Ministry of Higher Education and Scientific Research,
Mouloud Mammeri University, Tizi-Ouzou,
Faculty of Sciences,
Department of Mathematics.

Lecture notes

TRACK : MATHEMATICS

Second year Bachelor's Degree in Mathematics

Author : ATIL Lynda

Introduction to probability

(Course & exercises)

Preface

Dear Second-Year Mathematics Students,
Welcome to this module on Probability and Statistics, a cornerstone of modern mathematics with profound applications across science, engineering, finance, and data analysis. This field provides the essential language and tools for modelling uncertainty, analysing data, and making informed predictions in the face of randomness. For a mathematician, it represents a fascinating world where abstract theory meets tangible, real-world phenomena.

This manual has been carefully structured to guide you from fundamental concepts to more advanced topics, ensuring a solid and progressive understanding. The journey begins with Chapter 1, which serves as a crucial recap of fundamental probability notions. Here, we will revisit the axiomatic foundations of probability, combinatorial calculations, conditional probability, and Bayes' Theorem. This chapter is designed to solidify your intuition and ensure all students have a common, robust starting point. Numerous examples and solved exercises will help you master these essential tools.

In Chapter 2, we will move from abstract events to more manageable mathematical objects by introducing Random Variables. We will explore the key distinctions between discrete and continuous random variables, focusing on their characterization through probability mass functions, probability density functions, and cumulative distribution functions. This chapter will also introduce you to the important concepts of expectation and variance, which capture the "average" behaviour and the spread of a random variable, respectively.

Chapter 3 builds directly upon this foundation by delving into the most Common Probability Distributions. You will become acquainted with the essential families of discrete distributions (such as the Binomial, Poisson, and Geometric) and continuous distributions (such as the Uniform, Normal, and Exponential). A significant part of this chapter is dedicated to approximations, particularly the use of the Poisson distribution to approximate the Binomial, a powerful technique for simplifying complex calculations. Furthermore, we will explore the transformation of random variables, a key concept for understanding how distributions change under functional mappings, which is vital for more advanced statistical inference.

A distinctive feature of this manual is its strong emphasis on application and practice. Each theoretical concept is illustrated with concrete examples, and every chapter concludes with a set of carefully chosen solved exercises. We strongly encourage you to not simply read these solutions, but to attempt the problems independently first. The process of struggling with a problem and then understanding the detailed solution is where the deepest learning occurs.

Our goal is not only to equip you with the technical skills to solve problems but to also

foster a deeper appreciation for the power and elegance of probabilistic reasoning. We wish you success and an intellectually stimulating experience as you embark on this journey into the world of Probability and Statistics.

Contents

CHAPTER I. Preliminaries in Probability

1 Introduction

In this first chapter, we begin by axiomatically defining the notion of probability over a coherent set events (or σ-algebra). The idea of conditional probability then follows very simply. Among other things, it is linked to the notion of independence, which is fundamental in both probability and statistics.

1.1 What is probability?

In most branches of knowledge, experiments are a way of life. In probability and statistics, too, we concern ourselves with special types of experiments.
Randomness and uncertainty exist in our daily lives as well as in every discipline in science, engineering, and technology. Probability theory is a mathematical framework that allows us to describe and analyze random phenomena in the world around us. By random phenomena, we mean events or experiments whose outcomes we can't predict with certainty.
Let's consider a couple of specific applications of probability in order to get some intuition. First, let's think more carefully about what we mean by the terms "randomness" and "probability" in the context of one of the simplest possible random experiments: flipping a fair coin.
One way of thinking about "randomness" is that it's a way of expressing what we don't know. Perhaps if we knew more about the force I flipped the coin with, the initial orientation of the coin, the impact point between my finger and the coin, the turbulence in the air, the surface smoothness of the table, the coin lands on, the material characteristics of the coin and the table, and so on, we would be able to definitively say whether the coin would come up heads or tails. However, in the absence of all that information, we cannot predict the outcome of the coin flip. When we say that something is random, we are saying that our knowledge about the outcome is limited, so we can't be certain what will happen.
Since the coin is fair, if we don't know anything about how it was flipped, the probability that it will come up heads is 50%, or $\frac{1}{2}$. What exactly do we mean by this? There are two common interpretations of the word "probability." One is in terms of relative frequency. In other words, if we flip the coin a very large number of times, it will come up heads about $\frac{1}{2}$ of the time. As the number of coin flips increases, the proportion that come up heads will tend to get closer to $\frac{1}{2}$.

A second interpretation of probability is that it is a quantification of our degree of subjective personal belief that something will happen. To get a sense of what we mean by this, it may be helpful to consider a second example: predicting the weather. When we think about the chances that it will rain today, we consider things like whether there are clouds in the sky and the humidity. However, the beliefs that we form based on these

factors may vary from person to person - different people may make different estimates of the probability that it will rain. Often these two interpretations of probability coincide - for instance, we may base our personal beliefs about the chance that it will rain on an assessment of the relative frequency of rain on days with conditions like today.

The beauty of probability theory is that it is applicable regardless of the interpretation of probability that we use (i.e., in terms of long-run frequency or degree of belief). Probability theory provides a solid framework for studying random phenomena. It starts by assuming axioms of probability, and then builds the entire theory using mathematical arguments.

Before delving into studying probability theory, let us briefly look at the following examples.

Example 1.1.
A coin is tossed. Assuming that the coin does not land on the side, there are two possible outcomes of the experiment: heads and tails. On any performance of this experiment one does not know what the outcome will be. The coin can be tossed as many times as desired.

Example 1.2.
A roulette wheel is a circular disk divided into 38 equal sectors numbered from 0 to 36 and 00. A ball is rolled on the edge of the wheel, and the wheel is rolled in the opposite direction. One bets on any of the 38 numbers or some combinations of them. One can also bet on a color, red or black. If the ball lands in the sector numbered 32, say, anybody who bet on 32 or combinations including 32 wins, and so on. In this experiment, all possible outcomes are known in advance, namely 00, 0, 1, 2,. . ., 36, but on any performance of the experiment there is uncertainty as to what the outcome will be, provided, of course, that the wheel is not rigged in any manner. Clearly, the wheel can be spun any number of times.

1.2 Review of Set Theory

Probability theory uses the language of sets. As we will see later, probability is defined and calculated for sets. Thus, here we briefly review some basic concepts from set theory. We discuss set notations, definitions, and operations (such as intersections and unions). We then introduce countable and uncountable sets. Finally, we briefly discuss functions.

1.2.1 Some definitions

Definition 1.1.
A set is a collection of distinct objects called elements. We often use capital letters to denote a set. To define a set we can simply list all the elements in curly brackets.

Example 1.3.
We define a set A that consists of the two elements $\diamondsuit$ and $\clubsuit$, we write $A = \{\diamondsuit, \clubsuit\}$. To say that $\diamondsuit$ belongs to A, we write $\diamondsuit \in A$, where $\in$ is pronounced "belongs to". To say that an element does not belong to a set, we use $\notin$. For example, we may write $\heartsuit \notin A$.

Remark 1.1.
Note that ordering does not matter, so the two sets $\{\diamondsuit, \clubsuit\}$ and $\{\clubsuit, \diamondsuit\}$ are equal. We often work with sets of numbers. Some important sets are given in the following example.

Example 1.4.
Consider the following usual sets;

- The set of natural numbers, $\mathbb{N} = \{0, 1, 2, 3, \ldots, \}$
- The set of integers, $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$.
- The set of rational numbers $\mathbb{Q}$.
- The set of real numbers $\mathbb{R}$.
- Closed intervals on the real line. For example, $[2, 3]$ is the set of all real numbers x such that $2 \leq x \leq 3$.
- Open intervals on the real line. For example $(-1, 3)$ is the set of all real numbers x such that $-1 < x < 3$.
- Similarly, $[1, 2)$ is the set of all real numbers x such that $1 \leq x < 2$.
- The set of complex numbers $\mathbb{C}$ is the set of numbers in the form of $a + bi$, where $a, b \in \mathbb{R}$, and $i = \sqrt{-1}$.

Definition 1.2.
Set A is a subset of set B if every element of A is also an element of B. We write $A \subset B$, where " $\subset$ " indicates "subset". Equivalently, we say B is a superset of A, or $B \supset A$.

Remark 1.2.
Two sets are equal if they have the exact same elements. Thus, $A = B$ if and only if $A \subset B$ and $B \subset A$.
The set with no elements, i.e., $\emptyset = \{\}$ is the null set or the empty set. For any set A, $\emptyset \subset A$.

Definition 1.3.
The universal set is the set of all things that we could possibly consider in the context we are studying. Thus every set A is a subset of the universal set. We often denote the universal set by Ω (As we will see, in the language of probability theory, the universal set is called the sample space.)

Example 1.5.
if we are discussing rolling of a die, our universal set may be defined as $\Omega = \{1, 2, 3, 4, 5, 6\}$, or if we are discussing tossing of a coin once, our universal set might be $\Omega = \{H, T\}$ (H for heads and T for tails).

1.2.2 Set operations

Property 1.1.
The union of two sets is a set containing all elements that are in A or in B (possibly both). For example, $\{1, 2\} \cup \{2, 3\} = \{1, 2, 3\}$. Thus, we can write $x \in (A \cup B)$ if and only if $(x \in A)$ or $(x \in B)$. Note that $A \cup B = B \cup A$.

Remark 1.3.
Similarly we can define the union of three or more sets. In particular, if $A_1, A_2, A_3, \ldots, A_n$ are n sets, their union $A_1 \cup A_2 \cup A_3 \ldots \cup A_n$ is a set containing all elements that are in at least one of the sets. We can write this union more compactly by

$$\bigcup_{i=1}^{n} A_i.$$

Property 1.2.
The intersection of two sets A and B, denoted by $A \cap B$, consists of all elements that are both in A and B. For example, $\{1, 2\} \cap \{2, 3\} = \{2\}$.

Remark 1.4.
More generally, for sets $A_1, A_2, A_3, \ldots$, their intersection $\bigcap_i A_i$ is defined as the set consisting of the elements that are in all A_i's.

Property 1.3.
The complement of a set A, denoted by A^c or $\bar{A}$, is the set of all elements that are in the universal set Ω but are not in A.

Remark 1.5.
Two sets A and B are mutually exclusive or disjoint if they do not have any shared elements; i.e., their intersection is the empty set, $A \cap B = \emptyset$. More generally, several sets are called disjoint if they are pairwise disjoint, i.e., no two of them share a common elements.

Property 1.4.
The difference (subtraction) is defined as follows. The set $A - B$ consists of elements that are in A but not in B. For example if $A = \{1, 2, 3\}$ and $B = \{3, 5\}$, then $A - B = \{1, 2\}$. Note that $A - B = A \cap \bar{B}$.

Definition 1.4.
A collection of nonempty sets $A_1, A_2, \ldots$ is a partition of a set A if they are disjoint and their union is A.

Example 1.6.
• If the earth's surface is our sample space, we might want to partition it to the different continents.
• A country can be partitioned to different provinces.

Here are some rules that are often useful when working with sets. We will see examples of their usage shortly.

Theorem 1.1. ***De Morgan's law.***
For any sets $A_1, A_2, \ldots, A_n$, we have

- $\overline{(A_1 \cup A_2 \cup A_3 \ldots \cup A_n)} = \bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3 \ldots \cap \bar{A}_n$;
- $\overline{(A_1 \cap A_2 \cap A_3 \ldots \cap A_n)} = \bar{A}_1 \cup \bar{A}_2 \cup \bar{A}_3 \ldots \cup \bar{A}_n$;

Theorem 1.2. ***Distributive law.***
For any sets A, B, and C we have

- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$;
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;

Example 1.7.
If the universal set is given by $\Omega = \{1, 2, 3, 4, 5, 6\}$, and $A = \{1, 2\}$, $B = \{2, 4, 5\}$, $C = \{1, 5, 6\}$ are three sets, find the following set,

1. $A \cup B$.
2. $A \cap B$.
3. $\bar{A}$.
4. $\bar{B}$.
5. Check De Morgan's law by finding $\overline{(A \cup B)}$ and $\bar{A} \cap \bar{B}$.
6. Check the distributive law by finding $A \cap (B \cup C)$ and $(A \cap B) \cup (A \cap C)$

Solution

1. $A \cup B = \{1, 2, 4, 5\}$.
2. $A \cap B = \{2\}$.
3. $\bar{A} = \{3, 4, 5, 6\}$ ($\bar{A}$ consists of elements that are in Ω but not in A).
4. $\bar{B} = \{1, 3, 6\}$.

5. We have
$$\overline{(A \cup B)} = \{1,2,4,5\}^c = \{3,6\},$$
which is the same as
$$\bar{A} \cap \bar{B} = \{3,4,5,6\} \cap \{1,3,6\} = \{3,6\}.$$

6. We have
$$A \cap (B \cup C) = \{1,2\} \cap \{1,2,4,5,6\} = \{1,2\},$$
which is the same as
$$(A \cap B) \cup (A \cap C) = \{2\} \cup \{1\} = \{1,2\}.$$

Property 1.5.
A Cartesian product of two sets A and B, written as $A \times B$, is the set containing ordered pairs from A and B. That is, if $C = A \times B$, then each element of C is of the form (x, y), where $x \in A$ and $y \in B$:
$$A \times B = \{(x,y) | x \in A \quad and \quad y \in B\}.$$

Example 1.8.
Let be $A = \{1,2,3\}$ and $B = \{H,T\}$, then
$$A \times B = \{(1,H),(1,T),(2,H),(2,T),(3,H),(3,T)\}.$$
Note that here the pairs are ordered, so for example, $(1,H) \neq (H,1)$. Thus $A \times B$ is not the same as $B \times A$

Example 1.9.
An important example of sets obtained using a Cartesian product is $\mathbb{R}^n$, where n is a natural number. For $n = 2$, we have
$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x,y) | x \in \mathbb{R}, y \in \mathbb{R}\}.$$
Thus, $\mathbb{R}^2$ is the set consisting of all points in the two-dimensional plane. Similarly, $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ and so on.

1.2.3 Cardinality: Countable and Uncountable Sets

We talk about cardinality of a set, which is basically the size of the set. The cardinality of a set is denoted by $|A|$. We first discuss cardinality for finite sets and then talk about infinite sets.

a. Finite Sets
Consider a set A. If A has only a finite number of elements, its cardinality is simply the number of elements in A. For example, if $A = \{2,4,6,8,10\}$, then $|A| = 5$.

We consider a very useful rule: the inclusion-exclusion principle. For two finite sets A and B, we have

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

We can extend the same idea to three or more sets.

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

Generally, for n finite sets $A_1, A_2, A_3, \ldots, A_n$, we can write

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{i=1}^{n} |A_i| - \sum_{i<j} |A_i \cap A_j| + \sum_{i<j<k} |A_i \cap A_j \cap A_k| - \ldots + (-1)^{n+1} |A_1 \cap \ldots \cap A_n|.$$

b. Infinite Sets

What if A is an infinite set?

We need to distinguish between two types of infinite sets, where one type is significantly "larger" than the other.

In particular, one type is called countable, while the other is called uncountable. Sets such as $\mathbb{N}$ and $\mathbb{Z}$ are called countable, but "bigger" sets such as $\mathbb{R}$ are called uncountable. The difference between the two types is that you can list the elements of a countable set A, i.e., you can write $A = \{a_1, a_2, \ldots\}$, but you cannot list the elements in an uncountable set. For example, you can write

- $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$,
- $\mathbb{Z} = \{0, 1, -1, 2, -2, 3, -3, \ldots\}$.

The fact that you can list the elements of a countably infinite set means that the set can be put in one-to-one correspondence with natural numbers $\mathbb{N}$. On the other hand, you cannot list the elements in $\mathbb{R}$, so it is an uncountable set. To be precise, here is the definition.

Definition 1.5.
Set A is called countable if one of the following assertion is true

1. if it is a finite set, $|A| < \infty$; or
2. it can be put in one-to-one correspondence with natural numbers $\mathbb{N}$, in which case the set is said to be countably infinite.

A set is called uncountable if it is not countable.

Here is a simple guideline for deciding whether a set is countable or not. As far as applied probability is concerned, this guideline should be sufficient for most cases.

Remark 1.6.

- $\mathbb{N}$, $\mathbb{Z}$, and any of their subsets are countable.
- Any set containing an interval on the real line such as $[a, b]$, $(a, b]$, $[a, b)$, or (a, b), where $a < b$ is uncountable.

Theorem 1.3.
Any subset of a countable set is countable.
Any superset of an uncountable set is uncountable.

Proof
The intuition behind this theorem is the following: If a set is countable, then any "smaller" set should also be countable, so a subset of a countable set should be countable as well. ■

Theorem 1.4.
If $A_1, A_2, \ldots$ is a list of countable sets, then the set $\bigcup_i A_i = A_1 \cup A_2 \cup A_3 \ldots$ is also countable.

Theorem 1.5.
If A and B are countable, then $A \times B$ is also countable.

2 Random experiments

Definition 2.1.
A random experiment is a process by which we observe something uncertain.

Example 2.1.
Before rolling a die, you do not know the result. This is an example of a random experiment.

Definition 2.2.
After the experiment, the result of the random experiment is known. An outcome is a result of a random experiment.

Definition 2.3.
The set of all possible outcomes is called the sample space. Thus in the context of a random experiment, the sample space is our universal set.

Here are some examples of random experiments and their sample spaces.

Example 2.2.

- Random experiment: toss a coin; sample space: $\Omega = \{heads, tails\}$ or as we usually write it, $\{H, T\}$.
- Random experiment: roll a die; sample space: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Random experiment: observe the number of goals in a soccer match; sample space: $\Omega = \{0, 1, 2, 3, \ldots\}$.

Definition 2.4.
A subset of the sample space is called an event. So, if $A \subset \Omega$, then A is an event. Thus an event is a collection of possible outcomes.

Property 2.1.
If A and B are events, then $A \cup B$ and $A \cap B$ are also events.

Remark 2.1.

1. The empty event $\emptyset$ never occurs, we call it an impossible event.
2. The entire event, Ω always occurs, it represents a certain event.
3. The complement of an event A is the event A^c such that A and A^c cannot occur simultaneously. If A occurs, A^c does not occur, and vice versa.
4. The events A and B are compatible if they can occur simultaneously.
5. The events A and C are incompatible if they cannot occur simultaneously.
6. Events that consist of an individual outcome are sometimes referred to as elementary events or simple events.

Definition 2.5.
If the result of our random experiment (outcome) belongs to the set $E \subset \Omega$, we say that the event E has occurred.

Property 2.2.
By remembering the definition of union and intersection, we observe that $A \cup B$ occurs if A or B occurs. Similarly, $A \cap B$ occurs if both A and B occur.

Definition 2.6.
We assume that $\mathcal{F}$ is a σ-field (or σ-algebra), i.e., a (nonempty) collection of subsets of Ω that satisfy

1. $\Omega \in \mathcal{F}$.
2. if $A \in \mathcal{F}$, then $\bar{A} \in \mathcal{F}$.
3. if $A_i \in \mathcal{F}$ is a countable sequence of sets, then $\bigcup_i A_i \in \mathcal{F}$.

Remark 2.2.
Here and in what follows, countable means finite or countably infinite.

Example 2.3.
$\mathcal{F}_1 = \{\emptyset, \Omega\}$ is a trivial σ-field, it represents the smallest σ-field.
$\mathcal{F}_2 = \mathcal{P}(\sigma)$ is the biggest σ-field.

Definition 2.7.
Let be Ω the sample space, and $\mathcal{F}$ the σ-field related to Ω, we call $(\Omega, \mathcal{F})$ a measurable space

Definition 2.8.
Given a sample space Ω and an associated σ-field $\mathcal{F}$, a probability function P is a function with domain $\mathcal{F}$ and range $[0, 1]$ ($P : \mathcal{F} \longrightarrow [0, 1]$) such that

1. For every $A \in \mathcal{F}, P(A) \geq 0$.
2. $P(\Omega) = 1$
3. If $A_1, A_2 \ldots$ are pairwise disjoint (i.e, $A_i \cap A_j = \emptyset$, for all $i \neq j$) then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} A_i$

Definition 2.9.
Let be Ω the sample space, $\mathcal{F}$ the σ-field related to Ω, and P the probability function, a probability space is a triple $(\Omega, \mathcal{F}, P)$.

Property 2.3.

- ***Non-Negativity:*** *The probability of any event is always non-negative. For any event A, $P(A) \geq 0$.*
- ***Normalization:*** *The probability of the sure event (sample space) is 1. If Ω is the sample space, then $P(\Omega) = 1$.*

- ***Additivity (Sum Rule):*** *For any two mutually exclusive (disjoint) events A and B, the probability of their union is the sum of their individual probabilities:* $P(A \cup B) = P(A) + P(B)$.
- ***Complementary Rule:*** *The probability of an event not happening is 1 minus the probability of it happening. For an event A and its complement $\bar{A}$, this is* $P(\bar{A}) = 1 - P(A)$.
- $P(\emptyset) = 0$.
- *If $A \subset B$, then $P(A) \leq P(B)$. This is intuitive because A is contained in B.*

2.1 Construction of a probability measure

Consider an arbitrary countable set $\Omega = \{\omega_1, \omega_2, \ldots\}$ and an arbitrary collection $(p_1, p_2, \ldots)$ of non-negative numbers with sum $p_1 + p_2 + \ldots = 1$, such that $P(\{\omega_i\}) = p_i$. Put,

$$P(A) = \sum_{i:\omega_i \in A} p_i$$

Then P satisfies the above axioms. The numbers $(p_1, p_2, \ldots)$ are called a probability distribution.

Theorem 2.1.
For Ω finite or countable and $(\Omega, P(\Omega))$ a measurable space, if $(P_\omega)_{\omega\in\Omega}$ are non-negative numbers such that $\sum_{\omega\in\Omega} P_\omega = 1$, then $(P_\omega)_{\omega\in\Omega}$ represents a probability distribution.

Remark 2.3.
As mentioned above, if Ω is not finite then it may not be possible to let $\mathcal{F}$ be all subsets of Ω. For example, it can be shown that it is impossible to define a P for all possible subsets of the interval $[0, 1]$ that will satisfy the axioms. Instead we define P for special subsets, namely the intervals $[a, b]$, with the natural choice of $P([a, b]) = b - a$. We then use F_1, F_2, F_3 to construct $\mathcal{F}$ as the collection of sets that can be formed from countable unions and intersections of such intervals, and deduce their probabilities from the axioms.

Remark 2.4.
As a consequence of property 2.3, we say that P is a subadditive set function, as it is one for which

$$P(A \cup B) \leq P(A) + P(B),$$

for all A, B. It is also a submodular function, since

$$P(A \cup B) + P(A \cap B) \leq P(A) + P(B),$$

for all A, B.

2.2 Characterization of probabilities in the case of equiprobable events

Consider an arbitrary countable set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, and let be $(\Omega, P(\Omega), P)$ the probability space. The elementary events $\{\omega_i\}$ constitue a partition of Ω, $\{\omega_i\} \neq \emptyset$, for $i = 1, \ldots, n$ and $\cup_{i=1}^{n} \omega_i = \Omega$ and $\{\omega_i\} \cap \{\omega_j\} = \emptyset$ for $i \neq j$. We put $P(\{\omega_i\}) = p_i$.

$$P(A) = \sum_{i:\omega_i \in A} p_i$$

Definition 2.10.
An assignment of probability is said to be equally likely (or uniform) if each elementary event in Ω is assigned the same probability. Thus, if Ω contains n points ω_j, $P\{\omega_j\} = 1/n,\ j = 1, 2, \ldots, n$.
With this assignment

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{The number of outcomes in event } A}{\text{The total number of possible outcomes in the entire sample space } \Omega}$$

Remark 2.5.
This probability is called an uniform probability on $(\Omega, P(\Omega))$.

Example 2.4.
A coin is tossed twice. The sample space consists of four points. Under the uniform assignment, each of four elementary events is assigned probability 1/4.

Example 2.5.
Three dice are rolled. The sample space consists of 6^3 points. Each one-point set is assigned probability $1/6^3$.

3 Conditional probability

So far, we have computed probabilities of events on the assumption that no information was available about the experiment other than the sample space. Sometimes, however, it is known that an event B has happened. How do we use this information in making a statement concerning the outcome of another event A? Consider the following example.

Example 3.1.
Let urn 1 contains one white and two black balls, and urn 2, one black and two white balls. A fair coin is tossed. If a head turns up, a ball is drawn at random from urn 1 otherwise, from urn 2. Let E be the event that the ball drawn is black. The sample space is $\Omega = \{Hb_{11}, Hb_{12}, Hw_{11}, Tb_{21}, Tw_{21}, Tw_{22}\}$, where H denotes head, T denotes tail, b_{ij} denotes j^{th} black ball in i^{th} urn, $i = 1, 2$, and so on. Then

$$P(E) = P\{Hb_{11}, Hb_{12}, Tb_{21}\} = \frac{3}{6} = \frac{1}{2}.$$

If, however, it is known that the coin showed a head, the ball could not have been drawn from urn 2. Thus, the probability of E, conditional on information H, is 2/3. Note that this probability equals the ratio PHead and ball drawn black/PHead.

Definition 3.1.
Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $B \in \mathcal{F}$, with $P(B) > 0$. For an arbitrary $A \in \mathcal{F}$ we shall write

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

and call the quantity so defined the conditional probability of A, given B. Conditional probability remains undefined when $P(B) = 0$.

Theorem 3.1.
Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $B \in \mathcal{F}$, with $P(B) > 0$. Then $(\Omega, \mathcal{F}, P_B)$, where $P_B(A) = P(A|B)$ for all $A \in \mathcal{F}$, is a probability space.

Proof
Clearly $P_B(A) = P(A|B) \geq 0$ for all $A \in \mathcal{F}$. Also, $P_B(\Omega) = P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = 1$.
If $A_1, A_2, \ldots$ is a disjoint sequence of sets in $\mathcal{F}$, then

$$P_B\left(\sum_{i=1}^{\infty} A_i\right) = P\left(\sum_{i=1}^{\infty} A_i|B\right) = \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} = \sum_{i=1}^{\infty} P_B(A_i).$$

■

Remark 3.1.
Let A and B be two events with $P(A) > 0$, $P(B) > 0$. Then it follows from (1) that

$$\begin{cases} P(A \cap B) = & P(A)P(B|A) \\ P(A \cap B) = & P(B)P(A|B) \end{cases} \tag{2}$$

Equation (2) may be generalized to any number of events. The result is given in the following theorem.

Theorem 3.2. ***(The Multiplication Rule).***
Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $A_1, A_2, \ldots, A_n \in \mathcal{F}$, with

$$P\left(\bigcap_{j=1}^{k} A_j\right) > 0 \quad k = 1, \ldots, n-1.$$

Then

$$P\left\{\bigcap_{j=1}^{n} A_j\right\} = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\ldots P\left(A_n|\bigcap_{j=1}^{n-1} A_j\right). \tag{3}$$

Example 3.2.
Consider a hand of five cards in a game of poker. If the cards are dealt at random, there are $\begin{pmatrix} 52 \\ 5 \end{pmatrix}$ possible hands of five cards each. Let $A = \{\text{at least 3 cards of spades}\}$ and $B = \{\text{all 5 cards of spades}\}$. Then

$$P(A \cap B) = P\{\text{all 5 cards of spades}\} = \frac{\begin{pmatrix} 13 \\ 5 \end{pmatrix}}{\begin{pmatrix} 52 \\ 5 \end{pmatrix}}$$

and

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\begin{pmatrix} 13 \\ 5 \end{pmatrix} / \begin{pmatrix} 52 \\ 5 \end{pmatrix}}{\left[\begin{pmatrix} 13 \\ 3 \end{pmatrix}\begin{pmatrix} 39 \\ 2 \end{pmatrix} + \begin{pmatrix} 13 \\ 4 \end{pmatrix}\begin{pmatrix} 39 \\ 1 \end{pmatrix} + \begin{pmatrix} 13 \\ 5 \end{pmatrix}\right] / \begin{pmatrix} 52 \\ 5 \end{pmatrix}}$$

In fact, all rules that we have learned so far can be extended to conditional probability. We can see it in the following proposition,

Proposition 3.1.
Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $A, B, C \in \mathcal{F}$, three events, with $P(C) > 0$, we have

- $P(\bar{A}|C) = 1 - P(A|C)$;
- $P(\emptyset|C) = 0$;
- $P(A - B|C) = P(A|C) - P(A \cap B|C)$;
- $P(A \cup B|C) = P(A|C) + P(B|C) - P(A \cap B|C)$;
- *if $A \subset B$ then $P(A|C) \leq P(B|C)$.*

Let's look at some special cases of conditional probability.

Proposition 3.2.
Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $A, B \in \mathcal{F}$, with $P(B) > 0$. When A and B are disjoint: In this case $A \cap B = \emptyset$, so

$$P(A|B) = 0$$

.

Remark 3.2.
This makes sense. In particular, since A and B are disjoint they cannot both occur at the same time. Thus, given that B has occurred, the probability of A must be zero.

Proposition 3.3.
Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $A, B \in \mathcal{F}$, with $P(B) > 0$.
- *When B is a subset of A: If $B \subset A$, then whenever B happens, A also happens. Thus, given that B occurred, we expect that probability of A be one. In this case $A \cap B = B$, so*

$$P(A|B) = 1$$

- *When A is a subset of B: In this case $A \cap B = A$, so*

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

Example 3.3.
We roll a fair die twice and obtain two numbers X_1 = result of the first roll and X_2 = result of the second roll. Given that we know $X_1 + X_2 = 7$, what is the probability that $X_1 = 4$ or $X_2 = 4$?

Solution.
Let A be the event that $X_1 = 4$ or $X_2 = 4$ and B be the event that $X_1 + X_2 = 7$. We are interested in $P(A|B)$, so we can use

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

We note that

$$A = \{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),(1,4),(2,4),(3,4),(5,4),(6,4)\},$$

$$B = \{(6,1),(5,2),(4,3),(3,4),(2,5),(1,6)\},$$

$$A \cap B = \{(4,3),(3,4)\}.$$

We conclude

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{3}$$

3.1 Law of total probability

Let us start this section by asking a very simple question: In a certain country there are three provinces, call them B_1, B_2, and B_3 (i.e., the country is partitioned into three disjoint sets B_1, B_2, and B_3). We are interested in the total forest area in the country. Suppose that we know that the forest area in B_1, B_2, and B_3 are $100km^2$, $50km^2$, and $150km^2$, respectively. What is the total forest area in the country? If your answer is

$$100km^2 + 50km^2 + 150km^2 = 300km^2,$$

you are right. That is, you can simply add forest areas in each province (partition) to obtain the forest area in the whole country. This is the idea behind the law of total probability, in which the area of forest is replaced by probability of an event A. In particular, if you want to find $P(A)$, you can look at a partition of Ω, and add the amount of probability of A that falls in each partition. We have already seen the special case where the partition is B and $\bar{B}$: we saw that for any two events A and B,

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

and using the definition of conditional probability, $P(A \cap B) = P(A|B)P(B)$, we can write

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}).$$

We can state a more general version of this formula which applies to a general partition of the sample space Ω.

Theorem 3.3.
A (finite or countable) collection $\{B_i\}_i$ of disjoint events such that $\bigcup_i B_i = \Omega$ is said to be a partition of the sample space Ω. For any event A,

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$

where the second summation extends only over B_i for which $P(B_i) > 0$

Proof
Since $B_1, B_2, B_3, \ldots$ is a partition of the sample space Ω, we can write

$$A = A \cap \Omega = A \cap \left(\bigcup_i B_i \right) = \bigcup_i (A \cap B_i).$$

So,

$$P(A) = P\left(\bigcup_i (A \cap B_i) \right) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i).$$

■

Remark 3.3.
Here is a typical scenario in which we use the law of total probability. We are interested in finding the probability of an event A, but we don't know how to find $P(A)$ directly. Instead, we know the conditional probability of A given some events B_i, where the B_i's form a partition of the sample space. Thus, we will be able to find $P(A)$ using the law of total probability, $P(A) = \sum_i P(A|B_i)P(B_i)$.

3.2 Bayes'formula

Now we are ready to state one of the most useful results in conditional probability: Bayes' formula. Suppose that we know $P(A|B)$, but we are interested in the probability $P(B|A)$.

Theorem 3.4. ***(Bayes' formula).***
Suppose $\{B_i\}_i$ is a partition of the sample space and A is an event for which $P(A) > 0$. Then for any event B_j in the partition for which $P(B_j) > 0$,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

where the summation in the denominator extends only over B_i for which $P(B_i) > 0$

Example 3.4.
A screening test is 98% effective in detecting a certain disease when a person has the disease. However, the test yields a false positive rate of 1% of the healthy persons tested. If 0.1% of the population have the disease, what is the probability that a person who tests positive has the disease?
The event + means that a person is tested positive.

$$P(+|D) = 0.98, \quad P(+|\bar{D}) = 0.01, \quad P(D) = 0.001.$$

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|\bar{D})P(\bar{D})} \approx 0.09$$

4 Independence of events

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $A, B \in \mathcal{F}$, with $P(B) > 0$. By the multiplication rule we have

$$P(A \cap B) = P(B)P(A|B).$$

In many experiments the information provided by B does not affect the probability of event A, that is, $P(A|B) = P(A)$.

Example 4.1.
Let two fair coins be tossed, and let

$$A = \{\text{head on the second throw}\}, \quad B = \{\text{head on the first throw}\}.$$

Then

$$P(A) = P(HH, TH) = \frac{1}{2}, \quad P(B) = P(HH, HT) = \frac{1}{2},$$

and

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

Thus

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

Definition 4.1.
Two events, A and B, are said to be independent if and only if

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

Remark 4.1.
Note that we have not placed any restriction on $P(A)$ or $P(B)$. Thus conditional probability is not defined when $P(A)$ or $P(B) = 0$ but independence is. Clearly, if $P(A) = 0$, then A is independent of every $E \in \mathcal{F}$. Also, any event $A \in \mathcal{F}$ is independent of $\emptyset$ and Ω.

Theorem 4.1.
Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $A, B \in \mathcal{F}$. If A and B are independent events, then

$$P(A|B) = P(A) \quad \textit{if} \quad P(B) > 0$$

and

$$P(B|A) = P(B) \quad \textit{if} \quad P(A) > 0$$

Theorem 4.2.
Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $A, B \in \mathcal{F}$. If A and B are independent, so are A and $\bar{B}$, $\bar{A}$ and B, and $\bar{A}$ and $\bar{B}$.

Proof

$$\begin{aligned} P(\bar{A} \cap B) = P(B - (A \cap B)) &= P(B) - P(A \cap B) \quad \text{since} \quad (A \cap B) \subseteq B \\ &= P(B)(1 - P(A)) = P(\bar{A})P(B). \end{aligned}$$

Similarly, one proves that (i) $\bar{A}$ and $\bar{B}$ and (ii) A and $\bar{B}$ are independent. ■

Remark 4.2.
We wish to emphasize that independence of events is not to be confused with disjoint or mutually exclusive events. If two events, each with nonzero probability, are mutually exclusive, they are obviously dependent since the occurrence of one will automatically preclude the occurrence of the other. Similarly, if A and B are independent and $P(A) > 0$, $P(B) > 0$, then A and B cannot be mutually exclusive.

Example 4.2.
A card is chosen at random from a deck of 52 cards. Let A be the event that the card is an ace and B, the event that it is a club. Then

$$P(A) = \frac{4}{52} = \frac{1}{13}\ ,\ P(B) = \frac{13}{52} = \frac{1}{4}\ ,\ P(A \cap B) = P\{\text{ace of clubs}\} = \frac{1}{52},$$

so that A and B are independent.

Example 4.3.
Consider families with two children, and assume that all four possible distributions of sex $\{BB, BG, GB, GG\}$, where B stands for boy and G for girl are equally likely. Let E be the event that a randomly chosen family has at most one girl and F, the event that the family has children of both sexes. Then

$$P(E) = \frac{3}{4}, \quad P(F) = \frac{1}{2}, \quad \text{and} \quad P(E \cap F) = \frac{1}{2},$$

so that E and F are not independent.
Now consider families with three children. Assuming that each of the eight possible sex distributions is equally likely, we have

$$P(E) = \frac{4}{8}, \quad P(F) = \frac{6}{8}, \quad \text{and} \quad P(E \cap F) = \frac{3}{8},$$

so that E and F are independent.

An obvious extension of the concept of independence between two events A and B to a given collection $\mathfrak{U}$ of events is to require that any two distinct events in $\mathfrak{U}$ be independent

Definition 4.2.
Let $\mathfrak{U}$ be a family of events from $\mathcal{F}$. We say that the events $\mathfrak{U}$ are pairwise independent if and only if, for every pair of distinct events $A, B \in \mathfrak{U}$,

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

A much stronger and more useful concept is mutual or complete independence.

Definition 4.3.
A family of events $\mathfrak{U}$ is said to be a mutually or completely independent family if and only if, for every finite sub collection $\{A_1, A_2, \ldots, A_k\}$ of $\mathfrak{U}$, the following relation holds:

$$P(A_1 \cap A_2 \cap \ldots, \cap A_k) = \prod_{j=1}^{k} P(A_j).$$

In what follows we will omit the adjective mutual or complete and speak of independent events. It is clear from the above definition that in order to check the independence of n events $A_1, A_2, \ldots, A_n \in \mathcal{F}$, we must check the following $2^n - n - 1$ relations.

$$P(A_i \cap A_j) = P(A_i)P(A_j), \ i \neq j; \ i, j = 1, 2, \ldots, n.$$

$$P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k), \ i \neq j \neq k; \ i, j, k = 1, 2, \ldots, n.$$

$$\vdots$$

$$P(A_i \cap A_j \cap \ldots \cap A_n) = P(A_i)P(A_j) \ldots P(A_n)$$

The first of these requirements is pairwise independence. Independence therefore implies pairwise independence, but not conversely.

Example 4.4.
Take four identical marbles. On the first, write symbols $A_1A_2A_3$. On each of the other three, write A_1, A_2, A_3, respectively. Put the four marbles in an urn and draw one at random. Let E_i denote the event that the symbol A_i appears on the drawn marble. Then

$$P(E_1) = P(E_2) = P(E_3) = \frac{1}{2},$$

$$P(E_1 \cap E_2) = P(E_2 \cap E_3) = P(E_1 \cap E_3) = \frac{1}{4},$$

and

$$P(E_1 \cap E_2 \cap E_3) = \frac{1}{4}$$

It follows that although events E_1, E_2, E_3 are not independent, they are pairwise independent.

5 Some exercises

5.1

Suppose that the universal set Ω is defined as $\Omega = \{1, 2, \ldots, 10\}$ and $A = \{1, 2, 3\}$, $B = \{x \in \Omega : 2 \leq x \leq 7\}$, and $C = \{7, 8, 9, 10\}$.

1. Find $A \cup B$
2. Find $(A \cup C) - B$
3. Find $\bar{A} \cup (B - C)$
4. Do A, B and C form a partition of Ω?

Solution.

1. $A \cup B = \{1, 2, 3, 4, 5, 6, 7\}$
2. $A \cup C = \{1, 2, 3, 7, 8, 9, 10\}$ and $B = \{2, 3, \ldots, 7\}$ thus, $(A \cup C) - B = \{1, 8, 9, 10\}$
3. $\bar{A} = \{4, 5, \ldots, 10\}$, $B - C = \{2, 3, 4, 5, 6\}$ thus, $\bar{A} \cup (B - C) = \{2, 3, \ldots, 10\}$
4. No, since they are not disjoint. For example,
$$A \cap B = \{2, 3\} \neq \emptyset$$

5.2

Determine whether each of the following sets is countable or uncountable.

1. $A = \{1, 2, \ldots, 10^{10}\}$.
2. $B = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$.
3. $C = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$.

Solution.

1. A is countable because it is a finite set.
2. B is countable because we can create a list with all the elements. we can write any set B in the form of
$$B = \bigcup_i \bigcup_j = \{q_{ij}\}$$
where indices i and j belong to some countable sets, that set in this form is countable. For this case we can write
$$B = \bigcup_{i \in \mathbb{Q}} \bigcup_{j \in \mathbb{Q}} = \{a_i + b_j\sqrt{2}\}$$
So, we can replace q_{ij} by $a_i + b_j\sqrt{2}$
3. C is uncountable. To see this, note that for all $x \in [0, 1]$ then $(x, 0) \in C$.

5.3

Two teams A and B play a soccer match, and we are interested in the winner. The sample space can be defined as:

$$\Omega = \{a, b, d\}$$

where a shows the outcome that A wins, b shows the outcome that B wins, and d shows the outcome that they draw. Suppose that we know that (1) the probability that A wins is $P(a) = P(\{a\}) = 0.5$, and (2) the probability of a draw is $P(d) = P(\{d\}) = 0.25$.

1. Find the probability that B wins.
2. Find the probability that B wins or a draw occurs.

Solution.

1. $P(a) + P(b) + P(d) = 1, \quad P(a) = 0.5, \quad P(d) = 0.25$.
   Therefore $P(b) = 0.25$.
2. $P(\{b, d\}) = P(b) + P(d) = 0.5$

5.4

I roll a fair die twice and obtain two numbers. X_1 = result of the first roll, X_2 = result of the second roll.

1. Find the probability that $X_2 = 4$.
2. Find the probability that $X_1 + X_2 = 7$.
3. Find the probability that $X_1 \neq 2$ and $X_2 \leq 4$.

Solution.
The sample space has 36 elements:

$$\Omega = \{(1,1), (1,2), \ldots, (1,6), (2,1), (2,2), \ldots, (2,6), \ldots (6,1), (6,2), \ldots, (6,6)\}$$

1. The event $X_2 = 4$ can be represented by the set.

   $$A = \{(1,4), (2,4), (3,4), (4,4), (5,4), (6,4)\}$$

   Thus,

   $$P(A) = \frac{|A|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$$

2. $$B = \{(x_1, x_2) \mid x_1 + x_2 = 7\} = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$$

   Therefore

   $$P(B) = \frac{6}{36} = \frac{1}{6}$$

3.

$$C = \{(X_1, X_2) \mid X_1 \neq 2, X_2 \geq 4\}$$

$$= \{(1,4),(1,5),(1,6),(3,4),(3,5),(3,6),(4,4),(4,5),(4,6),(5,4),(5,5),(5,6),(6,4),(6,5),(6,6)\}$$

Therefore

$$|C| = 15$$

Which results in:

$$P(C) = \frac{15}{36} = \frac{5}{12}$$

5.5

Continuity of probability.
For any sequence of events $A_1, A_2, A_3, \ldots$. Prove

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} A_i\right)$$

$$P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} P\left(\bigcap_{i=1}^{n} A_i\right)$$

Solution.
Define the new sequence $B_1, B_2, \ldots,$ as

$$B_1 = A_1$$

$$B_2 = A_2 - A_1$$

$$B_3 = A_3 - (A_1 \cup A_2)$$

$$\cdots$$

$$B_i = A_i - \left(\bigcup_{j=1}^{i-1} A_j\right)$$

Then we have:
(a) B_i's are disjoint.
(b) $\cup_{i=1}^{n} B_i = \cup_{i=1}^{n} A_i$
(c) $\cup_{i=1}^{\infty} B_i = \cup_{i=1}^{\infty} A_i$
Then we can write:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i), \quad (B_i\text{'s are disjoint})$$

$$= \lim_{n\to\infty} P\left(\sum_{i=1}^{n} B_i\right), \quad (\text{definition of infinite sum})$$

$$= \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} B_i\right) \quad (B_i\text{'s are disjoint})$$

$$= \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} A_i\right)$$

To prove the second part, apply the result of the first part to $\bar{A}_1, \bar{A}_2, \ldots$.

5.6

A professor thinks students who live on campus are more likely to get As in the probability course. To check this theory, the professor combines the data from the past few years:

1. 600 students have taken the course.
2. 120 students have got As.
3. 200 students lived on campus.
4. 80 students lived off campus and got As. Does this data suggest that "getting an As" and "living on campus" are dependent or independent?

Solution.
From the data, you can see that 80 students out of the 400 off- campus students got an As (20%). Also, 40 students out of the 200 on campus students got an A (again 20%). Thus, the data suggests that "getting an As" and "living on campus" are independent. You can also see this using the definitions of independence in the following way: Let C be the event that a random student lives on campus and A be the event that he or she gets an As in the course. We have:

$$P(A) \approx \frac{120}{600} = \frac{1}{5}$$

$$P(C) \approx \frac{200}{600} = \frac{1}{3}$$

$$P(A \cap \bar{C}) \approx \frac{80}{600} = \frac{2}{15}$$

$$P(A \cap C) = P(A) - P(A \cap \bar{C}) = \frac{1}{5} - \frac{2}{15} = \frac{1}{15}$$

Therefore,

$$\frac{1}{15} = P(A \cap C) = P(A)P(C)$$

The data suggests that A and C are independent.

5.7

Consider a communication system. At any given time, the communication channel is in good condition with probability 0.8 and is in bad condition with probability 0.2. An error occurs in a transmission with probability 0.1 if the channel is in good condition and with probability 0.3 if the channel is in bad condition. Let G be the event that the channel is in good condition and E be the event that there is an error in transmission.

1. Find $P(G \cap E)$ and $P(\bar{G} \cap E)$.
2. Find $P(E)$.
3. Find $P(G|\bar{E})$.

Solution.

1. $P(G \cap E) = 0.08$ and $P(\bar{G} \cap E) = 0.06$
2. $$P(E) = P(G \cap E) + P(\bar{G} \cap E) = 0.08 + 0.06 = 0.14$$
3. $$P(G \cap \bar{E}) = \frac{P(G \cap \bar{E})}{P(\bar{E})} = \frac{0.72}{1 - 0.14} \approx 0.84$$

5.8

One way to design a spam filter is to look at the words in an email. In particular, some words are more frequent in spam emails. Suppose that we have the following information:

1. 50% of emails are spam.
2. 1% of spam emails contain the word "refinance".
3. 0.001% of non-spam emails contain the word "refinance".

Suppose that an email is checked and found to contain the word "refinance". What is the probability that the email is spam?

Solution.
Let S be the event that an email is spam and let R be the event that the email contains the word "refinance". Then,

$$P(S) = \frac{1}{2}, \quad P(R|S) = \frac{1}{100}, \quad P(R|\bar{S}) = \frac{1}{100000}$$

Then,

$$P(S|R) = \frac{P(R|S)P(S)}{P(R)} = \frac{P(R|S)P(S)}{P(R|S)P(S) + P(R|\bar{S})P(\bar{S})} = \frac{0.01 \times 0.5}{0.01 \times 0.5 + \frac{1}{100000} \times 0.5} \approx 0.999$$

5.9

A family has n children. We pick one of them at random and find out that she is a girl. What is the probability that all their children are girls?

Solution.
Let Gr be the event that a randomly chosen child is a girl. Let A be the event that all the children are girls. Then,

$$P(Gr|A) = 1, \quad P(A) = \frac{1}{2^n}, \quad P(Gr) = \frac{1}{2}$$

Thus,

$$P(A|Gr) = \frac{P(Gr|A)P(A)}{P(Gr)} = \frac{1 \times \frac{1}{2^n}}{\frac{1}{2}} = \frac{1}{2^{n-1}}$$

CHAPTER II. Random Variables

1 Introduction.

What are random variables?
The Holy Roman Empire was, in the words of the historian Voltaire, "neither holy, nor Roman, nor an empire".

Similarly, a random variable is neither random nor a variable:

A random variable is a function defined on a sample space. The values of the function can be anything at all, but for us they will always be numbers. The standard abbreviation for 'random variable' is r.v.

Example 1.1.
We select at random a student from the class and measure his or her height in centimetres. Here, the sample space is the set of students; the random variable is 'height', which is a function from the set of students to the real numbers: $h(\Omega)$ is the height of student S in centimetres. (Remember that a function is nothing but a rule for associating with each element of its domain set an element of its target or range set. Here the domain set is the sample space Ω, the set of students in the class, and the target space is the set of real numbers.)

Example 1.2.
We throw a six-sided die twice. We are interested in the sum of the two numbers. Here the sample space is

$$\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$$

and the random variable F is given by $F(i, j) = i + j$. The target set is the set $\{2, 3, \ldots, 12\}$.

The two random variables in the above examples are representatives of the two types of random variables that we will consider. These definitions are not quite precise, but more examples should make the idea clearer.

In Chapter 1, we dealt essentially with random experiments which can be described by finite sample spaces. We studied the assignment and computation of probabilities of events. In practice, one observes a function defined on the space of outcomes. Thus, if a coin is tossed n times, one is not interested in knowing which of the 2^n n-tuples in the sample space has occurred. Rather, one would like to know the number of heads in n tosses. In games of chance one is interested in the net gain or loss of a certain player. Actually, in Chapter 1 we were concerned with such functions without defining the term random variable. Here we study the notion of a random variable and examine some of its properties.

In this chapter, we define a random variable, while we study the notion of probability distribution of a random variable. We deal with some special types of random variables, and we consider functions of a random variable and their induced distributions.
The fundamental difference between a random variable and a real-valued function of a real variable is the associated notion of a probability distribution. Nevertheless our knowledge of advanced calculus or real analysis is the basic tool in the study of random variables and their probability distributions.

2 Random variables

In Chapter 1 we studied properties of a set function P defined on a sample space $(\Omega, \mathcal{F})$. Since P is a set function, it is not very easy to handle; we cannot perform arithmetic or algebraic operations on sets. Moreover, in practice one frequently observes some function of elementary events. When a coin is tossed repeatedly, which replication resulted in heads is not of much interest. Rather one is interested in the number of heads, and consequently the number of tails, that appear in, say, n tossings of the coin. It is therefore desirable to introduce a point function on the sample space. We can then use our knowledge of calculus or real analysis to study properties of P.

Definition 2.1.
Let $(\Omega, \mathcal{F})$ be a sample space. A finite, single-valued function which maps Ω into $\mathbb{R}$ is called a random variable (r.v.) if the inverse images under X of all Borel sets in $\mathbb{R}$ are events, that is, if

$$X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{F}, \quad \text{for all} \quad B \in \mathfrak{B} \tag{1}$$

Remark 2.1.
In order to verify whether a real-valued function on $(\Omega, \mathcal{F})$ is a r.v., it is not necessary to check that (1) holds for all Borel sets $B \in \mathfrak{B}$. It suffices to verify (1) for any class A of subsets of $\mathbb{R}$ which generates $\mathfrak{B}$. By taking A to be the class of semiclosed intervals $(-\infty, x],\ x \in \mathbb{R}$ we get the following result.

Theorem 2.1.
X is a r.v. if and only if for each $x \in \mathbb{R}$

$$\{\omega : X(\omega) \leq x\} = X \leq x \in \mathcal{F} \tag{2}.$$

Remark 2.2.
Note that the notion of probability does not enter into the definition of a r.v.

Remark 2.3.
If X is a r.v, the sets $\{X = x\}$, $\{a < X \leq b\}$, $\{X < x\}$, $\{a \leq X < b\}$, $\{a < X < b\}$,

$\{a \leq X \leq b\}$ are all events. Moreover, we could have used any of these intervals to define a r.v. For example, we could have used the following equivalent definition: X is a r.v. if and only if

$$\{\omega : X(\omega) < x\} \in \mathcal{F}, \quad \text{for all} \quad x \in \mathbb{R} \tag{3}.$$

We have

$$\{X < x\} = \bigcup_{n=1}^{\infty} \left\{X \leq x - \frac{1}{n}\right\} \tag{4}$$

and

$$\{X \leq x\} = \bigcap_{n=1}^{\infty} \left\{X \leq x + \frac{1}{n}\right\} \tag{5}$$

Remark 2.4.
In practice (1) or (2) is a technical condition in the definition of a r.v. which we may ignore and think of r.v.s simply as real-valued functions defined on Ω. It should be emphasized though that there do exist subsets of $\mathbb{R}$ which do not belong to $\mathfrak{B}$ and hence there exist real-valued functions defined on Ω which are not r.v.s but we will not encounter them in practical applications.

Example 2.1.
For any set $A \subseteq \Omega$, define

$$I_A(\omega) = \begin{cases} 0, & \omega \notin A; \\ 1, & \omega \in A. \end{cases}$$

$I_A(\omega)$ is called the indicator function of set A. I_A is a r.v. if and only if $A \in \mathcal{F}$.

Example 2.2.
Let $\Omega = \{H, T\}$ and $\mathcal{F}$ be the class of all subsets of Ω. Define X by $X(H) = 1$, $X(T) = 0$. Then

$$X^{-1}(-\infty, x] = \begin{cases} \emptyset, & \text{if } x < 0; \\ \{T\}, & \text{if } 0 \leq x < 1; \\ \{H, T\}, & \text{if } x \geq 1. \end{cases}$$

and we see that X is a r.v.

Example 2.3.
Let $\Omega = \{HH, TT, HT, TH\}$ and $\mathcal{F}$ be the class of all subsets of Ω. Define X by

$$X(\omega) = \text{number of } H\text{'s in } \omega.$$

Then $X(HH) = 2$, $X(HT) = X(TH) = 1$, and $X(TT) = 0$.

$$X^{-1}(-\infty, x] = \begin{cases} \emptyset, & \text{if } x < 0; \\ \{TT\}, & \text{if } 0 \leq x < 1; \\ \{TH, HT, TT\}, & \text{if } 1 \leq x < 2; \\ \Omega, & \text{if } x \geq 2. \end{cases}$$

Thus X is a r.v.

Remark 2.5.
Let $(\Omega, \mathcal{F})$ be a discrete sample space; that is, let Ω be a countable set of points and $\mathcal{F}$ be the class of all subsets of Ω. Then every numerical valued function defined on $(\Omega, \mathcal{F})$ is a r.v.

Example 2.4. Let $\Omega = [0,1]$ and $\mathcal{F} = \mathfrak{B} \cap [0,1]$ be the σ-field of Borel sets on [0,1]. Define X on Ω by

$$X(\omega) = \omega, \ \omega \in [0,1].$$

Clearly X is a r.v. Any Borel subset of Ω is an event.

Remark 2.6.
Let X be a r.v. defined on $(\Omega, \mathcal{F})$ and a, b be constants. Then $aX + b$ is also a r.v. on $(\Omega, \mathcal{F})$. Moreover, X^2 is a r.v. and so also is $1/X$, provided that $\{X = 0\} = \emptyset$.

3 Probability distribution of a random variable

In the above section, we introduced the concept of a r.v. and noted that the concept of probability on the sample space was not used in this definition. In practice, however, random variables are of interest only when they are defined on a probability space. Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let X be a r.v. defined on it.

Theorem 3.1.
The r.v. X defined on the probability space $(\Omega, \mathcal{F}, P)$ induces a probability space $(\mathbb{R}, \mathfrak{B}, Q)$ by means of the correspondence

$$Q(B) = P(X^{-1}(B)) = P\{\omega : X(\omega) \in B\} \quad \text{for all} \quad B \in \mathfrak{B}. \tag{1}$$

We write $Q = PX^{-1}$ and call Q or PX^{-1} the (probability) distribution of X.

Proof
Clearly $Q(B) \geq 0$ for all $B \in \mathfrak{B}$, and also $Q(\mathbb{R}) = P\{X \in \mathbb{R}\} = P(\Omega) = 1$. Let $B_i \in \mathfrak{B}, i = 1, 2, \ldots$ with $B_i \cap B_j = \emptyset, i \neq j$. Since the inverse image of a disjoint union of Borel sets is the disjoint union of their inverse images, we have

$$Q\left(\bigcup_{i=1}^{\infty} B_i\right) = P\left\{X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right)\right\} = P\left\{\bigcup_{i=1}^{\infty} X^{-1}(B_i)\right\}$$

$$= \sum_{i=1}^{\infty} P\left(X^{-1}(B_i)\right) = \sum_{i=1}^{\infty} Q(B_i)$$

It follows that $(\mathbb{R}, \mathfrak{B}, Q)$ is a probability space, and the proof is complete.

■

Let us first introduce and study some properties of a special point function on $\mathbb{R}$.

Definition 3.1.
A real valued function F defined on $(-\infty, \infty)$ that is nondecreasing, right continuous, and satisfies

$$F(-\infty) = 0 \quad \text{and} \quad F(+\infty) = 1$$

is called a distribution function (DF).

Theorem 3.2.
The set of discontinuity points of a DF is at most countable.

Definition 3.2.
Let X be a r.v. defined on $(\Omega, \mathcal{F}, P)$. Define a point function $F(.)$ on $\mathbb{R}$ by using (1),

$$F(x) = Q(-\infty, x] = P\{\omega : X(\omega) \leq x\} \quad \text{for all} \quad x \in \mathbb{R} \tag{2}$$

The function F is called the distribution function of r.v. X.
If there is no confusion, we will write

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

The following result justifies our calling F as defined by (2) a DF.

Theorem 3.3.
The function F defined in (2) is indeed a DF.

Proof
Let $x_1 < x_2$. Then $(-\infty, x_1] \subset (-\infty, x_2]$, and we have

$$F(x_1) = P\{X \leq x_1\} \leq P\{X \leq x_2\} = F(x_2).$$

Since F is nondecreasing, it is sufficient to show that for any sequence of numbers $x_n \to x$, $x_1 > x_2 > \ldots > x_n > \ldots > x$, $F(x_n) \to F(x)$. Let $A_n = \{\omega : X(\omega) \in (x, x_n]\}$. Then $A_n \in \mathcal{F}$ and $A_n \not\downarrow$. Also,

$$\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n = \emptyset,$$

since none of the intervals $(x, x_n]$ contains x. It follows that $\lim_{n\to\infty} P(A_n) = 0$. But,

$$P(A_n) = P\{X \leq x_n\} - P\{X \leq x\} = F(x_n) - F(x),$$

so that

$$\lim_{n\to\infty} F(x_n) = F(x)$$

and F is right continuous.
Finally, let $\{x_n\}$ be a sequence of numbers decreasing to $-\infty$. Then,

$$\{X \leq x_n\} \supseteq \{X \leq x_{n+1}\}$$

for each n and

$$\lim_{n\to\infty}\{X \le x_n\} = \emptyset$$

Therefore,

$$F(-\infty) = \lim_{n\to\infty} P\{X \le x_n\} = P\left\{\lim_{n\to\infty}\{X \le x_n\}\right\} = 0.$$

Similarly,

$$F(+\infty) = \lim_{x_n\to\infty} P\{X \le x_n\} = 1$$

and the proof is complete.

■

Theorem 3.4.
Every DF is the DF of a r.v. on some probability space.

Remark 3.1.
If X is a r.v. on $(\Omega, \mathcal{F}, P)$, we have seen that $F(x) = P\{X \le x\}$ is a DF associated with X. The above theorem assures us that to every DF F we can associate some r.v. Thus, given a r.v., there exists a DF, and conversely. In this chapter, when we speak of a r.v. we will assume that it is defined on some probability space.

Example 3.1.
Let X be defined on $(\Omega, \mathcal{F}, P)$ by

$$X(\omega) = c \text{ for all } \omega \in \Omega.$$

Then

$$P\{X = c\} = 1,$$

$$F(x) = Q(-\infty, x] = P\{X^{-1}(-\infty, x]\} = 0 \text{ if } x < c$$

and

$$F(x) = 1 \text{ if } x \ge c.$$

Example 3.2.
Let $\Omega = \{H, T\}$ and X be defined by

$$X(H) = 1, \quad X(T) = 0.$$

If P assigns equal mass to $\{H\}$ and $\{T\}$, then

$$P\{X = 0\} = \frac{1}{2} = P\{X = 1\}$$

and

$$F(x) = Q(-\infty, x] = \begin{cases} 0, & x < 0; \\ \frac{1}{2}, & 0 \le x < 1; \\ 1, & 1 \le x. \end{cases}$$

4 Discrete and continuous random variables.

Let X be a r.v. defined on some fixed, but otherwise arbitrary, probability space $(\Omega, \mathcal{F}, P)$, and let F be the DF of X. We shall consider two cases, namely, the case in which the r.v. assumes at most a countable number of values and hence its DF is a step function and that in which the DF F is (absolutely) continuous.

Definition 4.1.
A r.v. X defined on $(\Omega, \mathcal{F}, P)$ is said to be of the discrete type, or simply discrete, if there exists a countable set $E \subseteq \mathbb{R}$ such that $P\{X \in E\} = 1$. The points of E which have positive mass are called jump points or points of increase of the DF of X, and their probabilities are called jumps of the DF.
Note that $E \in \mathfrak{B}$ since every one-point is in $\mathfrak{B}$. Indeed, if $x \in \mathbb{R}$, then

$$\{x\} = \bigcap_{n=1}^{\infty} \left(x - \frac{1}{n}, x + \frac{1}{n}\right) \tag{1}$$

Thus $\{X \in E\}$ is an event. Let X take on the value x_i with probability p_i, $(i = 1, 2, \ldots)$. We have

$$P\{\omega : X(\omega) = x_i\} = p_i, \quad i = 1, 2, \ldots, \text{ and } p_i \geq 0 \text{ for all } i.$$

Then

$$\sum_{i=1}^{\infty} p_i = 1.$$

Definition 4.2.
The collection of numbers $\{p_i\}$ satisfying $P\{X = x_i\} = p_i \geq 0$, for all i and $\sum_{i=1}^{\infty} p_i = 1$, is called the probability mass function (pmf) of r.v. X.
The DF F of X is given by

$$F(x) = P\{X \leq x\} = \sum_{x_i \leq x} p_i. \tag{2}$$

Remark 4.1.
If I_A denotes the indicator function of the set A, we may write

$$X(\omega) = \sum_{i=1}^{\infty} x_i I_{[X=x_i]}(\omega). \tag{3}$$

Let us define a function $\varepsilon(x)$ as follows:

$$\varepsilon(x) = \begin{cases} 1, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

Then we have

$$F(x) = \sum_{i=1}^{\infty} p_i \varepsilon(x - x_i). \tag{4}$$

Example 4.1.
The simplest example is that of a r.v. X degenerate at c, $P\{X = c\} = 1$:

$$F(x) = \varepsilon(x - c) = \begin{cases} 0\ , & x < c; \\ 1\ , & x \geq c. \end{cases}$$

Example 4.2.
A box contains good and defective items. If an item drawn is good, we assign the number 1 to the drawing; otherwise, the number 0. Let p be the probability of drawing at random a good item. Then

$$P(X = x) = \begin{cases} 1 - p & \text{if } x = 0\ , \\ p & \text{if } x = 1 \end{cases}$$

and

$$F(x) = P\{X \leq x\} = \begin{cases} 0\ , & x < 0; \\ 1 - p\ , & 0 \leq x < 1; \\ 1\ , & x \geq 1. \end{cases}$$

Theorem 4.1.
Let $\{p_k\}$ be a collection of nonnegative real numbers such that $\sum_{k=1}^{\infty} p_k = 1$. Then $\{p_k\}$ is the PMF of some r.v. X.

We next consider r.v.s associated with DFs that have no jump points. The DF of such a r.v. is continuous. We shall restrict our attention to a special subclass of such r.v.s.

Definition 4.3.
Let X be a r.v. defined on $(\Omega, \mathcal{F}, P)$ with DF F. Then X is said to be of the continuous type (or, simply, continuous) if F is absolutely continuous, that is, if there exists a nonnegative function $f(x)$ such that for every real number x we have

$$F(x) = \int_{-\infty}^{x} f(t)dt, \tag{5}$$

The function f is called the probability density function (PDF) of the r.v. X.

Proposition 4.1.
Note that $f \geq 0$ and satisfies $\lim_{x \to +\infty} F(x) = F(+\infty) = \int_{-\infty}^{+\infty} f(t)dt = 1$. Let a and b be any two real numbers with $a < b$. Then

$$P\{a < X \leq b\} = F(b) - F(a) = \int_{a}^{b} f(t)dt.$$

In view of this proposition, the following result holds.

Theorem 4.2.
Let X be a r.v. of the continuous type with PDF f. Then for every Borel set $B \in \mathfrak{B}$

$$P(B) = \int_B f(t)dt. \tag{6}$$

If F is absolutely continuous and f is continuous at x, we have

$$F'(x) = \frac{dF(x)}{dx} = f(x). \tag{7}$$

Theorem 4.3.
Every nonnegative real function f that is integrable over $\mathbb{R}$ and satisfies

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

is the PDF of some continuous type r.v. X.

Proof
Define

$$F(x) = \int_{-\infty}^{x} f(t)dt, \ x \in \mathbb{R}.$$

Then $F(-\infty) = 0$, $F(+\infty) = 1$, and, if $x_2 > x_1$,

$$F(x_2) = \left(\int_{-\infty}^{x_1} + \int_{x_1}^{x_2}\right) f(t)dt \geq \int_{-\infty}^{x_1} f(t)dt = F(x_1).$$

Finally, F is (absolutely) continuous and hence continuous from the right.

■

Remark 4.2.
In the discrete case, $P\{X = a\}$ is the probability that X takes the value a. In the continuous case, $f(a)$ is not the probability that X takes the value a. Indeed, if X is of the continuous type, it assumes every value with probability 0.

Theorem 4.4.
Let X be any r.v. Then

$$P\{X = a\} = \lim_{t \to a, t<a} P\{t < X \leq a\} \tag{8}$$

Proof
Let $t_1 < t_2 < \ldots < a$, $t_n \to a$, and write

$$A_n = \{t_n < X \leq a\}.$$

Then A_n is a nonincreasing sequence of events which converges to $\bigcap_{n=1}^{\infty} A_n = \{X = a\}$. It follows that $\lim_{n\to\infty} P(A_n) = P\{X = a\}$. ■

Remark 4.3.
The set of real numbers x for which a DF F increases is called the support of F. Let X be the r.v. with DF F, and let S be the support of F. Then $P(X \in S) = 1$ and $P(X \in \bar{S}) = 0$. The open interval $(0, 1)$ is the support of F in the example below.

Example 4.3.
Let X be a r.v. with DF F given by

$$F(x) = \begin{cases} 0\ , & x \leq 0; \\ x\ , & 0 < x \leq 1; \\ 1\ , & x > 1. \end{cases}$$

Differentiating F with respect to x at continuity points of f , we get

$$f(x) = F'(x) = \begin{cases} 0\ , & x < 0 \ \text{ or } \ x > 1; \\ 1\ , & 0 < x < 1. \end{cases}$$

The function f is not continuous at $x = 0$ or at $x = 1$. We may define $f(0)$ and $f(1)$ in any manner. Choosing $f(0) = f(1) = 0$, we have

$$f(x) = \begin{cases} 1\ , & 0 < x < 1; \\ 0\ , & \text{otherwise.} \end{cases}$$

Then

$$P\{0.4 < X \leq 0.6\} = F(0.6) - F(0.4) = 0.2.$$

Example 4.4.
Let X have the triangular PDF

$$f(x) = \begin{cases} x\ , & 0 < x \leq 1; \\ 2 - x\ , & 1 \leq x \leq 2; \\ 0\ , & \text{otherwise.} \end{cases}$$

It is easy to check that f is a PDF. For the DF F of X we have

- $F(x) = 0, \ \text{ if } \ x \leq 0$
- $F(x) = \int_0^x t dt = \frac{x^2}{2} \ \text{ if } \ 0 < x \leq 1$
- $F(x) = \int_0^1 t dt + \int_1^x (2 - t) dt = 2x - \frac{x^2}{2} - 1 \ \text{ if } \ 1 < x \leq 2$
- $F(x) = 1 \ \text{ if } \ x \leq 2$

Then

$$P\{0.3 < X \leq 1.5\} = P\{X \leq 1.5\} - P\{X \leq 0.3\} = 0.83.$$

Example 4.5.
Let $k > 0$ be a constant, and

$$f(x) = \begin{cases} kx(1-x)\ , & 0 < x < 1,; \\ 0, & \text{otherwise.} \end{cases}$$

Then $\int_0^1 f(x)dx = k/6$. It follows that $f(x)$ defines a PDF if $k = 6$. We have

$$P\{X > 0.3\} = 1 - 6\int_0^{0.3} x(1-x)dx = 0.784.$$

5 Functions of a random variable

Let X be a r.v. with a known distribution, and let g be a function defined on the real line. We seek the distribution of $Y = g(X)$, provided that Y is also a r.v. We first prove the following result.

Theorem 5.1.
Let X be a r.v. defined on $(\Omega, \mathcal{F}, P)$. Also, let g be a Borel-measurable function on $\mathbb{R}$. Then $g(X)$ is also a r.v.

Proof
For $y \in \mathbb{R}$, we have

$$\{g(X) \leq y\} = \{X \in g^{-1}(-\infty, y]\},$$

and since g is Borel-measurable, $g^{-1}(-\infty, y]$ is a Borel set. It follows that $\{g(X) \leq y\} \in \mathcal{F}$ and the proof is complete. ■

Theorem 5.2.
Given a r.v. X with a known DF, the distribution of the r.v. $Y = g(X)$, where g is a Borel-measurable function, is determined.

Proof
Indeed, for all $y \in \mathbb{R}$

$$P\{Y \leq y\} = P\{X \in g^{-1}(-\infty, y]\}. \tag{1}$$

In what follows, we will always assume that the functions under consideration are Borel-measurable. ■

Example 5.1.
Let X be a r.v. with PMF

$$P\{X = k\} = \begin{cases} e^{-\lambda}\frac{\lambda^k}{k!}\ , & k = 0, 1, 2, \ldots; \text{ and } \lambda > 0 \\ 0\ , & \text{otherwise.} \end{cases}$$

Let $Y = X^2 + 3$. Then $y = x^2 + 3$ maps $A = \{0, 1, 2, \ldots\}$ onto $B = \{3, 4, 7, 12, 19, 28, \ldots\}$. The inverse map is $x = \sqrt{(y-3)}$, and since there are no negative values in A we take the positive square root of $y - 3$. We have

$$P\{Y = y\} = P\{X = \sqrt{y-3}\} = \frac{e^{-\lambda}\lambda^{\sqrt{y-3}}}{(y-3)!},\ y \in B,$$

and $P\{Y = y\} = 0$ elsewhere.

Remark 5.1.
Actually the restriction to a single-valued inverse on g is not necessary. If g has a finite (or even a countable) number of inverses for each y, from countable additivity of P we have

$$P\{Y=y\}=P\{g(X)=y\}=P\left(\bigcup_a [X=a, g(a)=y]\right)=\sum_a P\{X=a, g(a)=y\}.$$

Example 5.2.
Let X be a r.v. with PMF

$$P\{X=-2\}=\frac{1}{5},\quad P\{X=-1\}=\frac{1}{6},\quad P\{X=0\}=\frac{1}{5},$$

$$P\{X=1\}=\frac{1}{15},\quad \text{and}\quad P\{X=2\}=\frac{11}{30}.$$

Let $Y=X^2$. Then $A=\{-2,-1,0,1,2\}$ and $B=\{0,1,4\}$. We have

$$P\{Y=y\}=\begin{cases}\frac{1}{5}\ , & y=0;\\ \frac{1}{6}+\frac{1}{15}=\frac{7}{30}\ , & ;y=1\\ \frac{1}{5}+\frac{11}{30}=\frac{17}{30}\ , & y=4.\end{cases}$$

Remark 5.2.
The case in which X is a r.v. of the continuous type is not as simple. First we note that if X is a continuous type r.v. and g is some Borel-measurable function, $Y=g(X)$ may not be a r.v. of the continuous type.

Example 5.3.
Let X be a r.v. with a continuous distribution on $[-1,1]$, defined as follows, the PDF of X is $f(x)=1/2,\ -1\le x\le 1$, and $=0$ elsewhere. Let $Y=X^+$. and $X^+=\begin{cases}X\ , & X\ge 0;\\ 0\ , & X<0.\end{cases}$
Then,

$$P\{Y\le y\}=\begin{cases}0\ , & y<0;\\ \frac{1}{2}\ , & y=0;\\ \frac{1}{2}+\frac{1}{2}y\ , & 0<y\le 1;\\ 1\ , & y>1.\end{cases}$$

We see that the DF of Y has a jump at $y=0$ and that Y is neither discrete nor continuous. Note that all we require is that $P\{X<0\}>0$ for X^+ to be of the mixed type.

Remark 5.3.
The above example shows that we need some conditions on g to ensure that $g(X)$ is also a r.v. of the continuous type whenever X is continuous. This is the case when g is a continuous monotonic function. A sufficient condition is given in the following theorem.

Theorem 5.3.
Let X be a r.v. of the continuous type with PDF f. Let $y = g(x)$ be differentiable for all x and either $g'(x) > 0$ for all x or $g'(x) < 0$ for all x. Then $Y = g(X)$ is also a r.v. of the continuous type with PDF given by

$$h(y) = \begin{cases} f[g^{-1}(y)] \left| \frac{d}{dy} g^{-1}(y) \right|, & \alpha < y < \beta; \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

where $\alpha = \min\{g(-\infty), g(+\infty)\}$ and $\beta = \max\{g(-\infty), g(+\infty)\}$.

Remark 5.4.
The key to computation of the induced distribution of $Y = g(X)$ from the distribution of X is (1). If the conditions of the above Theorem 5.3 are satisfied, we are able to identify the set $\{X \in g^{-1}(-\infty, y]\}$ as $\{X \leq g^{-1}(y)\}$ or $\{X \geq g^{-1}(y)\}$, according to whether g is increasing or decreasing. In practice Theorem 5.3 is quite useful, but whenever the conditions are violated one should return to (1) to compute the induced distribution.

Remark 5.5.
If the PDF f of X vanishes outside an interval $[a, b]$ of finite length, we need only to assume that g is differentiable in (a, b) and either $g'(x) > 0$ or $g'(x) < 0$ throughout the interval. Then we take

$$\alpha = \min\{g(a), g(b)\} \quad \text{and} \quad \beta = \max\{g(a), g(b)\}$$

in Theorem 5.3.

Example 5.4.
Let X have the density $f(x) = 1, 0 < x < 1$, and $= 0$ otherwise. Let $Y = e^X$. Then $X = \log Y$, and we have

$$h(y) = \left| \frac{1}{y} \right|, \quad 0 < \log y < 1,$$

that is,

$$h(y) = \begin{cases} \frac{1}{y}, & 1 < y < e; \\ 0, & \text{otherwise.} \end{cases}$$

If $y = -2 \log x$, then $x = e^{-y/2}$ and

$$h(y) = \left| -\frac{1}{2} e^{-y/2} \right|, \ 0 < e^{-y/2} < 1,$$

$$= \begin{cases} \frac{1}{2} e^{-y/2}, & 0 < y < \infty; \\ 0, & \text{otherwise.} \end{cases}$$

Example 5.5.
Let X be a nonnegative r.v. of the continuous type with PDF f , and $\alpha > 0$. Let $Y = X^\alpha$. Then

$$P\{X^\alpha \le y\} = \begin{cases} P\{X \le y^{1/\alpha}\} & \text{if } y \ge 0 \\ 0\ , & \text{if } y < 0. \end{cases}$$

The PDF of Y is given by

$$h(y) = f(y^{1/\alpha}) \left| \frac{d}{dy} y^{1/\alpha} \right|$$

$$= \begin{cases} \frac{1}{\alpha} y^{\frac{1}{\alpha}-1} f(y^{1/\alpha}) & \text{if } y \ge 0 \\ 0\ , & \text{if } y < 0. \end{cases}$$

Example 5.6.
Let X be a r.v. with PDF

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty.$$

Let $Y = X^2$. In this case, $g'(x) = 2x$ which is > 0 for $x > 0$, and < 0 for $x < 0$, so that the conditions of Theorem 5.3 are not satisfied. But for $y > 0$

$$P\{Y \le y\} = P\{-\sqrt{y} \le X \le \sqrt{y}\} = F(\sqrt{y}) - F(-\sqrt{y}),$$

where F is the DF of X. Thus the PDF of Y is given by

$$h(y) = \begin{cases} \frac{1}{2\sqrt{y}} \{f(\sqrt{y}) + f(-\sqrt{y})\}\ , & y > 0; \\ 0\ , & y \le 0. \end{cases}$$

Thus

$$h(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}} e^{-y/2}\ , & y > 0; \\ 0\ , & y \le 0. \end{cases}$$

Remark 5.6.
In the above example, the function $y = g(x)$ can be written as the sum of two monotone functions. We applied Theorem 5.3 to each of these monotonic summands. This example is a special case of the following result.

Theorem 5.4.
Let X be a r.v. of the continuous type with PDF f. Let $y = g(x)$ be differentiable for all x, and assume that $g'(x)$ is continuous and nonzero at all but a finite number of values of x. Then, for every real number y,
(a) *there exists a positive integer $n = n(y)$ and real numbers (inverses) $x_1(y), x_2(y), \ldots, x_n(y)$ such that*

$$g[x_k(y)] = y,\ g'[x_k(y)] \ne 0,\ k = 1, 2, \ldots, n(y),$$

or
(b) *there does not exist any x such that $g(x) = y$, $g'(x) = 0$, in which case we write $n(y) = 0$.*
Then Y is a continuous r.v. with PDF given by

$$h(y) = \begin{cases} \sum_{k=1}^{n} f[x_k(y)] |g'[x_k(y)]|^{-1} & \text{if } n > 0 \\ 0 & \text{if } n = 0 \end{cases}$$

Example 5.7.
Let X be a r.v. with PDF f , and let $Y = |X|$. Here $n(y) = 2$, $x_1(y) = y$, $x_2(y) = -y$ for $y > 0$, and

$$h(y) = \begin{cases} f(y) + f(-y) , & y > 0; \\ 0 , & y \leq 0. \end{cases}$$

Thus, if $f(x) = 1/2$, $-1 \leq x \leq 1$, and $= 0$ otherwise, then

$$h(y) = \begin{cases} 1 , & 0 \leq y \leq 1; \\ 0 , & \text{otherwise.} \end{cases}$$

If $f(x) = (1/\sqrt{2\pi})e^{-(x^2/2)}$, $-\infty < x < \infty$, then

$$h(y) = \begin{cases} \frac{2}{\sqrt{2\pi}} e^{-(y^2/2)} , & y > 0; \\ 0 , & \text{otherwise.} \end{cases}$$

6 Moments of a random variable

The study of probability distributions of a random variable is essentially the study of some numerical characteristics associated with them. These so-called parameters of the distribution play a key role in mathematical statistics. In this section we introduce some of these parameters, namely, moments , and investigate their properties.

Definition 6.1.
Let X be a random variable of the discrete type with probability mass function $p_k = P\{X = x_k\}, k = 1, 2, \ldots$ If

$$\sum_{k=1}^{\infty} |x_k| p_k < \infty, \tag{1}$$

we say that the expected value (or the mean or the mathematical expectation) of X exists and write

$$\mu = E(X) = \sum_{k=1}^{\infty} x_k p_k. \tag{2}$$

Note that the series $\sum_{k=1}^{\infty} x_k p_k$ may converge but the series $\sum_{k=1}^{\infty} |x_k| p_k$ may not. In that case we say that $E(X)$ does not exist.

Example 6.1.
Let X have the PMF given by

$$p_j = P\left\{X = (-1)^{j+1} \frac{3^j}{j}\right\} = \frac{2}{3^j}, j = 1, 2, \ldots$$

Then

$$\sum_{j=1}^{\infty} |x_j| p_j = \sum_{j=1}^{\infty} \frac{2}{j} = \infty,$$

and $E(X)$ does not exist, although the series

$$\sum_{j=1}^{\infty} x_j p_j = \sum_{j=1}^{\infty} (-1)^{j+1} \frac{2}{j}$$

is convergent.

Definition 6.2.
If X is of the continuous type and has PDF f , we say that $E(X)$ exists and

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx,$$

provided that

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty.$$

Remark 6.1.
A similar definition is given for the mean of any Borel-measurable function $h(X)$ of X. Thus, if X is of the continuous type and has PDF f , we say that $E(h(X))$ exists and equals $\int_{-\infty}^{\infty} h(x) f(x) dx$, provided that

$$\int_{-\infty}^{\infty} |h(x)| f(x) dx < \infty$$

We emphasize that the condition $\int |x| f(x) dx < \infty$ must be checked before it can be concluded that $E(X)$ exists and equals $\int x f(x) dx$. Moreover, it is worthwhile to recall at this point that the integral $\int_{-\infty}^{\infty} \varphi(x) dx$ exists, provided that the limit $\lim_{b\to\infty} \lim_{a\to\infty} \int_{-b}^{a} \varphi(x) dx$ exists. It is quite possible for the limit $\lim_{a\to\infty} \int_{-a}^{a} \varphi(x) dx$ to exist without the existence of $\int_{-\infty}^{\infty} \varphi(x) dx$.

Example 6.2.
Consider the following PDF

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}, -\infty < x < \infty.$$

Clearly

$$\lim_{a\to\infty} \int_{-a}^{a} \frac{x}{\pi} \frac{1}{1+x^2} dx = 0.$$

However, $E(X)$ does not exist since the integral $(1/\pi) \int_{-\infty}^{\infty} |x|/(1+x^2) dx$ diverges.

Remark 6.2.
If a and b are constants and X is a r.v. with $E|X| < \infty$, then $E|aX + b| < \infty$ and $E\{aX + b\} = aE(X) + b$. In particular, when $\mu = E(X)$, then $E(X - \mu) = 0$, a fact that should not come as a surprise.

Remark 6.3.
If X is bounded, that is, if $P\{|X| < M\} = 1$, $0 < M < \infty$, then $E(X)$ exists.

Remark 6.4.
If $P\{X > 0\} = 1$, and $E(X)$ exists, then $E(X) > 0$.

Theorem 6.1.
Let X be a r.v., and $\{x_1, x_2, \dots, \}$ values of X and g be a Borel-measurable function on $\mathbb{R}$. Let $Y = g(X)$. If X is of discrete type then

$$E(Y) = \sum_{j=1}^{\infty} g(x_j)P(X = x_j) \tag{3}$$

in the sense that, if either side of (3) exists, so does the other, and then the two are equal. If X is of continuous type with PDF f then $E(Y) = \int_{-\infty}^{+\infty} g(x)f(x)dx$ in the sense that, if either of the two integrals converges absolutely, so does the other, and the two are equal.

6.1 Properties of expectation

Theorem 6.2.

1. *If $X \geq 0$ then $E(X) \geq 0$.*
2. *If $X \geq 0$ and $E(X) = 0$ then $P(X = 0) = 1$.*
3. *If a and b are constants then $E(a + bX) = a + bE(X)$.*
4. *For any random variables X, Y then $E(X + Y) = E(X) + E(Y)$.*
   Properties 3 and 4 show that E is a linear operator.

Theorem 6.3.
For any random variables $X_1, X_2, \dots, X_n$, for which all the following expectations exist,

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

Definition 6.3.
The quantity $E(X^n), n \geq 1$, is called the nth moment of X. We note that

$$E(X^n) = \begin{cases} \sum_{x:P(x)>0} x^n P(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x^n f(x)dx, & \text{if } X \text{ is continuous.} \end{cases}$$

The expected value of a random variable X, $E(X)$, is also referred to as the first moment of X.

Definition 6.4.
Let k be a positive integer and c be a constant. If $E(X-c)^k$ exists, we call it the moment of order k about the point c. If we take $c = E(X) = \mu$, which exists since $E|X| < \infty$, we call $E(X-\mu)^k$ the central moment of order k or the moment of order k about the mean. We shall write

$$\mu_k = E(X-\mu)^k.$$

6.2 Variance

Definition 6.5.
If $E(X^2)$ exists, we call $E(X-\mu)^2$ the variance of X, and we write

$$\sigma^2 = var(X) = E(X-\mu)^2.$$

(which we below show $var(X) = E(X^2) - (E(X))^2$). So, the variance is the central moment of order 2. The quantity σ is called the standard deviation (SD) of X.

Theorem 6.4. ***(Properties of variance).***
(i) *$var(X) \geq 0$. If $var(X) = 0$, then $P(X = E(X)) = 1$.*
(ii) *If a, b are constants, $var(a+bX) = b^2 var(X)$.*
(iii) *$var(X) = E(X^2) - (E(X))^2$.*

Proof
(i) From Theorem 6.2, properties 1 and 2.
(ii) $var(a+bX) = E\left((a+bX-a-bE(X))^2\right) = b^2 E\left((X-E(X))^2\right) = b^2 var(X)$
(iii) $E\left((X-E(X))^2\right) = E\left(X^2 - 2XE(X) + (E(X))^2\right) = E(X^2) - (E(X))^2$.

■

Example 6.3.

1. Find $E(X)$ where X is the outcome when we roll a fair die.
2. Calculate $var(X)$

Solution.
1. Since $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$, we obtain
$E(X) = 1(\frac{1}{6}) + 2(\frac{1}{6}) + 3(\frac{1}{6}) + 4(\frac{1}{6}) + 5(\frac{1}{6}) + 6(\frac{1}{6}) = \frac{7}{2}$

2. $E(X^2) = 1^2(\frac{1}{6}) + 2^2(\frac{1}{6}) + 3^2(\frac{1}{6}) + 4^2(\frac{1}{6}) + 5^2(\frac{1}{6}) + 6^2(\frac{1}{6}) = \frac{1}{6}(91)$,
$var(X) = \frac{35}{12}$

7 Some moment inequalities

In this section, we derive some important inequalities. These inequalities use the mean, and possibly the variance, of a random variable to draw conclusions on the probabilities of certain events. They are primarily useful in situations where the mean and variance of a random variable X are easily computable, but the distribution of X is either unavailable or hard to calculate.

7.1 Markov's inequality

Theorem 7.1.
If X is a random variable with $E|X| < \infty$ and $a > 0$, then

$$P(|X| \geq a) \leq \frac{E|X|}{a}.$$

Proof
$I[|X| \geq a] \leq |X|/a$ (as the left-hand side is 0 or 1, and if 1 then the right-hand side is at least 1). So

$$P(|X| \geq a) = E\left(I[|X| \geq a]\right) \leq E\left(|X|/a\right) = \frac{E|X|}{a}$$

■

Corollary 7.1.
If a random variable X can only take nonnegative values, then

$$P(X \geq a) \leq \frac{E(X)}{a}, \quad \textit{for all} \quad a > 0$$

Remark 7.1.
Loosely speaking, it asserts that if a nonnegative random variable has a small mean, then the probability that it takes a large value must also be small.

Example 7.1.
Let X be uniformly distributed on the interval $[0, 4]$ and note that $E(X) = 2$. Then, the Markov inequality asserts that

$$P(X \geq 2) \leq \frac{2}{2} = 1, \quad P(X \geq 3) \leq \frac{2}{3} = 0.67, \quad P(X \geq 4) \leq \frac{2}{4} = 0.5.$$

By comparing with the exact probabilities

$$P(X \geq 2) = 0.5, \quad P(X \geq 3) = 0.25, \quad P(X \geq 4) = 0,$$

we see that the bounds provided by the Markov inequality can be quite loose.

7.2 Chebyshev's Inequality

We continue with the Chebyshev inequality. Loosely speaking, it asserts that if the variance of a random variable is small, then the probability that it takes a value far from its mean is also small. Note that the Chebyshev inequality does not require the random variable to be nonnegative.

Theorem 7.2.
If X is a random variable with $E(X^2) < \infty$ and $\varepsilon > 0$, then

$$P(|X| \geq \varepsilon) \leq \frac{E(X^2)}{\varepsilon^2}.$$

Proof
Similarly to the proof of the Markov inequality,

$$I[|X| \geq \varepsilon] \leq \frac{|X|^2}{\varepsilon^2}$$

■

Corollary 7.2.
If X is a random variable with mean μ and variance σ^2, then

$$P\left(|X - \mu| \geq c\right) \leq \frac{\sigma^2}{c^2}, \quad \textit{for all} \quad c > 0.$$

Proof
To justify the Chebyshev's inequality, we consider the nonnegative random variable $(X - \mu)^2$ and apply the Markov inequality with $a = c^2$. We obtain

$$P\left((X - \mu)^2 \geq c^2\right) \leq \frac{E((X - \mu)^2)}{c^2} = \frac{\sigma^2}{c^2},$$

The derivation is completed by observing that the event $(X - \mu)^2 \geq c^2$ is identical to the event $|X - \mu| \geq c$ and

$$P\left(|X - \mu| \geq c\right) = P\left((X - \mu)^2 \geq c^2\right) \leq \frac{\sigma^2}{c^2}.$$

■

Remark 7.2.
The Chebyshev's inequality is generally more powerful than the Markov's inequality (the bounds that it provides are more accurate), because it also makes use of information on the variance of X. Still, the mean and the variance of a random variable are only a rough summary of the properties of its distribution, and we cannot expect the bounds to be close approximations of the exact probabilities.

Example 7.2.
As in the above example, let X be uniformly distributed on $[0, 4]$. Let us use the Chebyshev inequality to bound the probability that $|X - 2| \geq 1$. We have $\sigma^2 = 16/12 = 4/3$, and

$$P\left(|X - 2| \geq 1\right) \leq \frac{4}{3},$$

which is not particularly informative.

7.3 Jensen's Inequality

Jensen's Inequality gives a lower bound on expectations of convex functions. Recall that a function g is convex if, for $0 < \lambda < 1$,

$$g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y)$$

for all x and y. Alternatively, if the derivatives are well defined, function g is convex if

$$\frac{d^2}{dt^2}\{g(t)\}_{t=x} = g^{(2)}(x) \geq 0.$$

Conversely, g is concave if $-g$ is convex.

Theorem 7.3.
Suppose that X is a random variable with expectation μ , and function g is convex and finite. Then

$$E(g(X)) \geq g(E(X))$$

with equality if and only if, it exists at least $a + bx$ that is a tangent to g at μ

$$P_X\left(g(X) = a + bX\right) = 1.$$

that is, $g(x)$ is linear.

Proof
Let $l(x) = a + bx$ be the equation of the tangent at $x = \mu$. Then, for each x, $g(x) \geq a + bx$ Thus,

$$E(g(X)) \geq E(a + bX) = a + bE(X) = l(\mu) = g(\mu) = g(E(X))$$

as required. Also, if $g(x)$ is linear, then equality follows by properties of expectations. Suppose that

$$E(g(X)) = g(E(X)) = g(\mu)$$

so, $g(x)$ is convex, but not linear. Let $l(x) = a + bx$ be the tangent to g at μ. Then by convexity

$$g(x) - l(x) > 0 \Rightarrow \int (g(x) - l(x))dF_X(x) = \int g(x)dF_X(x) - \int l(x)dF_X(x) > 0$$

and hence

$$E(g(X)) > E(l(X)).$$

But $l(x)$ is linear, so $E(l(X)) = a + bE(X) = g(\mu)$, yielding the contradiction

$$E(g(X)) > g(E(X)).$$

and the result follows. ■

Example 7.3.

- $g(x) = x^2$ is convex, thus $E(X^2])geq \{E(X)\}^2$
- $g(x) = \log x$ is concave, thus $E(\log X) \leq \log(E(X))$

8 Some exercises

8.1

Find the range for each of the following random variables. The range of X is the set of possible values of X.

1. I toss a coin 100 times. Let X be the number of heads I observe.
2. I toss a coin until the first heads appears. Let Y be the total number of coin tosses.
3. The random variable T is defined as the time (in hours) from now until the next earthquake occurs in a certain city.

Solution.
1. The random variable X can take any integer from 0 to 100, so $R_X = \{0, 1, 2, \ldots, 100\}$.

2. The random variable Y can take any positive integer, so $R_Y = \{1, 2, 3, \ldots, \} = \mathbb{N}^*$.

3. The random variable T can in theory get any positive real number, so $R_T = [0, \infty)$.

8.2

I toss a fair coin twice, and let X be defined as the number of heads I observe. Find the range of X, R_X, as well as its probability mass function P_X. Find the CDF of X.

Solution.
Here, the sample space is given by

$$\Omega = \{HH, HT, TH, TT\}.$$

The number of heads will be 0, 1 or 2. Thus

$$R_X = \{0, 1, 2\}.$$

Since this is a finite (and thus a countable) set, the random variable X is a discrete random variable. Next, we need to find PMF of X. The PMF is defined as

$$P_X(k) = P(X = k) \quad \text{for} \quad k = 0, 1, 2.$$

We have

$$P_X(0) = P(X = 0) = P(\{TT\}) = \frac{1}{4},$$

$$P_X(1) = P(X = 1) = P(\{HT, TH\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

$$P_X(2) = P(X = 2) = P(\{HH\}) = \frac{1}{4}.$$
$$P_X(0) + P_X(1) + P_X(2) = 1$$

To find the CDF, we argue as follows. First, note that if $x < 0$, then

$$F_X(x) = P(X \leq x) = 0, \quad \text{for} \quad x < 0.$$

Next, if $x \geq 2$,

$$F_X(x) = P(X \leq x) = 1, \quad \text{for} \quad x \geq 2.$$

Next, if $0 \leq x < 1$,

$$F_X(x) = P(X \leq x) = P(X = 0) = \frac{1}{4}, \quad \text{for} \quad 0 \leq x < 1.$$

Finally, if $1 \leq x < 2$,

$$F_X(x) = P(X \leq x) = P(X = 0) + P(X = 1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}, \quad \text{for} \quad 1 \leq x < 2.$$

Thus, to summarize, we have

$$F_X(x) = \begin{cases} 0, & \text{for} \quad x < 0 \\ \frac{1}{4}, & \text{for} \quad 0 \leq x < 1 \\ \frac{3}{4}, & \text{for} \quad 1 \leq x < 2 \\ 1, & \text{for} \quad x \geq 2 \end{cases}$$

8.3

I have an unfair coin for which $P(H) = p$, where $0 < p < 1$. I toss the coin repeatedly until I observe a heads for the first time. Let Y be the total number of coin tosses. Find the distribution of Y .

Solution.
First, we note that the random variable Y can potentially take any positive integer, so we have $R_Y = \mathbb{N}^* = \{1, 2, 3, \ldots\}$. To find the distribution of Y , we need to find $P_Y(k) = P(Y = k)$ for $k = 1, 2, 3, \ldots$. We have

$$P_Y(1) = P(Y = 1) = P(\{H\}) = p,$$
$$P_Y(2) = P(Y = 2) = P(\{TH\}) = (1 - p)p,$$
$$P_Y(3) = P(Y = 3) = P(\{TTH\}) = (1 - p)^2 p,$$
$$\ldots$$
$$P_Y(k) = P(Y = k) = P(\{TT \ldots TH\}) = (1 - p)^{k-1} p$$

such that T appears $k - 1$ times. Thus, we can write the PMF of Y in the following way

$$P_Y(y) = \begin{cases} (1 - p)^{y-1} p, & \text{for} \quad y = 1, 2, 3, \ldots \\ 0, & \text{otherwise.} \end{cases}$$

8.4

Let X be a discrete random variable with the following PMF

$$P_X(k) = \begin{cases} 0.1, & \text{for } k = 0 \\ 0.4, & \text{for } k = 1 \\ 0.3, & \text{for } k = 2 \\ 0.2, & \text{for } k = 3 \\ 0, & \text{otherwise.} \end{cases}$$

1. Find $E(X)$.
2. Find $var(X)$.
3. If $Y = (X-2)^2$, find $E(Y)$.

Solution.
1.

$$E(X) = \sum_{x_k \in R_X} x_k P_X(x_k) = 0(0.1) + 1(0.4) + 2(0.3) + 3(0.2) = 1.6$$

2. We can use $var(X) = E(X^2) - (E(X))^2 = E(X^2) - (1.6)^2$ we have

$$E(X^2) = 0^2(0.1) + 1^2(0.4) + 2^2(0.3) + 3^2(0.2) = 3.4$$

Thus, we have

$$var(X) = (3.4) - (1.6)^2 = 0.84$$

3. We have

$$E[(X-2)^2] = (0-2)^2(0.1) + (1-2)^2(0.4) + (2-2)^2(0.3) + (3-2)^2(0.2) = 1.$$

8.5

Let X be a continuous random variable with the following PDF

$$f_X(x) = \begin{cases} ce^{-x}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

where c is a positive constant.

1. Find c.
2. Find the CDF of X, $F_X(x)$.
3. Find $P(1 < X < 3)$.

Solution.
1. To find c, we can use the following property of density

$$1 = \int_{-\infty}^{\infty} f_X(u)du = \int_0^{\infty} ce^{-u}du = c\left[-e^{-x}\right]_0^{\infty} = c$$

Thus, we must have $c = 1$.

2. To find the CDF of X, we use $F_X(x) = \int_{-\infty}^{x} f_X(u)du$, so for $x < 0$, we obtain $F_X(x) = 0$. For $x \geq 0$, we have

$$F_X(x) = \int_0^x e^{-u}du = 1 - e^{-x}.$$

Thus,

$$F_X(x) = \begin{cases} 1 - e^{-x}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

3. We can find $P(1 < X < 3)$ using either the CDF or the PDF. If we use the CDF, we have

$$P(1 < X < 3) = F_X(3) - F_X(1) = [1 - e^{-3}] - [1 - e^{-1}] = e^{-1} - e^{-3}.$$

Equivalently, we can use the PDF. We have

$$P(1 < X < 3) = \int_1^3 f_X(t)dt = \int_1^3 e^{-t}dt = e^{-1} - e^{-3}.$$

8.6

Let X be a continuous random variable with PDF

$$f_X(x) = \begin{cases} \frac{3}{x^4}, & x \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find the mean and variance of X.

Solution.

$$E(X) = \int_{-\infty}^{\infty} xf_X(x)dx = \int_1^{\infty} \frac{3}{x^3}dx = \frac{3}{2}$$

Next, we find $E(X^2)$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_1^{\infty} \frac{3}{x^2}dx = 3$$

Thus, we have

$$var(X) = E(X^2) - (E(X))^2 = 3 - \frac{9}{4} = \frac{3}{4}.$$

8.7

Let X a continuous random variable with PDF

$$f_X(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & \text{otherwise.} \end{cases}$$

and let $Y = e^X$.

1. Find the CDF of Y .
2. Find the PDF of Y .
3. Find $E(Y)$.

Solution.
We deduce easily the CDF of X.

$$F_X(x) = \begin{cases} 0, & \text{for } x < 0 \\ x, & \text{for } x \in [0,1] \\ 1, & \text{for } x > 1. \end{cases}$$

It is a good idea to think about the range of Y before finding the distribution. Since e^x is an increasing function of x and $R_X = [0,1]$, we conclude that $R_Y = [1,e]$. So we immediately know that

$$F_Y(y) = P(Y \leq y) = 0, \text{for } y < 1,$$

$$F_Y(y) = P(Y \leq y) = 1, \text{for } y \geq e.$$

1. To find $F_Y(y)$ for $y \in [1,e]$, we can write

$$F_Y(y) = P(Y \leq y) = P(e^X \leq y) = P(X \leq \ln y) = F_X(\ln y) = \ln y$$

To summarize

$$F_Y(y) = \begin{cases} 0, & \text{for } y < 1 \\ \ln x, & \text{for } y \in [1,e] \\ 1, & \text{for } y > e. \end{cases}$$

2. The above CDF is a continuous function, so we can obtain the PDF of Y by taking its derivative. We have

$$f_Y(y) = F'_Y(y) = \begin{cases} \frac{1}{y}, & \text{for } y \in [1,e] \\ 0, & \text{otherwise.} \end{cases}$$

3.

$$E(Y) = E(e^X) = \int_{-\infty}^{\infty} e^x f_X(x) dx = \int_0^1 e^x dx = e - 1.$$

we could also find E(Y) using the PDF of Y ,

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_1^e y \frac{1}{y} dy = e - 1$$

8.8

Let X be a random variable with PDF given by

$$f_X(x) = F'_X(x) = \begin{cases} cx^2, & \text{for } x \in [-1,1] \\ 0, & \text{otherwise.} \end{cases}$$

1. Find the constant c.
2. Find $E(X)$ and $var(X)$.
3. Find $P(X \geq \frac{1}{2})$.

Solution.
1. To find c, we can use $\int_{-\infty}^{\infty} f_X(u)du = 1$

$$1 = \int_{-\infty}^{\infty} f_X(u)du = \int_{-1}^{1} cu^2 du = \frac{2}{3}c.$$

Thus, we must have $c = \frac{3}{2}$.

2. To find $E(X)$, we can write

$$E(X) = \int_{-1}^{1} uf_X(u)du = \frac{3}{2}\int_{-1}^{1} u^3 du = 0.$$

$$var(X) = E(X^2) - (E(X))^2 = E(X^2) = \int_{-1}^{1} u^2 f_X(u)du = \frac{3}{2}\int_{-1}^{1} u^4 du = \frac{3}{5}.$$

3. To find $P(X \geq \frac{1}{2})$, we can write

$$P(X \geq \frac{1}{2}) = \frac{3}{2}\int_{\frac{1}{2}}^{1} x^2 dx = \frac{7}{16}.$$

8.9

Let X be a discrete random variable with the following PMF,

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

Using Markov's inequality, find an upper bound on $P(X \geq \alpha n)$, where $p < \alpha < 1$. Evaluate the bound for $p = \frac{1}{2}$ and $\alpha = \frac{3}{4}$.

Solution.
Note that X is a nonnegative random variable and $E(X) = np$. Applying Markov's inequality, we obtain

$$P(X \geq \alpha n) \leq \frac{E(X)}{\alpha n} = \frac{pn}{\alpha n} = \frac{p}{\alpha}.$$

For $p = \frac{1}{2}$ and $\alpha = \frac{3}{4}$, we obtain

$$P(X \geq \frac{3n}{4}) \leq \frac{2}{3}.$$

8.10

Let X be a discrete random variable with the following PMF,

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

Using Chebyshev's inequality, find an upper bound on $P(X \geq \alpha n)$, where $p < \alpha < 1$. Evaluate the bound for $p = \frac{1}{2}$ and $\alpha = \frac{3}{4}$.

Solution.
One way to obtain a bound is to write

$$P(X \geq \alpha n) = P(X - np \geq \alpha n - np) \leq P(|X - np| \geq n\alpha - np) \leq \frac{var(X)}{(n\alpha - np)^2} = \frac{p(1-p)}{n(\alpha - p)^2}.$$

For $p = \frac{1}{2}$ and $\alpha = \frac{3}{4}$, we obtain

$$P\left(X \geq \frac{3n}{4}\right) \leq \frac{4}{n}$$

The bound given by Markov is the "weakest" one. It is constant and does not change as n increases. The bound given by Chebyshev's inequality is "stronger" than the one given by Markov's inequality. In particular, note that $\frac{4}{n}$ goes to zero as n goes to infinity.

8.11

Let X be a continuous random variable with the following PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Using Markov's inequality find an upper bound for $P(X \geq a)$. Compare the upper bound with the actual value of $P(X \geq a)$.

Solution.
We find easily that $E(X) = \frac{1}{\lambda}$. Using Markov's inequality, $P(X \geq a) \leq \frac{E(X)}{a} = \frac{1}{\lambda a}$. The actual value of $P(X \geq a)$ is $e^{-\lambda a}$, and we always have $\frac{1}{\lambda a} \geq e^{-\lambda a}$.

8.12

Let X be a continuous random variable with the following PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Using Chebyshev's inequality find an upper bound for $P(|X - E(X)| \geq b)$.

Solution.
We have $E(X) = \frac{1}{\lambda}$ and $var(X) = \frac{1}{\lambda^2}$. Using Chebyshev's inequality, we have

$$P(|X - E(X)| \geq b) \leq \frac{var(X)}{b^2} = \frac{1}{\lambda^2 b^2}.$$

8.13

Let X be a positive random variable with $E(X) = 10$. What can you say about the following quantities?

1. $E\left(\frac{1}{X+1}\right)$
2. $E\left(e^{\frac{1}{X+1}}\right)$
3. $E(\ln\sqrt{X})$

Solution.
1.

$$g(x) = \frac{1}{x+1}, \quad g''(x) = \frac{2}{(1+x)^3} > 0 \quad \text{for} \quad x > 0$$

Thus g is convex on $(0, \infty)$

$$E\left(\frac{1}{X+1}\right) > \frac{1}{1+E(X)}, \text{(Jensen's inequality)}$$

$$= \frac{1}{1+10} = \frac{1}{11}$$

2. If we let $h(x) = e^x$, $g(x) = \frac{1}{x+1}$ then h is convex and non-decreasing and g is convex thus, $e^{\frac{1}{x+1}}$ is a convex function, thus

$$E\left(e^{\frac{1}{X+1}}\right) \geq e^{\frac{1}{1+E(X)}} = e^{\frac{1}{11}} \text{(by Jensen's inequality)}$$

3. If $g(x) = \ln\sqrt{x} = \frac{1}{2}$, then $g'(x) = \frac{1}{2x}$ for $x > 0$ and $g''(x) = -\frac{1}{2x^2}$. Thus g is concave on $(0, \infty)$. We conclude

$$E(\ln\sqrt{X}) = E\left(\frac{1}{2}\ln X\right) \leq \frac{1}{2}\ln E(X) = \frac{1}{2}\ln 10 \quad \text{(by Jensen's inequality)}$$

CHAPTER III. Some Special Distributions

1 Introduction.

In preceding chapters we studied probability distributions in general. In this chapter we will study some commonly occurring probability distributions and investigate their basic properties. As it turns out, there are some specific distributions that are used over and over in practice, thus they have been given special names. There is a random experiment behind each of these distributions. Since these random experiments model a lot of real life phenomenon, these special distributions are used frequently in different applications. That's why they have been given a name and we devote a section to study them. We will provide PMFs for all of these special random variables, but rather than trying to memorize the PMF, you should understand the random experiment behind each of them. If you understand the random experiments, you can simply derive the PMFs when you need them. Although it might seem that there are a lot of formulas in this section, there are in fact very few new concepts. We begin with some discrete distributions and follow with some continuous models.

2 Some discrete random variables

2.1 Bernoulli Distribution

The simplest discrete random variable is the Bernoulli distribution. A Bernoulli random variable is a random variable that can only take two possible values, usually 0 and 1. This random variable models random experiments that have two possible outcomes, sometimes referred to as "success" and "failure." Here are some examples:

- You take a pass-fail exam. You either pass (resulting in $X = 1$) or fail (resulting in $X = 0$).
- You toss a coin. The outcome is ether heads or tails.
- A child is born. The gender is either male or female.

Formally, the Bernoulli distribution is defined as follows:

Definition 2.1.
A random variable X is said to be a Bernoulli random variable with parameter p, shown as $X \sim Bernoulli(p)$, if its PMF is given by

$$P_X(x) = \begin{cases} p, & \text{for } x = 1 \\ 1-p, & \text{for } x = 0 \\ 0, & \text{otherwise.} \end{cases}$$

where $0 < p < 1$.

Remark 2.1.
A Bernoulli random variable is associated with a certain event A. If an event A occurs (for example, if you pass the test), then $X = 1$; otherwise $X = 0$. For this reason the Bernoulli random variable, is also called the indicator random variable. In particular, the indicator random variable I_A for an event A is defined by

$$I_A = \begin{cases} 1, & \text{if the event A occurs} \\ 0, & \text{otherwise.} \end{cases}$$

The indicator random variable for an event A has a Bernoulli distribution with parameter $p = P(A)$, so we can write

$$I_A \sim Bernoulli(P(A)).$$

Property 2.1.
If $X \sim Bernoulli(p)$ *then* $E(X) = p$ *and* $var(X) = p(1-p)$.

Example 2.1.
We have a biased coin with $P(H) = \frac{2}{3}$ and we toss it. We define the random variable X as a Bernoulli random variable associated with this coin toss, i.e., $X = 1$ if the result of the coin toss is heads and $X = 0$ otherwise. So $X \sim Bernoulli(\frac{2}{3})$, $E(X) = \frac{2}{3}$, $var(X) = \frac{2}{3}(1 - \frac{2}{3})$.

2.2 Binomial Distribution

The random experiment behind the binomial distribution is as follows. Suppose that we have a coin with $P(H) = p$. We toss the coin n times and define X to be the total number of heads that we observe. Then X is binomial with parameters n and p, and we write $X \sim Binomial(n, p)$. The range of X in this case is $R_X = \{0, 1, 2, \ldots, n\}$. The PMF of X in this case is given by the binomial formula

$$P_X(k) = C_n^k p^k (1-p)^{n-k}, \quad \text{for} \quad k = 0, 1, 2, \ldots, n.$$

We have the following definition:

Definition 2.2.
A random variable X is said to be a binomial random variable with parameters n and p, shown as $X \sim Binomial(n, p)$, if its PMF is given by

$$P_X(k) = \begin{cases} C_n^k p^k (1-p)^{n-k}, & \text{for} \quad k = 0, 1, 2, \ldots, n. \\ 0, & \text{otherwise.} \end{cases}$$

where $0 < p < 1$.

Property 2.2.
If $X \sim Binomial(n, p)$ *then* $E(X) = np$ *and* $var(X) = np(1-p)$.

Remark 2.2.
Binomial random variable as a sum of Bernoulli random variables. Here is a useful way of thinking about a binomial random variable. Note that a $Binomial(n,p)$ random variable can be obtained by n independent coin tosses. If we think of each coin toss as a $Bernoulli(p)$ random variable, the $Binomial(n,p)$ random variable is a sum of n independent $Bernoulli(p)$ random variables. This is stated more precisely in the following lemma.

Lemma 2.1.
If $X_1, X_2, \ldots, X_n$ are independent $Bernoulli(p)$ random variables, then the random variable X defined by $X = X_1 + X_2 + \ldots + X_n$ has a $Binomial(n,p)$ distribution.

To generate a random variable $X \sim Binomial(n,p)$, we can toss a coin n times and count the number of heads. Counting the number of heads is exactly the same as finding $X_1 + X_2 + \ldots + X_n$, where each X_i is equal to one if the corresponding coin toss results in heads and zero otherwise. This interpretation of binomial random variables is sometimes very helpful. Let's look at an example.

Example 2.2.
Let $X \sim Binomial(n,p)$ and $Y \sim Binomial(m,p)$ be two independent random variables. Define a new random variable as $Z = X + Y$. We find the PMF of Z.
Since $X \sim Binomial(n,p)$, we can think of X as the number of heads in n independent coin tosses, i.e., we can write

$$X = X_1 + X_2 + \ldots + X_n,$$

where the X_i's are independent $Bernoulli(p)$ random variables. Similarly, since $Y \sim Binomial(m,p)$, we can think of Y as the number of heads in m independent coin tosses, i.e., we can write

$$Y = Y_1 + Y_2 + \ldots + Y_m,$$

where the Y_j's are independent $Bernoulli(p)$ random variables. Thus, the random variable $Z = X + Y$ will be the total number of heads in $n + m$ independent coin tosses:

$$Z = X + Y = X_1 + X_2 + \ldots + X_n + Y_1 + Y_2 + \ldots + Y_m$$

where the X_i's and Y_j's are independent $Bernoulli(p)$ random variables. Thus, by Lemma 2.1, Z is a binomial random variable with parameters $m+n$ and p, i.e., $Binomial(m+n,p)$. Therefore, the PMF of Z is

$$P_Z(k) = \begin{cases} C_{m+n}^k p^k (1-p)^{m+n-k}, & \text{for } k = 0, 1, 2, \ldots, m+n. \\ 0, & \text{otherwise.} \end{cases}$$

2.3 Geometric Distribution

The random experiment behind the geometric distribution is as follows. Suppose that we have a coin with $P(H) = p$. We toss the coin until we observe the first heads. We define X as the total number of coin tosses in this experiment. Then X is said to have geometric distribution with parameter p. In other words, you can think of this experiment as repeating independent Bernoulli trials until observing the first success. The range of X here is $R_X = \{1, 2, 3, \ldots\}$. We have the following definition:

Definition 2.3.
A random variable X is said to be a geometric random variable with parameter p, shown as $X \sim Geometric(p)$, if its PMF is given by

$$P_X(k) = \begin{cases} p(1-p)^{k-1}, & \text{for } k = 1, 2, \ldots, . \\ 0, & \text{otherwise.} \end{cases}$$

where $0 < p < 1$.

Property 2.3.
If $X \sim Geometric(p)$ *then* $E(X) = \frac{1-p}{p}$ *and* $var(X) = \frac{(1-p)}{p^2}$.

Remark 2.3.
We should note that some books define geometric random variables slightly differently. They define the geometric random variable X as the total number of failures before observing the first success. By this definition the range of X is $R_X = \{0, 1, 2, \ldots\}$ and the PMF is given by

$$P_X(k) = \begin{cases} p(1-p)^{k}, & \text{for } k = 0, 1, 2, \ldots, . \\ 0, & \text{otherwise.} \end{cases}$$

Example 2.3.
if we toss a fair coin until heads is obtained, the expected number of tosses until the first head is 2 (so the expected number of tails is 1); and the variance of this number is also 2.

2.4 Negative Binomial (Pascal) Distribution

The negative binomial or Pascal distribution is a generalization of the geometric distribution. It relates to the random experiment of repeated independent trials until observing m successes. Here is how we define the Pascal distribution in this work. Suppose that we have a coin with $P(H) = p$. We toss the coin until we observe m heads, where $m \in \mathbb{N}$. We define X as the total number of coin tosses in this experiment. Then X is said to have Pascal distribution with parameter m and p. We write $X \sim Pascal(m, p)$. Note that $Pascal(1, p) = Geometric(p)$. Note that by our definition the range of X is given by $R_X = \{m, m+1, m+2, m+3, \ldots\}$.

Definition 2.4.
A random variable X is said to be a Pascal random variable with parameters m and p, shown as $X \sim Pascal(m,p)$, if its PMF is given by

$$P_X(k) = \begin{cases} C_{k-1}^{m-1} p^m (1-p)^{k-m}, & \text{for } k = m+1, m+2, \ldots, . \\ 0, & \text{otherwise.} \end{cases}$$

where $0 < p < 1$.

Property 2.4.
If $X \sim Pascal(m,p)$ *then* $E(X) = \frac{m}{p}$ *and* $var(X) = \frac{m(1-p)}{p^2}$.

Example 2.4.
Suppose that we toss a biased coin until I observe 5 heads, such that $P(H) = \frac{2}{3}$, and X is defined as the total number of coin tosses in this experiment. $X \sim Pascal(5, \frac{2}{3})$.

2.5 Hypergeometric Distribution

We consider the random experiment behind the hypergeometric distribution. We have a bag that contains b blue marbles and r red marbles. We choose $k \leq b + r$ marbles at random (without replacement). Let X be the number of blue marbles in our sample. By this definition, we have $X \leq \min(k, b)$. Also, the number of red marbles in our sample must be less than or equal to r, so we conclude $X \geq \max(0, k - r)$. Therefore, the range of X is given by

$$R_X = \{\max(0,k-r), \max(0,k-r)+1, \max(0,k-r)+2, \ldots, \min(k,b)\}.$$

The following definition summarizes the discussion above.

Definition 2.5.
A random variable X is said to be a Hypergeometric random variable with parameters b, r and k, shown as $X \sim Hypergeometric(b,r,k)$, if its range is $R_X = \{\max(0,k-r), \max(0,k-r)+1, \max(0,k-r)+2, \ldots, \min(k,b)\}$ and its PMF is given by

$$P_X(x) = \begin{cases} \dfrac{C_b^x C_r^{k-x}}{C_{b+r}^k}, & \text{for } x \in R_X. \\ 0, & \text{otherwise.} \end{cases}$$

Property 2.5.
If $X \sim Hypergeometric(b,r,k)$ *then* $E(X) = \dfrac{kb}{b+r}$ *and* $var(X) = \dfrac{kbr}{(b+r)^2} \frac{b+r-k}{b+r-1}$.

Example 2.5.
Consider an urn with n_1 red balls and n_2 black balls. Suppose n balls are drawn without replacement, $n \leq n_1 + n_2$. The probability of drawing exactly k red balls is given by the hypergeometric distribution.

2.6 Poisson Distribution

The Poisson distribution is one of the most widely used probability distributions. It is usually used in scenarios where we are counting the occurrences of certain events in an interval of time or space. In practice, it is often an approximation of a real-life random variable. Here is an example of a scenario where a Poisson random variable might be used. Suppose that we are counting the number of customers who visit a certain store from 13h00 to 14h00. Based on data from previous days, we know that on average $\lambda = 15$ customers visit the store. Of course, there will be more customers some days and fewer on others. Here, we may model the random variable X showing the number customers as a Poisson random variable with parameter $\lambda = 15$. Let us introduce the Poisson PMF.

Definition 2.6.
A random variable X is said to be a Poisson random variable with parameter *lambda*, shown as $X \sim Poisson(\lambda)$, if its range is $R_X = \{0, 1, 2, 3, \ldots\}$ and its PMF is given by

$$P_X(k) = \begin{cases} \frac{e^{-\lambda}\lambda^k}{k!}, & \text{for } k \in R_X. \\ 0, & \text{otherwise.} \end{cases}$$

Property 2.6.
If $X \sim Poisson(\lambda)$ *then* $E(X) = \lambda$ *and* $var(X) = \lambda$.

Example 2.6.
The number of emails that I get in a weekday can be modeled by a Poisson distribution with an average of 0.2 emails per minute.

1. What is the probability that I get no emails in an interval of length 5 minutes?
2. What is the probability that I get more than 3 emails in an interval of length 10 minutes?

Solution.

1. Let X be the number of emails that I get in the 5-minute interval. Then, by the assumption X is a Poisson random variable with parameter $\lambda = 5(0.2) = 1$,
$$P(X = 0) = P_X(0) = \frac{e^{-\lambda}\lambda^0}{0!} = 0.3679$$
2. Let Y be the number of emails that I get in the 10-minute interval. Then by the assumption Y is a Poisson random variable with parameter $\lambda = 10(0.2) = 2$,
$$P(Y > 3) = 1 - P(Y \leq 3) = 1 - (P_Y(0) + P_Y(1) + P_Y(2) + P_Y(3))$$
$$= 1 - e^{-\lambda} - \frac{e^{-\lambda}\lambda^1}{1!} - \frac{e^{-\lambda}\lambda^2}{2!} - \frac{e^{-\lambda}\lambda^3}{3!} = 0.1429$$

2.7 Multinomial Distribution

The binomial distribution is generalized in the following natural fashion. Suppose that an experiment is repeated n times. Each replication of the experiment terminates in one of k mutually exclusive and exhaustive events $A_1, A_2, \ldots, A_k$. Let p_j be the probability that the experiment terminates in A_j, $j = 1, 2, \ldots, k$, and suppose that p_j $(j = 1, 2, \ldots, k)$ remains constant for all n replications. We assume that the n replications are independent.

Let $x_1, x_2, \ldots, x_{k-1}$ be nonnegative integers such that $x_1 + x_2 + \ldots + x_{k-1} \leq n$. Then the probability that exactly x_i trials terminate in $A_i, i = 1, 2, \ldots, k-1$ and hence that $x_k = n - (x_1 + x_2 + \ldots + x_{k-1})$ trials terminate in A_k is clearly.
$$\frac{n!}{x_1!x_2!\ldots x_k!}p_1^{x_1}p_2^{x_2}\ldots p_k^{x_k}.$$
If $(X_1, X_2, \ldots, X_k)$ is a random vector such that $X_j = x_j$ means that event A_j has occurred x_j times, $x_j = 0, 1, 2, \ldots, n$, the joint PMF of $(X_1, X_2, \ldots, X_k)$ is given by
$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$$
$$= \begin{cases} \dfrac{n!}{x_1!x_2!\ldots x_k!}p_1^{x_1}p_2^{x_2}\ldots p_k^{x_k}, & \text{if } n = \sum_1^k x_i. \\ 0, & \text{otherwise.} \end{cases}$$

Definition 2.7.
A random vector $(X_1, X_2, \ldots, X_k)$ with joint PMF given by
$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$$
$$= \begin{cases} \dfrac{n!}{x_1!x_2!\ldots x_k!}p_1^{x_1}p_2^{x_2}\ldots p_k^{x_k}, & \text{if } n = \sum_1^k x_i. \\ 0, & \text{otherwise.} \end{cases}$$
is said to have a multinomial distribution. We write $X \sim M_k(n; p_1, \ldots, p_k)$.

Property 2.7.
If $X \sim M_k(n; p_1, \ldots, p_k)$ *then* $E(X_j) = np_j$ *and* $var(X_j) = np_j(1 - p_j)$.

Example 2.7.
Suppose n balls are tossed independently into k boxes such that the probability that a given ball goes in box i is p_i. The probability that there will be $n_1, \ldots, n_k$ balls in boxes $1, \ldots, k$, respectively, is

$$\frac{n!}{n_1!n_2!\ldots n_k!}p_1^{n_1}p_2^{n_2}\ldots p_k^{n_k}, \quad n_1 + \cdots + n_k = n.$$

3 Some continuous random variables

Remember that discrete random variables can take only a countable number of possible values. On the other hand, a continuous random variable X has a range in the form of an interval or a union of non-overlapping intervals on the real line (possibly the whole real line). Also, for any $x \in \mathbb{R}$, $P(X = x) = 0$. Thus, we need to develop new tools to deal with continuous random variables. The theory of continuous random variables is completely analogous to the theory of discrete random variables. Indeed, if we want to oversimplify things, we might say the following: take any formula about discrete random variables, and then replace sums with integrals, and replace PMFs with probability density functions (PDFs), and you will get the corresponding formula for continuous random variables. Of course, there is a little bit more to the story and that's why we need a section to discuss it. In this section, we will introduce usual continuous variables.

3.1 Uniform Distribution

We choose a real number uniformly at random in the interval $[a, b]$, and call it X. By uniformly at random, we mean all intervals in $[a, b]$ that have the same length must have the same probability. We have the following definition:

Definition 3.1.
A continuous random variable X is said to have a Uniform distribution over the interval $[a, b]$, shown as $X \sim Uniform(a, b)$, if its PDF is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } \ a \leq x \leq b, \\ 0, & x < a \ \text{ or } \ x > b. \end{cases}$$

Property 3.1.
If $X \sim Uniform(a,b)$, then its CDF is given by

$$F_X(x) = \begin{cases} 0, & for \;\; x < a \\ \frac{x-a}{b-a}, & for \;\; x \in [a,b] \\ 1, & for \;\; x > b. \end{cases}$$

and its mean and variance are

$$E(X) = \frac{a+b}{2} \quad and \quad var(X) = \frac{(b-a)^2}{12}$$

3.2 Exponential Distribution

The exponential distribution is one of the widely used continuous distributions. It is often used to model the time elapsed between events. We will now mathematically define the exponential distribution, and derive its mean and expected value.

Definition 3.2.
A continuous random variable X is said to have an exponential distribution with parameter $\lambda > 0$, shown as $X \sim Exponential(\lambda)$, if its PDF is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if} \;\; x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Property 3.2.
If $X \sim Exponential(\lambda)$, then its CDF is given by

$$F_X(x) = \begin{cases} 0, & for \;\; x < 0 \\ 1 - e^{-\lambda x}, & for \;\; x > 0. \end{cases}$$

and its mean and variance are

$$E(X) = \frac{1}{\lambda} \quad and \quad var(X) = \frac{1}{\lambda^2}.$$

Remark 3.1.
An interesting property of the exponential distribution is that it can be viewed as a continuous analogue of the geometric distribution.

Remark 3.2.
The most important property is that the exponential distribution is memoryless. To see this, think of an exponential random variable in the sense of tossing a lot of coins until observing the first heads. If we toss the coin several times and do not observe a heads, from now on it is like we start all over again. In other words, the failed coin tosses do not

impact the distribution of waiting time from now on. The reason for this is that the coin tosses are independent. We can state this formally as follows:

Property 3.3.
If X is exponential with parameter $\lambda > 0$, then X is a memoryless random variable, that is

$$P(X > x + a | X > a) = P(X > x), \quad \text{for } a, \quad x > 0.$$

3.3 Normal (Gaussian) Distribution

The normal distribution is by far the most important probability distribution. One of the main reasons for that is the Central Limit Theorem (CLT),is one of the most important theorems in probability.
To give you an idea, the CLT states that if you add a large number of random variables, the distribution of the sum will be approximately normal under certain conditions. The importance of this result comes from the fact that many random variables in real life can be expressed as the sum of a large number of random variables and, by the CLT, we can argue that distribution of the sum should be normal. The CLT is one of the most important results in probability and we will discuss it later on. Here, we will introduce normal random variables.
We first define the standard normal random variable. We will then see that we can obtain other normal random variables by scaling and shifting a standard normal random variable.

3.3.1 Standard normal random variable

Definition 3.3.
A continuous random variable Z is said to be a standard normal (standard Gaussian) random variable, shown as $Z \sim N(0, 1)$, if its PDF is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}, \quad \text{for all } z \in \mathbb{R}$$

Remark 3.3.
The $\frac{1}{\sqrt{2\pi}}$ is there to make sure that the area under the PDF is equal to one.

Property 3.4.
If $Z \sim N(0,1)$, then its CDF is given by

$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left\{-\frac{u^2}{2}\right\} du$$

and its mean and variance are

$$E(Z) = 0 \quad \textit{and} \quad var(Z) = 1.$$

Remark 3.4.
This integral does not have a closed form solution. Nevertheless, because of the importance of the normal distribution, the values of $F_Z(z)$ have been tabulated and many calculators and software packages have this function. We usually denote the standard normal CDF by Φ.

Property 3.5.
The CDF of the standard normal distribution is denoted by the Φ function:

$$\Phi(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{u^2}{2}\right\} du$$

As we will see in the rest of the section, the CDF of any normal random variable can be written in terms of the Φ function, so the Φ function is widely used in probability. Here are some properties of the Φ function that can be shown from its definition.

Property 3.6.

1. $$\lim_{x\to\infty} \Phi(x) = 1, \quad \lim_{x\to-\infty} \Phi(x) = 0;$$
2. $$\Phi(0) = \frac{1}{2}$$
3. $$\Phi(-x) = 1 - \Phi(x), \quad \textit{for all} \quad x \in \mathbb{R}.$$

3.3.2 Normal random variables

Now that we have seen the standard normal random variable, we can obtain any normal random variable by shifting and scaling a standard normal random variable. In particular, define

$$X = \sigma Z + \mu, \quad \text{where} \quad \sigma > 0.$$

Then

$$E(X) = \sigma E(Z) + \mu = \mu,$$
$$var(X) = \sigma^2 var(Z) = \sigma^2.$$

We say that X is a normal random variable with mean μ and variance σ^2. We write $X \sim N(\mu, \sigma^2)$.

Proposition 3.1.
If Z is a standard normal random variable and $X = \sigma Z + \mu$, then X is a normal random variable with mean μ and variance σ^2, i.e,

$$X \sim N(\mu, \sigma^2).$$

Conversely, if $X \sim N(\mu, \sigma^2)$, the random variable defined by $Z = \frac{X-\mu}{\sigma}$ is a standard normal random variable, i.e., $Z \sim N(0, 1)$.

Definition 3.4.
If X is a normal random variable with mean μ and variance σ^2 , i.e, $X \sim N(\mu, \sigma^2)$, then

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$
$$F_X(x) = P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$
$$P(a < X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Example 3.1.
Let $X \sim N(-5, 4)$.

1. Find $P(X < 0)$.
2. Find $P(-7 < X < -3)$.
3. Find $P(X > -3|X > -5)$.

Solution.
X is a normal random variable with $\mu = -5$ and $\sigma = \sqrt{4} = 2$, thus we have

1. Find $P(X < 0)$:

$$P(X < 0) = F_X(0) = \Phi\left(\frac{0+5}{2}\right) = \Phi(2.5) \approx 0.99$$

2. Find $P(-7 < X < -3)$:

$$P(-7 < X < -3) = F_X(-3) - F_X(-7) = \Phi\left(\frac{-3+5}{2}\right) - \Phi\left(\frac{-7+5}{2}\right) = 2\Phi(1) - 1 \approx 0.68$$

3. Find $P(X > -3 | X > -5)$:

$$P(X > -3 | X > -5) = \frac{P(X > -3, X > -5)}{P(X > -5)} = \frac{P(X > -3)}{P(X > -5)} = \frac{1 - \Phi(1)}{1 - \Phi(0)} \approx 0.32$$

An important and useful property of the normal distribution is that a linear transformation of a normal random variable is itself a normal random variable. In particular, we have the following theorem:

Theorem 3.1.
If $X \sim N(\mu_X, \sigma_X^2)$, and $Y = aX + b$, where $a, b \in \mathbb{R}$, then $Y \sim N(\mu_Y, \sigma_Y^2)$ where

$$\mu_Y = a\mu_X + b \ , \ \sigma_Y^2 = a^2\sigma_X^2.$$

3.4 Lognormal Distribution

Definition 3.5.
Let $X \sim N(0,1)$ and $X = a + b \log Y$. Then, Y is said to have a lognormal distribution with parameters a and b. By a simple transformation of random variables, we find the pdf of Y as

$$f_Y(y) = \frac{b}{\sqrt{2\pi}y} \exp\left\{-\frac{1}{2}(a + b \log y)^2\right\}, \quad y > 0$$

Remark 3.5.
We may take b to be positive without loss of any generality, since $-X$ has the same distribution as X. An alternative reparametrization is obtained by replacing the parameters a and b by the mean μ and standard deviation σ of the random variable $\log Y$. Then, the two sets of parameters satisfy the relationships

$$\mu = -\frac{a}{b} \quad \text{and} \quad \sigma = \frac{1}{b}$$

so that we have $X = (\log Y - \mu)/\sigma$, and the lognormal pdf under this reparametrization is given by

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma y} \exp\left\{-\frac{1}{2}\frac{(\log y - \mu)^2}{\sigma^2}\right\}, \quad y > 0.$$

We write $Y \sim lognormal(\mu, \sigma^2)$.

Property 3.7.
If $Y \sim lognormal(\mu, \sigma^2)$, then its mean and variance are

$$E(Y) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \quad and \quad var(Y) = \left[\exp(\sigma^2) - 1\right] \exp\left(2\mu + \sigma^2\right)$$

3.5 Gamma Distribution

The gamma distribution is another widely used distribution. Its importance is largely due to its relation to exponential and normal distributions. Here, we will provide an introduction to the gamma distribution. Before introducing the gamma random variable, we need to introduce the gamma function.

Definition 3.6. Gamma function.
The gamma function, shown by $\Gamma(x)$, is an extension of the factorial function to real (and complex) numbers. Specifically, if $n \in \{1, 2, 3, \ldots\}$, then

$$\Gamma(n) = (n-1)!$$

More generally, for any positive real number α, $\Gamma(\alpha)$ is defined as

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

Property 3.8.
For any positive real number α:

1. $$\Gamma(1) = \int_0^\infty e^{-x} dx = 1.$$

2. $$\int_0^\infty x^{\alpha-1} e^{-\lambda x}\, dx = \frac{\Gamma(\alpha)}{\lambda^\alpha}, \quad for \quad \lambda > 0;$$

3. $$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \quad for \quad \alpha > 0$$

4. $$\Gamma(n) = (n-1)!, \quad for \quad n = 1, 2, 3, \ldots$$

5. $$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

We now define the gamma distribution by providing its PDF:

Definition 3.7.
A continuous random variable X is said to have a gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$, shown as $X \sim Gamma(\alpha, \lambda)$, if its PDF is given by

$$f_X(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Remark 3.6.
• If we let $\alpha = 1$, we obtain

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, we conclude $Gamma(1, \lambda) = Exponential(\lambda)$.
• More generally, if you sum n independent $Exponential(\lambda)$ random variables, then you will get a $Gamma(n, \lambda)$ random variable.

Example 3.2.
Using the properties of the gamma function, show that the gamma PDF integrates to 1, i.e., show that for $\alpha, \lambda > 0$, we have

$$\int_0^\infty \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} dx = 1$$

Solution.
We can write

$$\int_0^\infty \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\lambda x} dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{\lambda^\alpha} = 1$$

Property 3.9.
If $X \sim Gamma(\alpha, \lambda)$, then its mean and variance are

$$E(X) = \frac{\alpha}{\lambda} \quad \text{and} \quad var(X) = \frac{\alpha}{\lambda^2}.$$

3.6 Beta Distribution

Before introducing the beta random variable, we need to introduce the beta function.

Definition 3.8. Beta function.
The integral

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx,$$

converges for $\alpha>0$ and $\beta>0$ and is called a beta function. For $\alpha\leq 0$ or $\beta\leq 0$ the integral above diverges. It is easy to see that for $\alpha>0$ and $\beta>0$.

$$B(\alpha,\beta)=B(\beta,\alpha),$$

$$B(\alpha,\beta)=\int_0^\infty x^{\alpha-1}(1+x)^{-\alpha-\beta}dx,$$

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Definition 3.9.
A random variable X with PDF given by

$$f_X(x)=\begin{cases}\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, & \text{if } 0<x<1,\\ 0, & \text{otherwise.}\end{cases}$$

is said to have a beta distribution with parameters α and β, $\alpha>0$ and $\beta>0$. We will write $X\sim B(\alpha,\beta)$ for a beta variable with density given above.

Property 3.10.
The CDF of a $B(\alpha,\beta)$ random variable is given by

$$F_X(x)=\begin{cases}0, & x\leq 0\\ \frac{1}{B(\alpha,\beta)}\int_0^x y^{\alpha-1}(1-y)^{\beta-1}dy, & \text{if } 0<x<1,\\ 1, & x\geq 1.\end{cases}$$

and its mean and variance are

$$E(X)=\frac{\alpha}{\alpha+\beta}\quad \text{and}\quad var(X)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Example 3.3.
Let X be distributed with PDF

$$f_X(x)=\begin{cases}\frac{1}{12}x^2(1-x), & \text{if } 0<x<1,\\ 0, & \text{otherwise.}\end{cases}$$

then, $X\sim B(3,2)$, $E(X)=\frac{12}{20}$ and $var(X)=\frac{1}{25}$.

3.7 Cauchy Distribution

Definition 3.10.
A random variable X is said to have a Cauchy distribution with parameters μ and θ if its PDF is given by

$$f_X(x) = \frac{\mu}{\pi}\frac{1}{\mu^2 + (x-\theta)^2}, \quad -\infty < x < \infty, \quad \mu > 0.$$

We will write $X \sim \mathcal{C}(\mu, \theta)$ for a Cauchy random variable with the density above.

Property 3.11.
The CDF of a $\mathcal{C}(1,0)$ random variable is given by

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi}tan^{-1}x, \quad -\infty < x < \infty$$

Theorem 3.2.
Let X be a Cauchy random variable with parameters μ and θ. The moments of order < 1 exist, but the moments of order ≥ 1 do not exist for the random variable X.

Proof
It suffices to consider the PDF

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \quad -\infty < x < \infty.$$

$$E|X|^\alpha = \frac{2}{\pi}\int_0^\infty x^\alpha \frac{1}{1+x^2}dx,$$

and letting $z = 1/(1+x^2)$ in the integral, we get

$$E|X|^\alpha = \frac{1}{\pi}\int_0^1 z^{\frac{(1-\alpha)}{2}-1}(1-z)^{[(\alpha+1)/2]-1}dz$$

which converges for $\alpha < 1$ and diverges for $\alpha \geq 1$. This completes the proof of the theorem.

■

Theorem 3.3.
Let $X \sim \mathcal{C}(\mu_1, \theta_1)$ and $Y \sim \mathcal{C}(\mu_2, \theta_2)$ be independent random variables. Then $X + Y$ is a $\mathcal{C}(\mu_1 + \mu_2, \theta_1 + \theta_2)$ random variable.

Theorem 3.4.
Let X be $\mathcal{C}(\mu, 0)$. Then λ/X, where λ is a constant, is a $\mathcal{C}(|\lambda|/\mu, 0)$ random variable.

3.8 Chi-square Distribution

The chi-square distribution is a continuous, positively skewed distribution defined for non-negative values, whose shape is determined by a single parameter, degrees of freedom (df). Its primary use is in statistical inference, particularly for the chi-square goodness-of-fit test and the chi-square test for independence, which compare observed categorical data to expected values to determine if there's a significant difference.

Definition 3.11.
A random variable X is said to have a Chi square distribution with parameter n which is the degree of freedom if its PDF is given by

$$f_X(x) = \begin{cases} \dfrac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases}$$

We write $X \sim \chi^2(n)$.

Property 3.12.
If $X \sim \chi^2(n)$, then its mean and variance are

$$E(X) = n \quad \text{and} \quad var(X) = 2n.$$

Theorem 3.5.
Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d) random variables with $N(0,1)$ distribution, then

$$X_1 \sim N(0,1) \Longleftrightarrow \sum_{k=1}^{n} X_k^2 \sim \chi^2(n).$$

3.9 Student Distribution

In probability theory and statistics, Student's t distribution (or simply the t distribution) is a continuous probability distribution that generalizes the standard normal distribution. Like the latter, it is symmetric around zero and bell-shaped.
The name "Student" is a pseudonym used by William Sealy Gosset in his scientific paper publications during his work at the Guinness Brewery in Dublin, Ireland.
The Student's t distribution plays a role in a number of widely used statistical analyses, including Student's t-test for assessing the statistical significance of the difference between two sample means, the construction of confidence intervals for the difference between two population means, and in linear regression analysis.

Definition 3.12.
A random variable X is said to have a Student distribution with parameter n which is the degree of freedom if its PDF is given by

$$f_n(x) = \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)} \frac{1}{\sqrt{n\pi}} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \quad \text{if } \infty < x < +\infty.$$

We write $X \sim t(n)$.

Property 3.13.
If $X \sim t(n)$, and $n > 2$ then its mean and variance are

$$E(X) = 0, \ n > 1 \quad \textit{and} \quad var(X) = n/(n-2), \ n > 2.$$

Theorem 3.6.
Let $X \sim t(n)$, $n > 1$. Then $E(X^r)$ exists for $r < n$. In particular, if $r < n$ is odd, $E(X^r) = 0$, and if $r < n$ is even,

$$E(X^r) = n^{r/2} \frac{\Gamma[(r+1)/2]\Gamma[(n-r)/2]}{\Gamma(1/2)\Gamma(n/2)}.$$

Theorem 3.7.
Let $X \sim N(0,1)$ and $Y \sim \chi^2(n)$, and let X and Y be independent. Then the statistic

$$T = \frac{X}{\sqrt{Y/n}}$$

is said to have a t-distribution with n d.f.

Remark 3.7.
For $n = 1$, T is a Cauchy random variable. We will therefore assume that $n > 1$.

Remark 3.8.
The PDF $f_n(t)$ is symmetric in t, and $f_n(t) \to 0$ as $t \to +\infty$.
For large n, the t-distribution is close to the normal distribution.

3.10 Fisher Distribution

Definition 3.13.
A random variable F is said to have a Fisher distribution or the F- distibution with (m, n) df if its PDF is given by

$$g(x) = \begin{cases} \dfrac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right) \left(\frac{m}{n}x\right)^{(m/2)-1} \left(1 + \frac{m}{n}x\right)^{-(m+n)/2}, & \text{if } x > 0. \\ 0, & \text{if } x < 0. \end{cases}$$

We write $F \sim F(m, n)$.

Theorem 3.8.
Let X and Y be independent χ^2 random variables with m and n d.f., respectively. The random variable

$$F = \frac{X/m}{Y/n}$$

is said to have an F-distribution with (m, n) d.f.

Property 3.14.
If $X \sim F(m, n)$, and $n > 2$ then its mean and variance are

$$E(X) = \frac{n}{n-2} \quad \text{if } n > 2,$$

and

$$var(X) = \frac{n^2(2m + 2n - 4)}{m(n-2)^2(n-4)} \quad \text{if } n > 4.$$

Theorem 3.9.
Let $X \sim F(m, n)$. Then, for $k > 0$, integral

$$E(X^k) = \left(\frac{n}{m}\right)^k \frac{\Gamma[k + (m/2)]\Gamma[(n/2) - k]}{\Gamma[(m/2)\Gamma(n/2)]} \quad \text{for } n > 2k.$$

Remark 3.9.
If $X \sim F(m,n)$, then $1/X \sim F(n,m)$. If we take $m = 1$, then $F = [t(n)]^2$, so that $F(1,n)$ and $t^2(n)$ have the same distribution. It also follows that, if Z is $\mathcal{C}(1,0)$ [which is the same as $t(1)$], Z^2 is $F(1,1)$.

3.11 Weibull distribution

In probability theory and statistics, the Weibull distribution is a continuous probability distribution. It models a broad range of random variables, largely in the nature of a time to failure or time between events. Examples are maximum one-day rainfalls and the time a user spends on a web page.

Definition 3.14.
A random variable X is said to have a Weibull distribution if its PDF is given by

$$f(x,\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right) e^{-(x/\lambda)^k}, & \text{if } x \geq 0. \\ 0, & \text{if } f < 0. \end{cases}$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the distribution We write $F \sim Weibull(k,\lambda)$.

Property 3.15.
The CDF of a $Weibull(k,\lambda)$ random variable is given by

$$F_X(x) = \begin{cases} 1 - e^{-(x/\lambda)^k}, & \text{if } x \geq 0, \\ 0, & x < 0. \end{cases}$$

and its mean and variance are

$$E(X) = \lambda\,\Gamma(1+\frac{1}{k}) \quad \text{and} \quad var(X) = \lambda^2\left[\Gamma\left(1+\frac{2}{k}\right) - \left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right].$$

Remark 3.10.
• The Weibull distribution can be characterized as the distribution of a random variable W such that the random variable

$$X = \left(\frac{W}{\lambda}\right)^k$$

is the standard exponential distribution with parameter 1.
• If $X \sim Weibull(\frac{1}{2},\lambda)$ then $X \sim Exponential(\frac{1}{\sqrt{\lambda}})$.

4 Approximation of some usual distributions

4.1 Poisson as an approximation for binomial distribution

The Poisson distribution can be viewed as the limit of binomial distribution. Suppose $X \sim Binomial(n, p)$ where n is very large and p is very small. In particular, assume that $\lambda = np$ is a positive constant. We show that the PMF of X can be approximated by the PMF of a $Poisson(\lambda)$ random variable. The importance of this is that Poisson PMF is much easier to compute than the binomial. Let us state this as a theorem.

Theorem 4.1.
Let $X \sim Binomial(n, p = p(n))$,such that $n \geq 30$ and $p \leq 0.1$, so $\lambda > 0$ is fixed and $\lim_{n\to\infty} np = \lambda$. Then, the PMF of X converges to a $Poisson(\lambda)$ PMF, as $n \to \infty$. That is, we have for any $k \in \{0, 1, 2, \ldots\}$,

$$\lim_{n\to\infty} P_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

4.2 Binomial as an approximation for Hypergeometric distribution

Intuitively, following the definition of the hypergeometric distribution, we deduce the convergence to the binomial distribution. Let us state this as a theorem.

Theorem 4.2.
Suppose X has a hypergeometric distribution with parameters a, b, and n and probability function f .

$$f(x) = \begin{cases} \dfrac{C_a^x C_b^{n-x}}{C_{a+b}^n}, & for \;\; x \in R_X. \\ 0, & otherwise. \end{cases}$$

Then, as $a \to \infty$ and $b \to \infty$ in such a way that

$$\lim_{a,b\to\infty} \frac{a}{a+b} = p,$$

we have that for $x = 0, 1, 2, \ldots, n$

$$\lim_{a,b\to\infty} f(x) = lim_{a,b\to\infty} P(X = x) = C_n^x p^x (1-p)^{n-x}.$$

4.3 Normal as an approximation for Poisson distribution

The Normal distribution can be viewed as the limit of Poisson distribution $Poisson(\lambda)$. If the parameter λ is very large. This approximation is useful because the normal distribution is easier to manipulate mathematically, especially for probability calculations. We consider the following theorem.

Theorem 4.3.
For sufficiently large values of λ, (say $\lambda > 15$) the normal distribution with mean λ and variance λ (standard deviation $\sqrt{\lambda}$), is an excellent approximation to the Poisson distribution. So

$$Poisson(\lambda) \rightsquigarrow Normal(\lambda, \lambda),$$

when λ is large.

4.4 The normal approximation to the binomial distribution

To assert this approximation, we consider the following result.

Theorem 4.4.
Consider the binomial distribution,

$$f(x) = C_n^x p^x (1-p)^{n-x},$$

so $X \sim Binomial(n, p)$. Remarkably, when n, np and $n(1-p)$ are large i.e as long as p is not too close to either 0 or 1 and n is taken large,
then the binomial distribution is well approximated by the normal distribution. So,

$$f(x) = C_n^x p^x (1-p)^{n-x} \approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-(x-np)^2/2np(1-p)}$$

i.e

$$Binomial(n, p) \rightsquigarrow Normal(np, np(1-p)).$$

5 Transformations of random variables

The topic for this section is transformations of random variables. After applying a function to a random variable X, the goal is to find the distribution of the transformed random variable.
Transformations of random variables appear all over the place in statistics. Here are a few examples, to preview the kinds of transformations we'll be looking at in this section.
• **Unit conversion**: In one dimension, we've already seen how standardization and location-scale transformations can be useful tools for learning about an entire family of distributions. A location-scale change is linear, converting a r.v X to the r.v. $Y = aX + b$ where a and b are constants (with $a > 0$).

• **Sums and averages as summaries**: It is common in statistics to summarize n observations by their sum or sample average. Turning $X_1, \ldots, X_n$ into the sum $T = X_1 + \ldots + X_n$. The term for a sum of independent random variables is convolution. In this section, convolution sums and integrals, which are based on the law of total probability, will give us another way of obtaining the distribution of a sum of random variables.

• **Extreme values**: In many contexts, we may be interested in the distribution of the most extreme observations. For disaster preparedness, government agencies may be concerned about the most extreme flood or earthquake in a 100-year period; in finance, a portfolio manager with an eye toward risk management will want to know the worst 1% or 5% of portfolio returns. In these applications, we are concerned with the maximum or minimum of a set of observations. The transformation that sorts observations, turning $X_1, \ldots, X_n$ into the order statistics $\min(X_1, \ldots, X_n)$,..., $\max(X_1, \ldots, X_n)$, is a transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$ that is not invertible. Order statistics are addressed in the last section in this section.

If we need the full distribution of $g(X)$, not just its expectation, our approach depends on whether X is discrete or continuous.
• In the discrete case, we get the PMF of $g(X)$ by translating the event $g(X) = y$ into an equivalent event involving X. To do so, we look for all values x such that $g(x) = y$, as long as X equals any of these x's, the event $g(X) = y$ will occur. This gives the formula

$$P(g(X) = y) = \sum_{x:g(x)=y} P(X = x),$$

For a one-to-one g, the situation is particularly simple, because there is only one value of x such that $g(x) = y$, namely $g^{-1}(y)$. Then we can use

$$P(g(X) = y) = P(X = g^{-1}(y))$$

to convert between the PMFs of X and g(X).

• In the continuous case, a universal approach is to start from the CDF of $g(X)$, and translate the event $g(X) \leq y$ into an equivalent event involving X. For general g, we may have to think carefully about how to express $g(X) \leq y$ in terms of X, and there is no easy formula we can plug into. But when g is continuous and strictly increasing, the translation is easy: $g(X) \leq y$ is the same as $X \leq g^{-1}(y)$, so

$$F_{g(X)}(y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)),$$

We can then differentiate with respect to y to get the PDF of $g(X)$.

5.1 Change of variables

Theorem 5.1.
Let X be a continuous random variable with PDF f_X, and let $Y = g(X)$, where g is differentiable and strictly increasing (or strictly decreasing). Then the PDF of Y is given by

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|,$$

where $x = g^{-1}(y)$. The support of Y is all $g(x)$ with x in the support of X.

Proof
Let g be strictly increasing. The CDF of Y is

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) = F_X(x)$$

so by the chain rule, the PDF of Y is

$$f_Y(y) = f_X(x)\frac{dx}{dy}.$$

The proof for g strictly decreasing is analogous. In that case the PDF ends up as $-f_X(x)\frac{dx}{dy}$, which is nonnegative since $\frac{dx}{dy} < 0$ if g is strictly decreasing. Using $\left|\frac{dx}{dy}\right|$, as in the statement of the theorem, covers both cases.

■

When applying the change of variables formula, we can choose whether to compute $\frac{dx}{dy}$, or compute $\frac{dy}{dx}$ and take the reciprocal. By the chain rule, these give the same result, so we can do whichever is easier.

Remark 5.1.
When finding the distribution of Y, be sure to:

- Check the assumptions of the change of variables theorem carefully if you wish to apply it (if it doesn't apply, a good strategy is to start with the CDF of Y).
- Express your final answer for the PDF of Y as a function of y.
- Specify the support of Y .

The change of variables formula (in the strictly increasing g case) is easy to remember when written in the form

$$f_Y(y)dy = f_X(x)dx,$$

which has an aesthetically pleasing symmetry to it.

The next two examples derive the PDFs of two r.v.s that are defined as transformations of a standard Normal r.v. In the first example the change of variables formula applies; in the second example it does not.

Example 5.1. **(Log-Normal PDF)**.
Let $X \sim N(0,1)$, $Y = e^X$, we use the change of variables formula to find the PDF of Y , since $g(x) = e^x$ is strictly increasing. Let $y = e^x$, so $x = \log y$ and $dy/dx = e^x$. Then

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = \varphi(x)\frac{1}{e^x} = \varphi(\log y)\frac{1}{y},\ y > 0.$$

Note that after applying the change of variables formula, we write everything on the right-hand side in terms of y, and we specify the support of the distribution. To determine the support, we just observe that as x ranges from $-\infty$ to ∞, e^x ranges from 0 to ∞.
We can get the same result by working from the definition of the CDF, translating the event $Y \leq y$ into an equivalent event involving X. For $y > 0$,

$$F_Y(y) = P(Y \leq y) = P(e^X \leq y) = P(X \leq \log y) = \Phi(\log y),$$

so the PDF is again

$$f_Y(y) = \frac{d}{dy}\Phi(\log y) = \varphi(\log y)\frac{1}{y},\ y > 0.$$

Example 5.2. **(Chi-Square PDF).**
Let $X \sim N(0,1),\ Y = X^2$. The distribution of Y is an example of a Chi-Square distribution. To find the PDF of Y , we can no longer apply the change of variables formula because $g(x) = x^2$ is not one-to-one; instead we start from the CDF. By drawing the graph of $y = x^2$, we can see that the event $X^2 \leq y$ is equivalent to the event $-\sqrt{y} \leq X \leq \sqrt{y}$. Then

$$F_Y(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1,$$

so

$$f_Y(y) = 2\varphi(\sqrt{y})\frac{1}{2}y^{-1/2} = \varphi(\sqrt{y})y^{-1/2},\ y > 0,$$

5.2 Convolutions

A convolution is a sum of independent random variables. As we mentioned earlier, we often add independent r.v.s because the sum is a useful summary of an experiment (in n Bernoulli trials, we may only care about the total number of successes), and because sums lead to averages, which are also useful (in n Bernoulli trials, the proportion of successes).
The main task in this section is to determine the distribution of $T = X + Y$, where X and Y are independent r.v.s whose distributions are known.
A method for obtaining the distribution of T is by using a convolution sum or integral. The formulas are given in the following theorem. As we'll see, a convolution sum is nothing more than the law of total probability, conditioning on the value of either X or Y ; a convolution integral is analogous.

Theorem 5.2. ***(Convolution sums and integrals)***.
Let X and Y be independent r.v.s and $T = X + Y$ be their sum. If X and Y are discrete, then the PMF of T is

$$P(T = t) = \sum_x P(Y = t - x)P(X = x) = \sum_y P(X = t - y)P(Y = y).$$

If X and Y are continuous, then the PDF of T is

$$f_T(t) = \int_{-\infty}^{\infty} f_Y(t - x)f_X(x)dx = \int_{-\infty}^{\infty} f_X(t - y)f_Y(y)dy.$$

Proof
For the discrete case, we use the following method.

$$P(T = t) = \sum_x P(X + Y = t|X = x)P(X = x) = \sum_x P(Y = t - x|X = x)P(X = x)$$

$$P(T = t) = \sum_x P(Y = t - x)P(X = x).$$

Conditioning on Y instead, we obtain the second formula for the PMF of T.
We use the assumption that X and Y are independent in order to get from $P(Y = t - x|X = x)$ to $P(Y = t - x)$ in the last step. We are only justified in dropping the condition $X = x$ if the conditional distribution of Y given $X = x$ is the same as the marginal distribution of Y , i.e., X and Y are independent. A common mistake is to assume that after plugging in x for X, we've "already used the information" that $X = x$, when in fact we need an independence assumption to drop the condition. Otherwise we destroy information without justification.

In the continuous case, since the value of a PDF at a point is not a probability, we first find the CDF, and then differentiate to get the PDF.

$$F_T(t) = P(X+Y \leq t) = \int_{-\infty}^{\infty} P(X+Y \leq t|X = x) f_X(x)dx = \int_{-\infty}^{\infty} P(Y \leq t-x) f_X(x)dx$$

$$F_T(t) = \int_{-\infty}^{\infty} F_Y(t-x) f_X(x)dx.$$

Again, we need independence to drop the condition $X = x$. To get the PDF, we then differentiate with respect to t, interchanging the order of integration and differentiation. This gives

$$f_T(t) = \int_{-\infty}^{\infty} f_Y(t-x) f_X(x)dx.$$

■

In the following examples, we find the distribution of a sum of Exponentials and a sum of Uniforms using a convolution integral.

Example 5.3. **(Exponential convolution).**
Let $X, Y \sim Exp(\lambda)$. Find the distribution of $T = X + Y$.
Solution:
For $t > 0$, the convolution formula gives

$$f_T(t) = \int_{-\infty}^{\infty} f_Y(t-x) f_X(x)dx = \int_0^t \lambda e^{-\lambda(t-x)} \lambda e^{-\lambda x} dx,$$

where we restricted the integral to be from 0 to t since we need $t - x > 0$ and $x > 0$ for the PDFs inside the integral to be nonzero. Simplifying, we have

$$f_T(t) = \lambda^2 \int_0^t e^{-\lambda t} dx = \lambda^2 t e^{-\lambda t}, \quad \text{for} \quad t > 0.$$

This is known as the $Gamma(2, \lambda)$ distribution.

Example 5.4. **(Uniform convolution)**.
Let $X, Y \sim Uniform(0, 1)$. Find the distribution of $T = X + Y$.

Solution: The PDF of X (and of Y) is

$$g(x) = \begin{cases} 1, & x \in (0,1) \\ 0, & \text{otherwise.} \end{cases}$$

The convolution formula gives

$$f_T(t) = \int_{-\infty} \int^{\infty} f_Y(t-x) f_X(x)dx = \int_{-\infty} \int^{\infty} g(t-x) g(x)dx.$$

The integrand is 1 if and only if $0 < t-x < 1$ and $0 < x < 1$; this is a parallelogram shaped constraint. Equivalently, the constraint is $\max(0, t-1) < x < \min(t, 1)$. Therefore, the PDF of T is a piecewise linear function:

$$f_T(t) = \begin{cases} \int_0^t dx = t, & \text{for } 0 < t \leq 1 \\ \int_{t-1}^1 dx = 2 - t, & \text{for } 1 < t < 2 \end{cases}$$

5.3 Order statistics

The final transformation we will consider in this section is the transformation that takes n random variables $X_1, \ldots, X_n$ and sorts them in order, producing the transformed r.v.s $\min(X_1, \ldots, X_n)$, ..., $\max(X_1, \ldots, X_n)$. The transformed r.v.s are called the order statistics, and they are often useful when we are concerned with the distribution of extreme values, as we alluded to earlier.

Furthermore, like the sample mean $\bar{X}_n$, the order statistics serve as useful summaries of an experiment, since we can use them to determine the cutoffs for the worst 5% of observations, the worst 25%, the best 25%, and so forth (such cutoffs are called the quantiles of the sample).

Definition 5.1. (Order statistics).
For r.v.s $X_1, X_2 \ldots, X_n$, the order statistics are the random variables $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, where

$$X_{(1)} = \min(X_1, \ldots, X_n)$$

$$X_{(2)} \text{ is the second-smallest of } X_1, \ldots X_n.$$

$$\vdots$$

$$X_{(n-1)} \text{ is the second-largest of } X_1, \ldots X_n.$$

$$X_{(n)} = \max(X_1, \ldots, X_n)$$

Note that $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ by definition. We call $X_{(j)}$ the jth order statistic. If n is odd, $X_{((n+1)/2)}$ is called the sample median of $X_1, \ldots, X_n$.

Remark 5.2.
The order statistics $X_{(1)}, \ldots, X_{(n)}$ are r.v.s, and each $X_{(j)}$ is a function of $X_1, \ldots, X_n$. Even if the original r.v.s are independent, the order statistics are dependent: if we know that $X_{(1)} = 100$, then $X_{(n)}$ is forced to be at least 100.

Let's start with the CDF of $X_{(n)} = \max(X_1, \ldots, X_n)$. Since $X_{(n)}$ is less than x if and only if all of the X_j are less than x, the CDF of $X_{(n)}$ is

$$F_{X_{(n)}}(x) = P(\max(X_1, \ldots, X_n) \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) = P(X_1 \leq x) \ldots P(X_n \leq x)$$

$$F_{X_{(n)}}(x) = (F(x))^n.$$

where F is the CDF of the individual X_i. Similarly, $X_{(1)} = \min(X_1, \ldots, X_n)$ exceeds x if and only if all of the X_j exceed x, so the CDF of $X_{(1)}$ is

$$F_{X_{(1)}}(x) = 1 - P(\min(X_1, \ldots, X_n) > x) = 1 - P(X_1 > x, \ldots, X_n > x) = 1 - (1 - F(x))^n.$$

The same logic lets us find the CDF of $X_{(j)}$. For the event $X_{(j)} \leq x$ to occur, we need at least j of the X_i to fall to the left of x.
Since it appears that the number of X_i to the left of x will be important to us, let's define a new random variable, N, to keep track of just that: define N to be the number of X_i that land to the left of x. Each X_i lands to the left of x with probability $F(x)$, independently. If we define success as landing to the left of x, we have n independent Bernoulli trials with probability $F(x)$ of success, so $N \sim B(n, F(x))$. Then, by the Binomial PMF,

$$P(X_{(j)} \leq x) = P(\text{at least } j \text{ of the } X_i \text{ are to the left of } x) = P(N \geq j)$$

$$P(X_{(j)} \leq x) = \sum_{k=j}^{n} C_n^k F(x)^k (1 - F(x))^{n-k}.$$

We thus have the following result for the CDF of $X_{(j)}$.

Theorem 5.3. ***(CDF of order statistic)***.
Let $X_1, \ldots, X_n$ be i.i.d. continuous r.v.s with CDF F. Then the CDF of the jth order statistic $X_{(j)}$ is

$$P(X_{(j)} \leq x) = \sum_{k=j}^{n} C_n^k F(x)^k (1 - F(x))^{n-k}.$$

To get the PDF of $X_{(j)}$, we can differentiate the CDF with respect to x, but the resulting expression is ugly (though it can be simplified).

Remark 5.3.
What is the probability of this extremely specific event? Let's break up the experiment into stages.
• First, we choose which one of the X_i will fall into the infinitesimal interval around x. There are n such choices, each of which occurs with probability $f(x)dx$, where f is the PDF of the X_i.
• Next, we choose exactly $j - 1$ out of the remaining $n - 1$ to fall to the left of x. There

are C_{n-1}^{j-1} such choices, each with probability $F(x)^{j-1}(1-F(x))^{n-j}$ by the $B(n,F(x))$ PMF. We multiply the probabilities of the two stages to get

$$f_{X_{(j)}}(x)dx = nf(x)\ dx\ C_{n-1}^{j-1}F(x)^{j-1}(1-F(x))^{n-j}.$$

Dropping the dx's from both sides gives us the PDF we desire.

Theorem 5.4. ***(PDF of order statistic)***.
Let $X_1,\dots,X_n$ be i.i.d. continuous r.v.s with CDF F and PDF f. Then the marginal PDF of the jth order statistic $X_{(j)}$ is

$$f_{X_{(j)}}(x) = nf(x)C_{n-1}^{j-1}F(x)^{j-1}(1-F(x))^{n-j}.$$

In general, the order statistics of $X_1,\dots,X_n$ will not follow a named distribution, but the order statistics of the standard Uniform distribution are an exception.

Example 5.5. **(Order statistics of Uniforms)**.
Let $U_1,\dots,U_n$ be i.i.d. $Uniform(0,1)$. Then for $0\le x\le 1$, $f(x)=1$ and $F(x)=x$, so the PDF of $U_{(j)}$ is

$$f_{U_{(j)}}(x) = nC_{n-1}^{j-1}x^{j-1}(1-x)^{n-j}.$$

This is the $Beta(j,n-j+1)$ PDF. So $U_{(j)}\sim Beta(j,n-j+1)$, and $E(U_{(j)})=\frac{j}{n+1}$.
We show that for i.i.d. $U_1,U_2\sim Uniform(0,1)$,

$$E(\max(U_1,U_2)) = 2/3,\ E(\min(U_1,U_2)) = 1/3.$$

Now that we know $\max(U_1,U_2)$ and $\min(U_1,U_2)$ follow Beta distributions, the expectation of the Beta distribution confirms our earlier findings.

6 Some exercises

6.1

Suppose the number of customers arriving at a store obeys a Poisson distribution with an average of λ customers per unit time. That is, if Y is the number of customers arriving in an interval of length t, then $Y \sim Poisson(\lambda t)$. Suppose that the store opens at time $t = 0$. Let X be the arrival time of the first customer. Show that $X \sim Exponential(\lambda)$.

Solution.
We first find $P(X > t)$:

$$P(X > t) = P(\text{No arrival in } [0, t]) = e^{-\lambda t}\frac{(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Thus, the CDF of X for $x > 0$ is given by

$$F_X(x) = 1 - P(X > x) = 1 - e^{-\lambda x},$$

which is the CDF of $Exponential(\lambda)$. Note that by the same argument, the time between the first and second customer also has $Exponential(\lambda)$ distribution. In general, the time between the k'th and $k + 1$'th customer is $Exponential(\lambda)$.

6.2

(Exponential as the limit of Geometric)
Let $Y \sim Geometric(p)$, where $p = \lambda\Delta$. Define $X = Y\Delta$, where $\lambda, \Delta > 0$. Prove that for any $x \in (0, \infty)$, we have

$$\lim_{\Delta \to 0} F_X(x) = 1 - e^{-\lambda x}.$$

Solution
If $Y \sim Geometric(p)$ and $q = 1 - p$, then

$$P(Y \leq n) = \sum_{k=1}^{n} pq^{k-1} = p.\frac{1 - q^n}{1 - q} = 1 - (1 - p)^n.$$

Then for any $y \in (0, \infty)$, we can write

$$P(Y \leq y) = 1 - (1 - p)^{\lfloor y \rfloor},$$

where $\lfloor y \rfloor$ is the largest integer less than or equal to y. Now, since $X = Y\Delta$, we have

$$F_X(x) = P(X \leq x) = P\left(Y \leq \frac{x}{\Delta}\right) = 1 - (1 - p)^{\lfloor \frac{x}{\Delta} \rfloor} = 1 - (1 - \lambda\Delta)^{\lfloor \frac{x}{\Delta} \rfloor}$$

Now, we have

$$\lim_{\Delta \to 0} F_X(x) = \lim_{\Delta \to 0} 1 - (1 - \lambda\Delta)^{\lfloor \frac{x}{\Delta} \rfloor} = 1 - e^{-\lambda x}.$$

The last equality holds because $\frac{x}{\Delta} - 1 \leq \lfloor \frac{x}{\Delta} \rfloor \leq \frac{x}{\Delta}$, and we know

$$\lim_{\Delta \to 0} (1 - \lambda\Delta)^{\frac{1}{\Delta}} = e^{-\lambda}.$$

6.3

Let $U \sim Uniform(0, 1)$ and $X = -\ln(1 - U)$. Show that $X \sim Exponential(1)$.

Solution.
First note that since $R_U = (0, 1)$, $R_X = (0, \infty)$. We will find the CDF of X. For $x \in (0, \infty)$, we have

$$F_X(x) = P(X \leq x) = P(-\ln(1-U) \leq x) = P\left(\frac{1}{1-U} \leq e^x\right) = P(U \leq 1-e^{-x}) = 1-e^{-x},$$

which is the CDF of an $Exponential(1)$ random variable.

6.4

Let $X \sim N(2, 4)$ and $Y = 3 - 2X$.

1. Find $P(X > 1)$.
2. Find $P(-2 < Y < 1)$.
3. Find $P(X > 2|Y < 1)$.

Solution.

1. Find $P(X > 1)$:
   We have $\mu_X = 2$ and $\sigma_X = 2$. Thus,
   $$P(X > 1) = 1 - \Phi\left(\frac{1-2}{2}\right) = 1 - \Phi(-0.5) = \Phi(0.5) = 0.6915$$
2. Find $P(-2 < Y < 1)$:
   Since $Y = 3 - 2X$, we have $Y \sim N(-1, 16)$. Therefore,
   $$P(-2 < Y < 1) = \Phi\left(\frac{1-(-1)}{4}\right) - \Phi\left(\frac{-2-(-1)}{4}\right) = \Phi(0.5) - \Phi(-0.25) = 0.29$$
3. Find $P(X > 2|Y < 1)$:
   $$P(X > 2|Y < 1) = P(X > 2|3 - 2X < 1) = P(X > 2|X > 1)$$
   $$= \frac{P(X > 2, X > 1)}{P(X > 1)} = \frac{P(X > 2)}{P(X > 1)} = \frac{1 - \Phi(\frac{2-2}{2})}{1 - \Phi(\frac{1-2}{2})} \approx 0.72$$

6.5

Let $X \sim N(0,2)$. Find $E|X|$.

Solution.
We can write $X = \sigma Z$, where $Z \sim N(0,1)$. Thus, $E|X| = \sigma E|Z|$. We have

$$E|Z| = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |t| e^{-\frac{t^2}{2}} dt = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} |t| e^{-\frac{t^2}{2}} dt \text{ (integral of an even function)}$$

$$= \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} t e^{-\frac{t^2}{2}} dt = \sqrt{\frac{2}{\pi}}$$

Thus, we conclude $E|X| = \sigma E|Z| = \sigma\sqrt{\frac{2}{\pi}}$.

6.6

Let $X \sim Gamma(\alpha, \lambda)$, where $\alpha, \lambda > 0$. Find $E(X)$, and $var(X)$.

Solution.
To find $E(X)$ we can write

$$E(X) = \int_0^{\infty} x f_X(x) dx = \int_0^{\infty} x \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx$$

$$= \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} x^{\alpha} e^{-\lambda x} dx = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}} \text{ (using Property of the gamma function)}$$

$$= \frac{\alpha\Gamma(\alpha)}{\lambda\Gamma(\alpha)} = \frac{\alpha}{\lambda}.$$

Similarly, we can find $E(X^2)$:

$$E(X) = \int_0^{\infty} x^2 f_X(x) dx = \int_0^{\infty} x^2 \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx$$

$$= \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} x^{\alpha+1} e^{-\lambda x} dx = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \frac{\Gamma(\alpha+2)}{\lambda^{\alpha+2}} \text{ (using Property of the gamma function)}$$

$$= \frac{\alpha+1\Gamma(\alpha+1)}{\lambda^2\Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

So, we conclude

$$var(X) = E(X^2) - (E(X))^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2}$$

$$var(X) = \frac{\alpha}{\lambda^2}$$

References

[1] Billingsley, P. (1986). Probability and Measure, 2nd ed., Wiley, New York.

[2] Chung, K. L. (1968). A Course in Probability Theory, Harcourt, Brace & World, New York.

[3] Durrett, R. (2009). Elementary Probability for Applications. Cambridge, NY: Cambridge University Press.

[4] Feller, W. (1968). An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd ed., Wiley, New York.

[5] Feller, W. (1971). An Introduction to Probability Theory and Its Applications, Vol. 2, 2nd ed., Wiley, New York.

[6] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics, Phil. Trans. R. Soc. A, 222, 309-386.

[7] Ghahramani, S. (2005). Fundamentals of Probability with Stochastic Processes, 3e. London: Pearson Education Ltd.

[8] Grimmett, G. and Stirzaker, D.R. (2001a). One Thousand Exercises in Probability. Oxford: Oxford University Press.

[9] Grimmett, G. and Stirzaker, D.R. (2001b). Probability and Random Processes, 3e. Oxford: Oxford University Press.

[10] Grimmett, G. and Welsh, D. (1986). Probability: An introduction. Oxford: Clarendon Press.

[11] Pitman, J. (1993). Probability. New York: Springer.

[12] Ross, S.M. (2006). A First Course in Probability, 7e. London, New Jersey: Pearson Prentice Hall (International Edition).

[13] Ross, S.M. (2007). An Introduction to Probability Models, 9e. Academic Press.

[14] Weiss, N.A. (2006). A Course in Probability. Pearson Education.