t)$ denote the number of earthquakes from time 0 up to time $t$, inclusively. For convenience, assume that the unit of time is the year. Then, $N(t)$ is a Poisson process with rate $\lambda = 7$:

$$P(N(t) = n) = \frac{(7t)^n e^{-7t}}{n!}, \quad n = 0, 1, 2, ...$$

(a) We seek

$$P(N(1) = 0) = e^{-7} \approx 9.1 \cdot 10^{-4}$$

(b) Because of the stationarity assumption with Poisson process:

$$P(\text{no earthquake in one year}) = P(N(1) = 0)$$

regardless of the specific one year period being considered. Because of the independence assumption, the number of earthquakes in consecutive years are independent random variables.

Thus, each of the 8 consecutive years may be viewed as an independent Bernouilli trial, where a success is defined as 0 earthquake with probability

$$p = P(N(1) = 0) = 9.1 \cdot 10^{-4}$$

Let $X$ be the number of years, over the next 8 years, with no earthquakes. It follows that $X$ is binomial with parameters $p$ and $n = 8$. Therefore

$$P(X = 3) = \binom{8}{3} p^3 (1 - p)^5 \approx 4.2 \cdot 20^{-8}$$

◄

### 13.5.1 Property of interarrival time

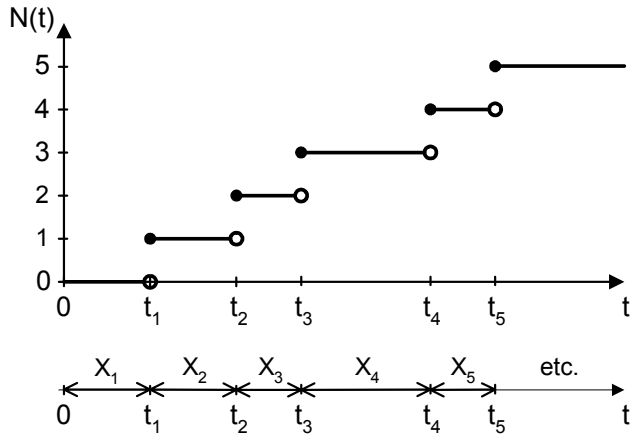
*Definition:* Consider a counting process $N(t)$ and let $t_i$ $(i = 1, 2, ...)$ denote the time of occurrence of the $i$th event (for convenience, set $t_0 = 0$). The continuous RVs

$$X_i \;=\; t_i - t_{i-1} \quad \text{for} \quad i \in \mathbb{N} \tag{13.40}$$

are called interarrival times.

*Remarks:*

- This situation is illustrated below:




- It is often of interest to characterize the distributions of the RV $X_i$. For the Poisson process, this is relatively straightforward.

*Theorem 13.5:* The interarrival times $X_i$ $(i = 1, 2, ...)$ are exponential RVs with parameter $\lambda$.

*Proof:* First consider $X_1$. Let $F_1(x)$ denote its CDF. We show below that $F_1(x)$ has the general form (7.80):

- Observe that for $x > 0$, the two events $X_1 > x$ and $N(x) = 0$ are identical. Thus, for $x > 0$, we have

$$P(X_1 > x) = P(N(x) = 0) = e^{-\lambda x}, \tag{13.41}$$

  which implies

$$F(x) = P(X_1 \leq x) = 1 - P(X_1 > x) = 1 - e^{-\lambda x} \tag{13.42}$$

- By definition, $X_1 = t_1 > 0$. Thus, for $x \leq 0$, we have

$$F_1(x) = P(X_1 \leq x) = 0 \tag{13.43}$$

This approach may be generalized to show that the other RVs $X_i$ $(i > 1)$ are also exponential with parameter $\lambda$. □

*Remarks:*

- As a result of this connection with the Poisson process, the exponential RV is extremely important.
- Examples of exponential RV include:
  - time between phone calls,
  - interarrival time of binary packets at a routing node in a network,
  - time interval between earthquakes.

*Example 13.10:*

► Suppose that on average, a 128-bit packet arrives at a switching node of a digital communication network every $10\mu$s. Assuming that these arrivals can be modeled as a Poisson process, find the probability that the next packet arrives within $5\mu$s.

*Solution:* Let $X$ denote the arrival time of the next packet, measured relative to the arrival time of the most recent packet. Since the arrivals are Poisson, it follows that $X$ is exponential with parameter $\lambda$. To find $\lambda$, note that

$$E(X) = \frac{1}{\lambda} = 10\mu\text{s} \quad \Longrightarrow \quad \lambda = 0.1\mu\text{s}^{-1}$$

Finally, we seek

$$\begin{aligned} P(X \leq 5\mu\text{s}) &= F(5\mu\text{s}) \\ &= 1 - e^{-\lambda \cdot 5\mu\text{s}} \\ &= 1 - e^{-1/2} \approx 0.39 \end{aligned}$$

◄