Introduction to probability

Xiaolong Han

Department of Mathematics, California State University, Northridge, CA 91330, USA

Email address: xiaolong.han@csun.edu

Contents

Preface

Every branch of mathematics, whether it is geometry, algebra, or any other, comprises both an abstract theoretical aspect and a practical applied aspect. There is a dynamic and mutually influential relationship between these two facets. Advancements in theory often open up new avenues for practical applications, and each fresh application gives rise to novel theoretical challenges, steering the course of ongoing research. Paradoxically, the applications of mathematics to a wide array of different fields stems from its inherent abstraction. – Its theories are not confined to any single specific application. Take, for instance, the concept of parabolas in algebra. They can describe the arc of a baseball in flight as well as the trajectory of a spacecraft bound for Mars, demonstrating the versatility and universality of mathematical principles.

The dynamic interaction between abstract theory and practical application is particularly pronounced in the branch of probability, and this interaction may be even more significant here than in geometry and algebra. The reason for this is twofold. First, probability applications have a rich historical foundation and are pervasive in our daily lives, such as coin tossing and medical trials. As a result, one often has formed an intuitive understanding of probability (which, I assure the readers, is accurate) before encountering the formal theory of probability. Second, the formal theory of probability is relatively young (comparing to, say, geometry and algebra) and only became possible after Lebesgue's theory of measure and integration[i]. In light of these factors, my objective in this book of *Introduction to probability* is to strike a delicate balance between these two aspects: Enable readers to comprehend and apply the theory of probability while also preserving and reinterpreting their pre-existing intuition about probability.

On the side of intuition and applications of probability, I include a diverse collection of examples. These examples range from hands-on models like determining the probability of getting two heads in three independent coin tosses to more general models like determining the probability of achieving r successes in n independent medical trials. Collectively, these examples serve as a bridge, facilitating readers in transitioning from an intuitive understanding of probability to its formal theory. Another example is the computation of the expectation (i.e., mean) of test scores in a class, which is considered as a random variable. Through two curving methods of the scores, readers can appreciate that, within the abstract probability theory, taking expectation commutes with linear operations (e.g., curving each score by $X \to 0.9X + 10$) but does not commute with nonlinear operations (e.g., $X \to \sqrt{100X}$).

On the side of abstract theory of probability, the foundation of this book follows the modern approach established by Kolmogorov[ii]. However, it is important to note that the level of abstraction in the theory of probability within this book remains absolutely *minimal*. This approach is intentional, as it aligns with our goal of applying probability theory to practical problems while elucidating the underlying intuition.

[i] Henri Lebesgue, *Leçons sur l'integration et la recherche des fonctions primitives.* [Lessons on integration and analysis of primitive functions]. (1904).

[ii] Andrey Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung.* [Foundations of the theory of probability]. (1933).

At this minimal level of abstraction, the mathematics required (and is covered) in this book includes the following.

- The language of sets and functions: It is indispensable for the mathematical theory of probability. Without this language, the seemingly elementary ideas like "mutually exclusiveness" in probability are impossible to define.
- Combinatorics (i.e., the theory of counting): It supplies the critical tools to probability in the *discrete* setting.
- Integration in calculus: It supplies the critical tools to probability in the *continuous* setting.

Message to Students:

- Do not get discouraged by the abstraction of abstract theories. A concept is abstract because there is a distance to our common intuition and practical experiences. Understanding them is typically a graduate process that evolves as you learn a mathematical branch. You can enhance your comprehension by studying more examples, checking on your intuition, and gaining practical experience applying the theory. Keep simple examples readily available as we introduce the abstract concepts.
- You can skip sections marked with stars ($\star$) during your initial reading.
- You are entitled to a reward of one point toward a Test if you can report a mistake, a typo, or any other inaccuracies.

CHAPTER 1

Preliminaries: Sets and Combinatorics

There are two types of mind, the mathematical, and what might be called the intuitive. The former arrives at its views slowly, but they are firm and rigid; the latter is endowed with greater flexibility and applies itself simultaneously to the diverse lovable parts of that which it loves.

Blaise Pascal, 1653[i]

In this chapter, we review the preliminaries of probability:

- Sets and functions: This is the language of mathematics, including probability.
- Combinatorics, i.e., the mathematical theory of counting: This supplies the main tools for probability in the *discrete* setting.

1.1. Sets and functions

Definition (Sets).

- Given a set A, $x \in A$ denotes that x is a member (or element, point) of A and $x \notin A$ denotes that x is not an member of A.
- We say that two sets A and B are equal, denoted by $A = B$, if they have the same members.
- Given two sets A and B, we say that A is a subset of B, denoted by $A \subset B$, if each member of A is a member of B, that is, $x \in A$ implies that $x \in B$.

Remark. Given two sets A and B, $A = B$ if and only if (denoted by "iff") $A \subset B$ and $B \subset A$. Therefore, to show that $A = B$, we need to prove that, on one hand, $x \in A$ implies that $x \in B$, so $A \subset B$, and on the other hand, $x \in B$ implies that $x \in A$, so $B \subset A$.

Example.

- We say that a set A is finite, if there are finite number of members of A. In this case, we use $|A|$ to denote the size (i.e., number of members) of A.
- $\mathbb{N} = \{1, 2, 3, ...\}$ denotes the set of natural numbers.[ii]
- We say that a set is countably infinite if its members can be arranged into an infinite sequence. For example, $\mathbb{N}$ is countably infinite.
- We say that a set is discrete if it is finite or countably infinite.

Definition (The empty set). The set that has no members is called the empty set and is denoted by $\emptyset$. A set that is not equal to the empty set is said to be nonempty.

Definition (A singleton set). A set that has a single member is called a singleton set.

Definition (The power set). Given a set A, the set of all the subsets of A is called the power set of A and is denoted by $\mathcal{P}(A)$.

Example. Let $A = \{1, 2, 3\}$. Then

$$\mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

[i] Blaise Pascal, *Discours sur les passions de l'amour*. (1653).

[ii] Notice that in the literature, $\mathbb{N}$ may refer to the set $\{0, 1, 2, 3, ...\}$.

Remark. If A has n members, then $\mathcal{P}(A)$ has 2^n members. This is because that each subset of A can be constructed via a selection process through all of its members (i.e., selected or not). In the example above, we find all the subsets of $\{1,2,3\}$ via a selection process through members 1 (selected or not), 2 (selected or not), and 3 (selected or not).

Definition (Union, intersection, and complement). Let A and B be two sets.

- The union of A and B is $A \cup B = \{x : x \in A \text{ or } x \in B\}$.
- The intersection of A and B is $A \cap B = \{x : x \in A \text{ and } x \in B\}$. We say that A and B are disjoint if $A \cap B = \emptyset$.
- The complement of A in B is $B \setminus A = \{x : x \in B \text{ and } x \notin A\}$, and is also denoted by $B \setminus A$. In particular, if all the set operations are within a universal set X, then for a set $A \subset X$, we simply call $X \setminus A$ the complement of A, and is denoted by A^c.

PROPOSITION 1.1. *Let A, B be sets. Then $A \cup B = A \cup (B \cap A^c)$.*

PROOF. First we show that $A \cup B \subset A \cup (B \cap A^c)$. Pick $x \in A \cup B$. Then $x \in A$ or $x \in B$. If $x \in A$, then $x \in A \cup (B \cap A^c)$; if $x \notin A$, then $x \in B$ (otherwise $x \notin A \cup B$), hence, $x \in B \cap A^c$ so $x \in A \cup (B \cap A^c)$. Therefore, $A \cup B \subset A \cup (B \cap A^c)$.

Next we show that $A \cup (B \cap A^c) \subset A \cup B$. Pick $x \in A \cup (B \cap A^c)$. Then $x \in A$ or $x \in B \cap A^c$. If $x \in A$, then $x \in A \cup B$; if $x \in B \cap A^c$, then $x \in B$ so $x \in A \cup B$. Therefore, $A \cup (B \cap A^c) \subset A \cup B$. □

Remark (Disjoint partition). The proposition above provides a disjoint partition of $A \cup B$ into A and $B \cap A^c$, in the sense that A and $B \cap A^c$ are disjoint and that $A \cup (B \cap A^c) = A \cup B$. See Figure 1.1.

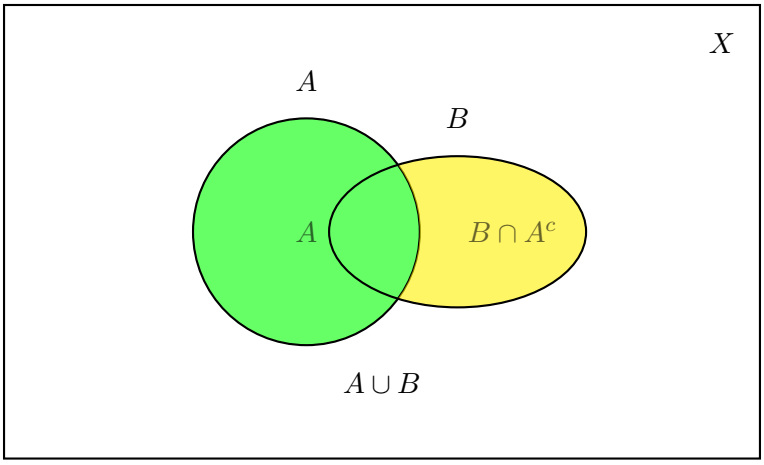


FIGURE 1.1. A disjoint partition of $A \cup B$ into A and $B \cap A^c$

Such disjoint partition can be generalized: Let $A_1, ..., A_n$ be sets. Define

$$E_1 = A_1, \quad E_2 = A_2 \cap A_1^c, \quad E_3 = A_3 \cap A_1^c \cap A_2^c, \quad ..., \quad E_n = A_n \cap A_1^c \cap \cdots \cap A_{n-1}^c,$$

that is,

$$E_i = A_i \cap \left(\bigcap_{j=1}^{i-1} A_j^c \right).$$

Then $E_1, ..., E_n$ are pairwise disjoint and that $E_1 \cup \cdots \cup E_n = A_1 \cup \cdots \cup A_n$.

THEOREM (De Morgan's laws[i]). *Let A and B_i, $i \in \mathbb{N}$, be sets. Then*

$$A \setminus \left(\bigcup_{i=1}^{\infty} B_i \right) = \bigcap_{i=1}^{\infty} A \setminus B_i \quad \text{and} \quad A \setminus \left(\bigcap_{i=1}^{\infty} B_i \right) = \bigcup_{i=1}^{\infty} A \setminus B_i.$$

[i] Augustus De Morgan, *Formal logic: or, The calculus of inference, necessary and probable.* (1847).

In particular,

$$\left(\bigcup_{i=1}^{\infty} B_i\right)^c = \bigcap_{i=1}^{\infty} B_i^c \quad \textit{and} \quad \left(\bigcap_{i=1}^{\infty} B_i\right)^c = \bigcup_{i=1}^{\infty} B_i^c.$$

Definition (Functions). Let A be a set and $\mathbb{R}$ denote the set of real numbers. A function $f : A \to \mathbb{R}$ is a correspondence that assigns to each member $x \in A$ a number in $\mathbb{R}$, denoted by $f(x)$.

Definition (Inverse image). Let $f : A \to \mathbb{R}$. Given a subset $E \subset \mathbb{R}$, we define $f^{-1}(E) = \{x \in A : f(x) \in E\}$ the inverse image of E.

For $a < b$, we introduce the following notations.

- $a < f < b$ denotes
$$f^{-1}((a,b)) = \{x \in A : a < f(x) < b\}.$$
- $a \le f < b$ denotes
$$f^{-1}([a,b)) = \{x \in A : a \le f(x) < b\}.$$
- $a < f \le b$ denotes
$$f^{-1}((a,b]) = \{x \in A : a < f(x) \le b\}.$$
- $a \le f \le b$ denotes
$$f^{-1}([a,b]) = \{x \in A : a \le f(x) \le b\}.$$
- $f > a$ denotes
$$f^{-1}((a,\infty)) = \{x \in A : f(x) > a\}.$$
- $f \ge a$ denotes
$$f^{-1}([a,\infty)) = \{x \in A : f(x) \ge a\}.$$
- $f < a$ denotes
$$f^{-1}((-\infty,a)) = \{x \in A : f(x) < a\}.$$
- $f \le a$ denotes
$$f^{-1}((-\infty,a]) = \{x \in A : f(x) \le a\}.$$
- $f = a$ denotes
$$f^{-1}(\{a\}) = \{x \in A : f(x) = a\}.$$

1.2. Combinatorics

Definition (Permutations). Let A be a discrete set. A permutation of A is an arrangement of the members of A into a sequence.

Remark. Let A be a set of size n. Then a permutation of A is a (way of) selection of the members of A *with order*. Hence, the number of permutations of A is $n!$, the factorial of n, which is defined as

$$n! = n \cdot (n-1) \cdot \cdots \cdot 2 \cdot 1.$$

Definition (Combinations). Let A be a set of size n. Suppose that $0 \le r \le n$. A r-combination of A is a subset of A of size r.

Remark. A r-combination of A with size n is a (way of) selection of r members out of the n members of A (*without order*). Hence, the number of r-combinations of A is $\binom{n}{r}$, the binomial coefficient, which is defined as

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

See Theorem 1.4. In particular,

- $\binom{n}{0} = 1$, which indicates that there is only one way to select 0 member, i.e., no member is selected.
- $\binom{n}{n} = 1$, which indicates that there is only one way to select n members, i.e., all members are selected.

- $\binom{n}{1} = n$, which indicates that there are n ways to select one member out of n members, i.e., exactly one member is selected in such a selection.
- $\binom{n}{n-1} = n$, which indicates that there are n ways to select $n-1$ members out of n members, i.e., exactly one member is not selected in such a selection.

PROPOSITION 1.2. *Let* $0 \leq r \leq n$. *Then*

$$\binom{n}{r} = \binom{n}{n-r}.$$

PROOF.

- Analytical argument: Compute that
$$\binom{n}{n-r} = \frac{n!}{(n-r)!\,(n-(n-r))!} = \frac{n!}{(n-r)!r!} = \binom{n}{r}.$$
- Combinatorial argument: Any selection of r of the n members is equivalent to a selection of $n-r$, namely, those members not selected.

□

PROPOSITION 1.3 (Pascal's identity[i]). *Let* $1 \leq r \leq n$. *Then*

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}.$$

PROOF.

- Analytical argument: Compute that
$$\begin{aligned}
\binom{n-1}{r-1} + \binom{n-1}{r} &= \frac{(n-1)!}{(r-1)!\,((n-1)-(r-1))!} + \frac{(n-1)!}{r!(n-1-r)!} \\
&= \frac{(n-1)!}{(r-1)!(n-r)!} + \frac{(n-1)!}{r!(n-r-1)!} \\
&= \frac{(n-1)!}{(r-1)!\cdot(n-r-1)!\cdot(n-r)} + \frac{(n-1)!}{(r-1)!\cdot r\cdot(n-r-1)!} \\
&= \frac{(n-1)!}{(r-1)!(n-r-1)!}\left(\frac{1}{n-r} + \frac{1}{r}\right) \\
&= \frac{(n-1)!}{(r-1)!(n-r-1)!}\cdot\frac{n}{r(n-r)} \\
&= \frac{(n-1)!\cdot n}{(r-1)!\cdot r\cdot(n-r-1)!\cdot(n-r)} \\
&= \frac{n!}{r!(n-r)!} \\
&= \binom{n}{r}.
\end{aligned}$$
- Combinatorial argument: Consider the selections of r members from a set of size n.

  On one hand, the number of the selections is $\binom{n}{r}$.

  On the other hand, fix our attention on one member, say a, from the set. There are $\binom{n-1}{r-1}$ selections such that a is selected, each of which corresponds to a selection of $r-1$ of the remaining $n-1$ members of the set; there are $\binom{n-1}{r}$ selections such that a is not selected, each of which corresponds to a selection of r of the remaining $n-1$ members of the set. The equation therefore follows.

□

[i] Blaise Pascal, *Traité du triangle arithmétique*. (1665).

Remark (Pascal's triangle). Pascal's identity is a mathematical formulation of the famous Pascal's triangle (or pyramid), a triangular arrangement of numbers that gives the coefficients in the expansion of the binomial expression $(x+y)^n$:

$$\begin{aligned}
(x+y)^0 &= 1,\\
(x+y)^1 &= x+y,\\
(x+y)^2 &= x^2+2xy+y^2,\\
(x+y)^3 &= x^3+3x^2y+3xy^2+y^3,\\
(x+y)^4 &= x^4+4x^3y+6x^2y^2+4xy^3+y^4,\\
(x+y)^5 &= x^5+5x^4y+10x^2y^3+10x^3y^3+5xy^4+y^5,\\
(x+y)^6 &= x^6+6x^5y+15x^2y^4+20x^3y^3+15x^4y^2+6xy^5+y^6.\\
&\cdots\ .
\end{aligned}$$

The coefficients are given by

$$\begin{array}{ll}
n=0: & \qquad\qquad\qquad\qquad\qquad 1\\
n=1: & \qquad\qquad\qquad\qquad 1 \qquad\qquad 1\\
n=2: & \qquad\qquad\qquad 1 \qquad\qquad 2 \qquad\qquad 1\\
n=3: & \qquad\qquad 1 \qquad\qquad 3 \qquad\qquad 3 \qquad\qquad 1\\
n=4: & \qquad\quad 1 \qquad\quad 4 \qquad\quad 6 \qquad\quad 4 \qquad\quad 1\\
n=5: & \quad 1 \qquad 5 \qquad 10 \qquad 10 \qquad 5 \qquad 1\\
n=6: & 1 \qquad 6 \qquad 15 \qquad 20 \qquad 15 \qquad 6 \qquad 1
\end{array}$$

Every term in the expansion of $(x+y)^n$ is a product of n factors, each of which is either x or y. Therefore, the number of terms which contains r factors of x (and $n-r$ factors of y) is given by $\binom{n}{r}$. It is thus called the binomial coefficient. Indeed, the above triangle in the binomial form is

$$\begin{array}{ccccccccccc}
 & & & & & \binom{0}{0} & & & & & \\
 & & & & \binom{1}{0} & & \binom{1}{1} & & & & \\
 & & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & \\
 & & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & \\
 & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & \\
\binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5}\\
\binom{6}{0} & \binom{6}{1} & & \binom{6}{2} & & \binom{6}{3} & & \binom{6}{4} & & \binom{6}{5} & \binom{6}{6}
\end{array}$$

Zooming in one of the smallest triangles as shown above, we see Pascal's identity that

$$\begin{array}{ccc}
\binom{n-1}{r-1} & & \binom{n-1}{r}\\
& \searrow \quad \swarrow & \\
& \binom{n}{r} &
\end{array}$$

In fact, we use Pascal's identity to prove the important binomial theorem:

THEOREM 1.4 (Binomial theorem).

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

PROOF. We prove by mathematical induction on n. If $n = 1$, then

$$\sum_{i=0}^{1} \binom{1}{i} x^i y^{n-i} = \binom{1}{0} x^0 y^1 + \binom{1}{1} x^1 y^0 = y + x = (x+y)^1.$$

Assume the theorem for $n-1$, i.e.,

$$(x+y)^{n-1} = \sum_{i=0}^{n-1} \binom{n-1}{i} x^i y^{n-1-i}.$$

Then

$$\begin{aligned}
(x+y)^n &= (x+y)(x+y)^{n-1} \\
&= (x+y)\left(\sum_{i=0}^{n-1} \binom{n-1}{i} x^i y^{n-1-i}\right) \\
&= x \cdot \sum_{i=0}^{n-1} \binom{n-1}{i} x^i y^{n-1-i} + y \cdot \sum_{i=0}^{n-1} \binom{n-1}{i} x^i y^{n-1-i} \\
&= \sum_{i=0}^{n-1} \binom{n-1}{i} x^{i+1} y^{n-1-i} + \sum_{i=0}^{n-1} \binom{n-1}{i} x^i y^{n-i} \\
&= \sum_{j=1}^{n} \binom{n-1}{j-1} x^j y^{n-j} + \sum_{j=0}^{n-1} \binom{n-1}{j} x^j y^{n-j} \\
&= x^n + \sum_{j=1}^{n-1} \left(\binom{n-1}{j-1} + \binom{n-1}{j}\right) x^j y^{n-j} + y^n \\
&= x^n + \sum_{j=1}^{n-1} \binom{c}{j} x^j y^{n-j} + y^n \\
&= \sum_{j=0}^{n} \binom{n}{j} x^j y^{n-j},
\end{aligned}$$

in which we used Pascal's identity in Proposition 1.3. □

COROLLARY 1.5.

$$\sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i} = 2^n.$$

PROOF.

- Analytical argument: Set $x = y = 1$ in the binomial theorem. Then

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

- Combinatorial argument: Consider the number of subsets of a set of size n.

  On one hand, the number the subsets is 2^n, since each member either belongs to a subset or does not.

  On the other hand, for each $0 \le i \le n$, there are $\binom{n}{i}$ subsets of size i. The equation therefore follows by summing $\binom{n}{i}$ with respect to $i = 0, ..., n$.

□

Remark (Partition a collection of identical objects into sub-collections). Let X be a collection of k identical objects. Consider a partition of X into n sub-collections, which are allowed to be empty. To this end, list $k+n-1$ stars in a sequence. Select $n-1$ stars and change them to bars, which then separate the remaining k stars (i.e., the original k objects) into n sub-collections. Therefore, the number of partitions of X into n sub-collections is

$$\binom{k+n-1}{n-1} = \binom{k+n-1}{k}.$$

For example, to obtain a partition of six objects by four sub-collections, among a total $6+4-1=9$ stars, select $4-1=3$ ones and change them to bars. See below for the partition by sub-collections of one object, two objects, no object, and three objects.

$$\begin{array}{ccccccccc} \star & \star & \star & \star & \star & \star & \star & \star & \star \\ \star & | & \star & \star & | & | & \star & \star & \star \end{array}$$

Homework Assignment.

Question 1.1. Let $A \subset B$. Show that $B^c \subset A^c$.

Proof. Pick $x \in B^c$. Then $x \notin B$. Hence, $x \notin A$ (otherwise, $x \in A$ so $x \in B$, because $A \subset B$). Therefore, $x \in A^c$ so $B^c \subset A^c$. □

Question 1.2. Show that $A = (A \cap B) \cup (A \cap B^c)$.

Proof. First we show that $A \subset (A \cap B) \cup (A \cap B^c)$. Pick $x \in A$. If $x \in B$, then $x \in A \cap B$; if $x \notin B$, then $x \in B^c$ so $x \in A \cap B^c$. Therefore, $A \subset (A \cap B) \cup (A \cap B^c)$.

Next we show that $(A \cap B) \cup (A \cap B^c) \subset A$. Pick $x \in (A \cap B) \cup (A \cap B^c)$. Then $x \in A \cap B$ or $x \in A \cap B^c$. If $x \in A \cap B$, then $x \in A$; if $x \in A \cap B^c$, then $x \in A$. Therefore, $(A \cap B) \cup (A \cap B^c) \subset A$. □

Question 1.3. Let $0 \leq k \leq n$. Determine the number of vectors $(x_1, ..., x_n)$ such that each x_i is either 0 or 1, and the following inequality is true.

$$\sum_{i=1}^{n} x_i \geq k.$$

Answer. Each vector $(x_1, ..., x_n)$ such that x_i is either 0 or 1, and that $x_1 + \cdots + x_n = k$, corresponds to a selection of r of its n components (and assign them as 1's). There are $\binom{n}{k}$ such vectors. Therefore, the number of vectors $(x_1, ..., x_n)$ such that each x_i is either 0 or 1, and that $x_1 + \cdots + x_n \geq k$, is

$$\binom{n}{k} + \binom{n}{k+1} + \cdots + \binom{n}{n} = \sum_{i=k}^{n} \binom{n}{i}.$$

Question 1.4. Show that

$$\sum_{i=0}^{n} (-1)^i \binom{n}{i} = 0.$$

Proof. Set $x = -1$ and $y = 1$ in the binomial theorem. Then

$$0 = (-1+1)^n = \sum_{i=0}^{n} \binom{n}{i} (-1)^i 1^{n-i} = \sum_{i=0}^{n} (-1)^i \binom{n}{i}.$$

□

Question 1.5. Use a combinatorial argument to show that

$$\binom{n+m}{r} = \binom{n}{0}\binom{m}{r} + \binom{n}{1}\binom{m}{r-1} + \cdots + \binom{n}{r-1}\binom{m}{1} + \binom{n}{r}\binom{m}{0} = \sum_{i=0}^{r} \binom{n}{i}\binom{m}{r-i}.$$

PROOF. Select r members from a group which is consisted of n men and m women.

On one hand, the number of selections is $\binom{n+m}{r}$.

On the other hand, for each $0 \leq i \leq r$, there are $\binom{n}{i}\binom{m}{r-i}$ selections for which i men and $r-i$ women are selected. The equation therefore follows by summing $\binom{n}{i}\binom{m}{r-i}$ with respect to $i = 0, ..., r$. □

Question 1.6. Consider a smooth function $f(x_1, ..., x_n)$ of n variables. Find the number of partial derivatives of f of order k.

Answer. Each partial derivative of order k corresponds to a partition of a collection of k objects (i.e., orders here) into n sub-collections (which are allowed to be empty). Therefore, the number of partial derivatives of f of order k is

$$\binom{k+n-1}{n-1}.$$

For example, the partition of six objects into four sub-collections of one object, two objects, no objects, and three objects, as demonstrated above, gives a sixth order partial derivative $\partial_{x_1}\partial_{x_2}^2\partial_{x_4}^3 f$ of a function $f(x_1, x_2, x_3, x_4)$ of four variables.

Question 1.7. Determine the number of vectors $(x_1, ..., x_n)$ such that each x_i is a non-negative integer, and the following inequality is true.

$$\sum_{i=1}^{n} x_i \leq k.$$

Answer. Each vector $(x_1, ..., x_n)$ such that x_i is a non-negative integer, and that $x_1 + \cdots + x_n = k$, corresponds to a partition of a collection of k objects (1's here) into n sub-collections. There are $\binom{k+n-1}{n-1}$ such vectors. Therefore, the number of vectors $(x_1, ..., x_n)$ such that each x_j is a non-negative integer, and that $x_1 + \cdots + x_n \leq k$, is

$$\binom{n-1}{n-1} + \binom{1+n-1}{n-1} + \cdots + \binom{k+n-1}{n-1} = \sum_{i=0}^{k}\binom{i+n-1}{n-1} = \sum_{i=0}^{k}\binom{i+n-1}{i}.$$

CHAPTER 2

Axioms of probability

The theory of probability, as a mathematical discipline, can and should be developed from axioms in exactly the same way as Geometry and Algebra. This means that after we have defined the elements to be studied and their basic relations, and have stated the axioms by which these relations are to be governed, all further exposition must be based exclusively on these axioms, independent of the usual concrete meaning of these elements and their relations.

Andrey Kolmogorov, 1933[i]

All possible definitions of probability fall short of the actual practice.

William Feller, 1968[ii]

The history of probability (and of mathematics in general) shows a stimulating interplay of theory and applications: progress in theory opens new fields of applications, and each new application creates new theoretical problems and influences the direction of research. Today applications of the theory of probability extend over many fields of different natures, and the number of applications is fast increasing. Only a general mathematical theory is flexible enough to provide proper tools for such a variety of problems, and we must withstand the temptation (and the pressure) to keep our notions, pictures, and terms too close to one particular field of experience. We require a rigorous mathematical theory proceeding along the lines which are generally accepted in Geometry and Algebra.

We study in Geometry the idealized and abstract shapes such as triangles, and the theories that govern the relations among the geometric quantities such as the Pythagorean theorem. The power of Geometry lies in the fact that the shapes and theories are not tied with a particular field of applications, say, building a bridge. Though the example of building a bridge certainly helps us understand a triangle, and also convinces us the usefulness of Geometry.

We study in Algebra the idealized and abstract variables, and the theories that govern the relations among the variables such as the quadratic formula. The power of Algebra lies in the fact that the variables and formulas are not tied with a particular field of applications, say, throwing a baseball. Though the example of throwing a baseball certainly helps us understand a parabola as the curve of a quadratic function, and also convinces us the usefulness of Algebra.

The modern mathematical theory of probability is concerned with one particular aspect of "chance". In a rough way, we may characterize this aspect by the probabilities of possible outcomes of "a conceptual experiment", such as tossing a coin. At the outset, we must agree on the possible *outcomes* of this experiment (i.e., the *sample space*) and the *probabilities* associated with them.

The power of Probability lies in the fact that the outcomes, sample spaces, and probabilities are not tied with a particular field of applications, say, counting heads in coin tosses. Though the example of coin tosses certainly helps us understand these concepts. We therefore begin from introducing several examples, through which the intuition of the axioms of probability that are defined later.

Example (Toss a coin). Consider the experiment of tossing a coin. Then the sample space $S = \{\mathrm{h}, \mathrm{t}\}$ is the set of two possible outcomes of heads (h) and tails (t). Therefore, the probability that

[i] Andrey Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung.* [Foundations of the theory of probability]. (1933).

[ii] William Feller, *An introduction to probability theory and its applications.* (1968).

the experiment lands on heads or on tails is $1 = 100\%$. Moreover, if the probability of the experiment landing on heads is $p \in [0, 1]$, then the one on tails is $1 - p$. This indicates that, if we toss the coin for a great many times[i], then the percentage of landing on heads in these experiments tends to p, and the one of landing on tails tends to $1 - p$. We say that the coin is fair if $p = \frac{1}{2}$ and is unfair if $p \neq \frac{1}{2}$.

The coin toss can be considered as an idealized model of experiments with only two outcomes, whose probabilities of occurrence are p and $1 - p$, respectively. Such examples, in addition to the coin toss, include success or failure of a medicine trial, win or loss of a sport game, pass or fail in a course, etc. Models with more than two outcomes include the die rolls.

Example (Roll a die). Consider the experiment of rolling a die. Then the sample space $S = \{1, 2, 3, 4, 5, 6\}$ is the set of six possible outcomes. Therefore, the probability that the experiment lands on one of the six numbers is 1. We say that the die is fair if the probability that the experiment lands on each number is $\frac{1}{6}$. This indicates that, if we roll a fair die for a great many times, then the percentage of rolls which land on any one number among $1, 2, 3, 4, 5, 6$ in these experiments tends to $\frac{1}{6}$.

In this setup, we can compute the probability of any event. It is represented by a subset of S, which should always have probability bounded between 0 and 1. For example, the event that the die lands on an even number is represented by $E = \{2, 4, 6\}$, which has a probability of $\frac{3}{6} = \frac{1}{2}$; whereas the event that the die lands on a number smaller than 4 is represented by $F = \{1, 2, 3\}$, which also has a probability of $\frac{3}{6} = \frac{1}{2}$. However, the event $E \cup F$, i.e., that the die lands on an even number or a number smaller than 4, is represented by $\{1, 2, 3, 4, 6\}$, which has a probability of $\frac{5}{6} \neq \frac{1}{2} + \frac{1}{2}$. These examples show that the probability is *not* additive in general, and is so if the events in question are mutually exclusive.

Definition (Probability spaces). A probability space is a set S which is equipped with a function $P : \mathcal{P}(S) \to [0, 1]$ such that the following conditions hold. (Here, $\mathcal{P}(S)$ is the power set of S, i.e., the set of all subsets of S, see Chapter 1.)

(i). For each $E \subset S$, $0 \leq P(E) \leq 1$.
(ii). $P(S) = 1$.
(iii). If $E_1, E_2, ...$ are pairwise disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P\left(E_i\right).$$

We call S the sample space, $s \in S$ an outcome, and $E \subset S$ an event and $P(E)$ the probability that E occurs.

The following table demonstrates the correspondence of languages in the set theory and in the probability theory.

[i]Strictly speaking, we would need to assume that these experiments are independent, see Chapter 3 for the formal discussion of independence in probability. Moreover, the percentage of tosses which land on heads in these experiments tends to p is a phenomenon of "the law of large numbers", that is, the average of the results obtained from a large number of independent and identical experiments should be close to the expected value and tends to become closer to the expected value as more experiments are performed.

	Set theory	Probability theory
S	the universal set	the sample space
$s \in S$	s is a member or an element.	s is an outcome.
$E \subset S$	E is a subset.	E is an event.
$\{s\} \subset S$	$\{s\}$ is a singleton set.	$\{s\}$ is a sample point.
$E \subset F$	E is a subset of F.	If E has occurred, then F occurs.
$E \cap F$	the intersection of E and F.	the event that both E and F occur
$E \cap F = \emptyset$	E and F are disjoint.	E and F are mutually exclusive.
$E \cup F$	the union of E and F.	the event that either E or F occurs
$E \setminus F$	the complement of F in E.	the event that E occurs and F does not occur
E^c	the complement of E.	the event that E does not occur

Definition (Almost surely events). Let S be a probability space with a probability P. We say that an event $E \subset S$ occurs almost surely if $P(E) = 1$.

Remark (Discrete probability space). In this book, our main focus is on the discrete probability spaces, that is, the sample space $S = \{s_1, ..., s_n\}$ for some $n \in \mathbb{N}$ or $S = \{s_1, s_2, ...\}$. In this chapter, all examples, except the last one, are discrete.

Remark (Finite probability space with equally likely outcomes). Let $S = \{s_1, ..., s_n\}$ be a finite probability space such that each sample point has the same probability, i.e., all outcomes are equally likely to occur. Since $S = \{s_1\} \cup \cdots \cup \{s_n\}$ is a disjoint union,

$$1 = P(S) = P(\{s_1\}) + \cdots + P(\{s_n\}) = n \cdot P(\{s_1\}).$$

which implies that

$$P(\{s_1\}) = \cdots = P(\{s_n\}) = \frac{1}{n}.$$

Therefore, given $E \subset S$,

$$P(E) = \frac{|E|}{n}.$$

Example. Suppose that the sample space S is the College of Mathematics and Computer Science (CS) consisted of 100 students. Equip S with a probability P such that each outcome has the same probability (of $\frac{1}{100}$). This indicates that, if we perform an experiment of drawing a student from the College according to the probability P, then each student has an equal chance of being chosen.

Assume that 20 students in the College are mathematics-majored. Denote this set by E. Then $P(E) = \frac{20}{100} = \frac{1}{5}$. This indicates that, if we perform such experiments for a great many times, then the percentage of meeting a mathematics-majored student in these experiments tends to $\frac{1}{5}$.

PROPOSITION 2.1. *Let S be a probability space with probability P. Suppose that $E, F \subset S$.*

(i). $P(E^c) = 1 - P(E)$.
(ii). $P(\emptyset) = 0$.
(iii). $P(E) = P(E \cap F) + P(E \cap F^c)$.
(iv). If $E \subset F$, then $P(E) = P(F) - (F \cap E^c)$, in particular, $P(E) \leq P(F)$.

PROOF.

(i). Since $S = E \cup E^c$ is a disjoint union,

$$1 = P(S) = P(E \cup E^c) = P(E) + P(E^c).$$

Hence,

$$P(E^c) = 1 - P(E).$$

(ii). Take $E = S$ in (i). Since $S^c = \emptyset$,

$$P(\emptyset) = P(S^c) = 1 - P(S) = 0.$$

(iii). Since $E = (E \cap F) \cup (E \cap F^c)$ is a disjoint union,

$$P(E) = P\left((E \cap F) \cup (E \cap F^c)\right) = P(E \cap F) + (E \cap F^c)\,.$$

(iv). Since $E \subset F$, $F \cap E = E$. By (iii),

$$P(F) = P(F \cap E) + P\left(F \cap E^c\right) = P(E) + P\left(F \cap E^c\right).$$

Hence,

$$P(E) = P(F) - P\left(F \cap E^c\right) \leq P(F),$$

because $(F \cap E^c) \geq 0$.

□

Example. Consider the length of calls (in minutes) a call center receives in a day. Then the sample space is $S = (0, \infty)$. For $a < b$, $(a, b) \subset S$ denotes the event that the length of calls is between a minutes and b minutes. Let

$$P((a, b)) = \int_a^b e^{-x}\, dx.$$

Then P defines a probability on S. For example,

(a). The probability that the length of calls is shorter than two minutes is

$$P((0, 2)) = \int_0^2 e^{-x}\, dx = 1 - e^{-2} \approx 0.865.$$

(b). The probability that the length of calls is shorter than three minutes is

$$P((0, 3)) = \int_0^3 e^{-x}\, dx = 1 - e^{-3} \approx 0.950.$$

(c). The probability that the length of calls is between two and three minutes is

$$P((2, 3)) = \int_2^3 e^{-t}\, dt = e^{-2} - e^{-3} \approx 0.086.$$

Homework Assignment.

Question 2.1. Let E, F, and G be three events. Find expressions for the events so that, of E, F, and G,

(a). only E occurs;
(b). both E and G, but not F, occur;
(c). at least one of the events occurs;
(d). at least two of the events occur;
(e). all three events occur;
(f). none of the events occurs;
(g). at most one of the events occurs;
(h). at most two of the events occur;
(i). exactly two of the events occur.

Answer.

(a). $E \cap F^c \cap G^c$;
(b). $E \cap F^c \cap G$;
(c). $E \cup F \cup G$;
(d). $(E \cap F) \cup (E \cap G) \cup (F \cap G)$;
(e). $E \cap F \cap G$;
(f). $E^c \cap F^c \cap G^c (= (E \cup F \cup G)^c)$;
(g). $(E \cap F^c \cap G^c) \cup (E^c \cap F \cap G^c) \cup (E^c \cap F^c \cap G) \cap (E^c \cap F^c \cap G^c)$;
(h). $(E \cap F \cap G)^c (= E^c \cup F^c \cup G^c)$;
(i). $(E^c \cap F \cap G) \cup (E \cap F^c \cap G) \cup (E \cap F \cap G^c)$.

Question 2.2. Let S be a probability space with probability P. Suppose that $E, F \subset S$. Show that $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

PROOF. Since $E \cup F = E \cup (F \cap E^c)$, the latter of which is a disjoint union,

$$P(E \cup F) = P\left(E \cup (F \cap E^c)\right) = P(E) + P\left(F \cap E^c\right) = P(E) + P(F) - P(F \cap E),$$

Here, we used the fact that

$$P\left(F \cap E^c\right) = P(F) - P(F \cap E),$$

by (iii) in Proposition 2.1 that $P(F) = P(F \cap E) + P(F \cap E^c)$. □

Question 2.3. Let S be a probability space with probability P. Suppose that $E, F \subset S$. Describe the event that exactly one of E or F occurs, and find the probability of such event.

Answer. The event that exactly one of E or F occurs is $(E \cap F^c) \cup (F \cap E^c)$. It is a disjoint union. Hence,

$$\begin{aligned} P\left((E \cap F^c) \cup (F \cap E^c)\right) &= P\left((E \cap F^c)\right) + P\left((F \cap E^c)\right) \\ &= P(E) - P(E \cap F) + P(F) - P(F \cap E) \\ &= P(E) + P(F) - 2P(E \cap F). \end{aligned}$$

CHAPTER 3

Conditional probability and independence

It seems that to make a correct conjecture about any event whatever, it is necessary to calculate exactly the number of possible cases and then to determine how much more likely it is that one case will occur than another.

Jacob Bernoulli, 1713[i]

The concept of "Events occur independently with each other" has its intuitive meaning in practice. In fact, we have already encountered this concept in Chapter 2, for example, perform two "independent" experiments of coin tosses, the intuitive meaning of such is that the outcomes of landing on heads and on tails in the first experiment do not change the chances of the outcomes in the second experiment. In other words, on the condition that the first experiment lands on heads or on tails, the probability of landing on heads or on tails in the second experiment stays the same (as if the first experiment has not been performed). This, in turn, demands an introduction of conditional probability, which is itself an intuitive concept but can be clumsy in terminology.

Let S be a probability space with a probability P throughout the chapter.

3.1. Conditional probability

Definition (Conditional probability). Let $E, F \subset S$ and $P(F) > 0$. Then the conditional probability of E for given F (also called the conditional probability that E occurs given that F has occurred) is defined as

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

Remark. The assumption that $P(F) > 0$ in the definition of conditional probability is only a technical one, since it appears in the denominator.

The conditional probability for given an event F can be understood as the probability in which F is regarded as the new sample space. All events E are now considered in this sample space so $E \cap F$, and we then need to rescale P by a factor $P(F)$.

Example. Suppose that the College of Mathematics and Computer Science (CS) is consisted of 20 mathematics-majored students and 80 CS-majored students. Equip this sample space S with a probability P such that each sample point has the same probability (of $\frac{1}{100}$). Let E be the set of mathematics-majored students and $F = E^c$ be the set of CS-majored students.

Suppose that the class of *Introduction to Probability* is consisted of four mathematics-majored students and six CS-majored students, denoted by the set G. Then

(a).

$$P(E|G) = \frac{P(E \cap G)}{P(G)} = \frac{|E \cap G|}{|G|} = \frac{4}{10} = 0.4,$$

which indicates that on the condition that a student enrolls the class, the conditional probability that she is mathematics-majored is 0.4. Equivalently, the class G is now the new sample space and

[i]Jacob Bernoulli, *Ars Conjectandi*. (1713).

the conditional probability of E is basically the probability that a student is mathematics-majored in this sample space.

(b).
$$P(F|G) = \frac{P(F \cap G)}{P(G)} = \frac{|F \cap G|}{|G|} = \frac{6}{10} = 0.6,$$
which indicates that on the condition that a student enrolls the class, the conditional probability that she is CS-majored is 0.6. Equivalently, the class G is now the new sample space and the conditional probability of F is basically the probability that a student is CS-majored in this sample space.

(c).
$$P(G|E) = \frac{P(G \cap E)}{P(E)} = \frac{|G \cap E|}{|E|} = \frac{4}{20} = 0.2,$$
which indicates that on the condition that a student is mathematics-majored, the conditional probability that she enrolls the class is 0.2. Equivalently, the mathematics department E is now the new sample space and the conditional probability of G is basically the probability that a student enrolls the class in this sample space.

(d).
$$P(G|F) = \frac{P(G \cap F)}{P(F)} = \frac{|G \cap F|}{|F|} = \frac{6}{80} = 0.075,$$
which indicates that on the condition that a student is CS-majored, the probability that she enrolls the class is 0.075. Equivalently, the CS department F is now the new sample space and the conditional probability of G is basically the probability that a student enrolls the class in this sample space.

Remark. Let $E \subset F$ such that $P(F), P(F^c) > 0$. Since $E \cap F = E$,
$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{P(E)}{P(E)} = 1,$$
that is, the conditional probability of F given that E has occurred is 1. This is natural since $E \subset F$ means that if E has occurred, then F occurs.

On the other hand, since $E \cap F^c = \emptyset$,
$$P\left(E|F^c\right) = \frac{P\left(E \cap F^c\right)}{P\left(F^c\right)} = \frac{P(\emptyset)}{P\left(F^c\right)} = 0,$$
that is, the conditional probability of E occurs given that F^c has occurred (i.e., F has not occurred) is 0. This is again natural since $E \subset F$, $F^c \subset E^c$. So if F^c has occurred, then E^c occurs, hence E does not occur.

PROPOSITION 3.1. *Let $E, F \subset S$. Then*
$$P(E \cap F) = P(E|F)P(F).$$
More generally, for $E_1, ..., E_n \subset S$,
$$P\left(E_1 \cap \cdots \cap E_n\right) = P\left(E_1\right) P\left(E_2|E_1\right) P\left(E_3|E_1 \cap E_2\right) \cdots P\left(E_n|E_1 \cap \cdots \cap E_{n-1}\right).$$

This proposition offers a useful method to compute probabilities, especially in the case when the sample space is difficult to determine, but the conditional probabilities can be found more easily (because the new sample spaces in the conditional probability are easy to describe.)

Example. Four of the eight teams in the quarterfinal round of the 2023 European Champions League Football tournament were the acknowledged strong teams Real Madrid, Bayern Munich, Milan, and Inter. All possible pairings in this round are equally likely. Find the probability that none of the strong teams play each other.

• Method I without conditional probability: Let $a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4$ denote the eight teams, in which a_1, a_2, a_3, a_4 are the strong teams and b_1, b_2, b_3, b_4 are the weak teams. Let S be the sample space of outcomes. Each outcome is an arrangement of the eight teams into four pairs:

$$\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ y_1 & y_2 & y_3 & y_4 \end{array}$$

in which $x_1, x_2, x_3, x_4, y_1, y_2, y_3, y_4$ are placeholders that are taken by the eight teams.

To determine an outcome, we first select four teams to occupy x_1, x_2, x_3, x_4, which are $\binom{8}{4}$ ways, then we select their opponents (i.e., to occupy y_1, y_2, y_3, y_4), which are 4! ways. But there are 2^4 ways to produce the same outcome in pairing, each of which corresponds to a switch of places between x_1 and y_1, x_2 and y_2, x_3 and y_3, x_4 and y_4. Therefore, the number of outcomes in the sample space S is

$$|S| = \frac{\binom{8}{4} \cdot 4!}{2^4} = 105.$$

Let E be the event that strong teams do not play each other, i.e., strong teams play weak teams. To determine an outcome in E, we let $x_1 = a_1$, $x_2 = a_2$, $x_3 = a_3$, and $x_4 = a_4$, then we select their opponents from the four weak teams (i.e., to occupy y_1, y_2, y_3, y_4), which corresponds to a permutation of four objects. Hence, there are $4! = 24$ outcomes in E. Therefore,

$$P(E) = \frac{|E|}{|S|} = \frac{24}{105} = \frac{8}{35} \approx 0.229.$$

• Method II with conditional probability: For $i = 1, 2, 3, 4$, let E_i be the event that the strong team a_i plays one of the four weak teams. Then the event that strong teams do not play each other, i.e., strong teams play weak teams, is $E_1 \cap E_2 \cap E_3 \cap E_4$. Hence,

$$\begin{aligned} & P\left(E_1 \cap E_2 \cap E_3 \cap E_4\right) \\ = \; & P\left(E_1\right) P\left(E_2|E_1\right) P\left(E_3|E_1 \cap E_2\right) P\left(E_4|E_1 \cap E_2 \cap E_3\right) \\ = \; & \frac{4}{7} \cdot \frac{3}{5} \cdot \frac{2}{3} \cdot \frac{1}{1} \\ = \; & \frac{8}{35}, \end{aligned}$$

which follows by examining the conditional probabilities.

(a). Since there are four weak teams out of seven (i.e., the eight teams minus a_1), the probability $P(E_1) = \frac{4}{7}$.

(b). On the condition that a_1 has been assigned a weak team, there are three weak teams out of five (i.e., the eight teams minus a_1 and its opponent, and a_2). The conditional probability $P(E_2|E_1) = \frac{3}{5}$.

(c). On the condition that a_1 and a_2 have been assigned two weak teams, there are two weak teams out of three (i.e., the eight teams minus a_1 and its opponent, a_2 and its opponent, and a_3). The conditional probability $P(E_3|E_1 \cap E_2) = \frac{2}{3}$.

(d). On the condition that a_1, a_2, a_3 have been assigned three weak teams, there is one weak team out of one (i.e., the eight teams minus a_1 and its opponent, a_2 and its opponent, a_3 and its opponent, and a_4). The conditional probability $P(E_4|E_1 \cap E_2 \cap E_3) = \frac{1}{1}$.

THEOREM 3.2 (Bayes' formula[i]). *Let $E, F \subset S$ such that $P(F), P(F^c) > 0$. Then*

$$P(E) = P(E|F)P(F) + P\left(E|F^c\right) P\left(F^c\right) = P(E|F)P(F) + P\left(E|F^c\right)\left(1 - P(F)\right),$$

that is, the probability of E is the sum of the conditional probability of E for given F and the conditional probability of E for given F^c.

[i]Thomas Bayes, *An essay towards solving a problem in the doctrine of chances.* (1763).

PROOF. Since E is a disjoint union of $E \cap F$ and $E \cap F^c$,

$$\begin{aligned} P(E) &= P(E \cap F) + P(E \cap F^c) \\ &= P(E|F)P(F) + P(E|F^c)\,P(F^c) \\ &= P(E|F)P(F) + P(E|F^c)\,(1 - P(F)). \end{aligned}$$

□

3.2. Independence

Definition (Independence). Let $E, F \subset S$ such that $P(F) > 0$. We say that E is independent of F, if

$$P(E|F) = P(E),$$

or equivalently,

$$P(E \cap F) = P(E)P(F).$$

Remark. If E is independent of F and $P(E) > 0$, then $P(E \cap F) = P(E)P(F)$ so

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{P(F)P(E)}{P(E)} = P(F),$$

which means that F is independent of E. Therefore, we can say that E and F are independent.

PROPOSITION 3.3. *Let E and F be independent. Then E and F^c are independent.*

PROOF. Since E is a disjoint union of $E \cap F$ and $E \cap F^c$,

$$\begin{aligned} P(E) &= P(E \cap F) + P(E \cap F^c) \\ &= P(E)P(F) + P(E \cap F^c), \end{aligned}$$

in which we used the fact that E and F are independent so $P(E \cap F) = P(E)P(F)$. Therefore,

$$P(E \cap F^c) = P(E) - P(E)P(F) = P(E)\,(1 - P(F)) = P(E)P(F^c),$$

which means that E and F^c are independent. □

Remark. Let $E, F \subset S$ such that E is independent of F. Then it follows that E and F are independent, E and F^c are independent, and E^c and F^c are independent. That is, knowing that E has occurred (or has not) occurred does not change the probability of occurrence of F (or F^c), and vice versa. This explains the meaning of "Two events occur independently."

Remark. Independence and mutual exclusiveness are different concepts. On one hand, two mutually exclusive events are dependent. Indeed, knowing that one event has occurred changes the probability of the occurrence of the other to 0, since they are mutually exclusive.

On the other hand, two independent events E and F are in general not mutually exclusive. Indeed, since $E \cap F = \emptyset$, we have that $P(E \cap F) = P(E)P(F)$ only if $P(E) = 0$ or $P(F) = 0$.

Example. Suppose that the College of Mathematics and Computer Science (CS) is consisted of 20 mathematics-majored students and 80 CS-majored students. Equip this sample space S with a probability P such that each sample point has the same probability (of $\frac{1}{100}$). Let E be the set of mathematics-majored students and $F = E^c$ be the set of CS-majored students. Then

$$P(E) = \frac{|E|}{|S|} = \frac{20}{100} = 0.2 \quad \text{and} \quad P(F) = \frac{|F|}{|S|} = \frac{80}{100} = 0.8.$$

Notice that since $E \cap F = \emptyset$, E and F are mutually exclusive, and $P(E|F) = \frac{P(E \cap F)}{P(F)} = 0$, which does not equal $P(E) = 0.2$. This indicates that E and F are dependent. Indeed, knowing that a student is CS-majored changes the probability of she being mathematics-majored (from 0.2) to 0, and vice versa.

Situation I. Suppose that the class of *Introduction to Probability* is consisted of two mathematics-majored students and eight CS-majored students, denoted by the set G. Then

$$P(G) = \frac{|G|}{|S|} = \frac{10}{100} = 0.1.$$

(a).
$$P(E|G) = \frac{P(E \cap G)}{P(G)} = \frac{|E \cap G|}{|G|} = \frac{2}{10} = 0.2 = P(E),$$
which indicates that E is independent of G. That is, knowing that a student enrolls the class does not change the probability of she being mathematics-majored (in the College).

(b).
$$P(F|G) = \frac{P(F \cap G)}{P(G)} = \frac{|F \cap G|}{|G|} = \frac{8}{10} = 0.8 = P(F),$$
which indicates that F and G are independent. That is, knowing that a student enrolls the class does not change the probability of she being CS-majored (in the College).

(c).
$$P(G|E) = \frac{P(G \cap E)}{P(E)} = \frac{|G \cap E|}{|E|} = \frac{2}{20} = 0.1 = P(G),$$
which indicates that G is independent of E. That is, knowing that a student is mathematics-majored does not change the probability of she enrolling the class.

(d).
$$P(G|F) = \frac{P(G \cap F)}{P(F)} = \frac{|G \cap F|}{|F|} = \frac{8}{80} = 0.1 = P(G),$$
which indicates that G is independent of E. That is, knowing that a student is CS-majored does not change the probability of she enrolling the class.

In Situation I, we also verify Bayes' formula, noting that $E^c = F$:

$$P(G) = P(G|E)P(E) + P(G|F)P(F) = 0.1 \cdot 0.2 + 0.1 \cdot 0.8 = 0.1.$$

Situation II. Suppose that the class of *Introduction to Probability* is consisted of four mathematics-majored students and six CS-majored students, denoted by the set G. Then

$$P(G) = \frac{|G|}{|S|} = \frac{10}{100} = 0.1.$$

(a).
$$P(E|G) = \frac{P(E \cap G)}{P(G)} = \frac{|E \cap G|}{|G|} = \frac{4}{10} = 0.4 \neq P(E) = 0.2,$$
which indicates that E is dependent of G. That is, knowing that a student enrolls the class changes the probability of she being mathematics-majored. Indeed, this knowledge increases the probability from 0.2 to 0.4, i.e., a student in the class is more likely mathematics-majored (compared with a student from the College).

(b).
$$P(F|G) = \frac{P(F \cap G)}{P(G)} = \frac{|F \cap G|}{|G|} = \frac{6}{10} = 0.6 \neq P(F) = 0.8,$$
which indicates that F and G are dependent. That is, knowing that a student enrolls the class changes the probability of she being CS-majored. Indeed, this knowledge reduces the probability from 0.8 to 0.6, i.e., a student in the class is less likely CS-majored (compared with a student from the College).

(c).
$$P(G|E) = \frac{P(G \cap E)}{P(E)} = \frac{|G \cap E|}{|E|} = \frac{4}{20} = 0.2 \neq P(G) = 0.1,$$
which indicates that G is dependent of E. That is, knowing that a student is mathematics-majored changes the probability of she enrolling the class. Indeed, this knowledge increases the probability from 0.1 to 0.2, i.e., a mathematics-majored student is more likely to enroll the class (compared with a student from the College).

(d).
$$P(G|F) = \frac{P(G \cap F)}{P(F)} = \frac{|G \cap F|}{|F|} = \frac{6}{80} = 0.075 \neq P(G) = 0.1,$$
which indicates that G is dependent of E. That is, knowing that a student is CS-majored changes the probability of she enrolling the class. Indeed, this knowledge reduces the probability from 0.1 to 0.075, i.e., a CS-majored student is less likely to enroll the class (compared with a student from the College).

In Situation II, we also verify Bayes' formula, noting that $E^c = F$:
$$P(G) = P(G|E)P(E) + P(G|F)P(F) = 0.2 \cdot 0.2 + 0.075 \cdot 0.8 = 0.1.$$

Definition. Let $E, F, G \subset S$. We say that E, F, G are pairwise independent if
$$P(E \cap F) = P(E)P(F), \quad P(E \cap G) = P(E)P(G), \quad P(F \cap G) = P(F)P(G),$$
and if, in addition,
$$P(E \cap F \cap G) = P(E)P(F)P(G).$$
then we say that they are independent.

Example. Consider two independent tosses of a fair coin. Then the sample space is
$$S = \{(\mathrm{h}, \mathrm{h}), (\mathrm{h}, \mathrm{t}), (\mathrm{t}, \mathrm{h}), (\mathrm{t}, \mathrm{t})\}.$$
Let A be the event that the first toss results in heads, let B be the event that the second toss results in heads, and let C be the event that in both tosses the coin lands on the same side. Then
$$A = \{(\mathrm{h}, \mathrm{h}), (\mathrm{h}, \mathrm{t})\}, \quad B = \{(\mathrm{h}, \mathrm{h}), (\mathrm{t}, \mathrm{h})\}, \quad C = \{(\mathrm{h}, \mathrm{h}), (\mathrm{t}, \mathrm{t})\},$$
each of which has probability $\frac{1}{2}$. In addition,
$$A \cap B = \{(\mathrm{h}, \mathrm{h})\}, \quad A \cap C = \{(\mathrm{h}, \mathrm{h})\}, \quad B \cap C = \{(\mathrm{h}, \mathrm{h})\}.$$
each of which has probability $\frac{1}{4}$. This shows that A, B, C are pairwise independent. However,
$$A \cap B \cap C = \{(\mathrm{h}, \mathrm{h})\},$$
whose probability is $\frac{1}{4}$, and is not equal to
$$P(A)P(B)P(C) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}.$$
This shows that A, B, C are dependent. This is reflected in the fact that knowing A and B changes the probability of C (to 1).

The concept of independence can be generalized to any collection of events:

Definition. Let $E_1, ..., E_n \subset S$. We say that $E_1, ..., E_n$ are pairwise independent if
$$P(E_i \cap E_j) = P(E_i) P(E_j) \quad \text{for all } i, j = 1, ..., n.$$
If for each sub-collection of sets $E_{i_1}, ..., E_{i_k}$,
$$P(E_{i_1} \cap \cdots \cap E_{i_k}) = P(E_{i_1}) \cdots P(E_{i_k}).$$
then we say that they are independent.

3.3. Independent trials

The notion of independence enables us to formulate analytically the intuitive concept of "independent experiments".

Example. Consider two experiments, first of which is a toss of a fair coin and second is a roll of a fair die. Then the probability space for the first experiment is $S_1 = \{\mathrm{h}, \mathrm{t}\}$ equipped with a probability P_1 that $P_1(\{\mathrm{h}\}) = P_1(\{\mathrm{t}\}) = \frac{1}{2}$ and the second is $S_2 = \{1, 2, 3, 4, 5, 6\}$ equipped with a probability P_2 that $P_2(\{1\}) = P_2(\{2\}) = P_2(\{3\}) = P_2(\{4\}) = P_2(\{5\}) = P_2(\{6\}) = \frac{1}{6}$. Hence, the sample space for two experiments is

$$S = \{(s_1, s_2) : s_1 \in \{\mathrm{h}, \mathrm{t}\}, s_2 \in \{1, 2, 3, 4, 5, 6\}\} = S_1 \times S_2.$$

The two experiments are independent, if the outcomes of the first experiment do not change the chances of the outcomes in the second one, and vice versa. Hence, as an example, the event that that the first toss lands on heads and the second lands on 1 is represented by $\{(\mathrm{h}, 1)\}$, which has probability $\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$.

Indeed, we say that the two experiments are independent if S is equipped with a probability[i] P that is defined by

$$P(\{s_1, s_2\}) = P_1(\{s_1\}) P_2(\{s_2\}) = \frac{1}{12} \quad \text{for all } s_1 \in S_1, s_2 \in S_2.$$

In particular, the probabilities of the events in each experiment stay the same, *independent* of the other one, for instance,

$$E = \{(\mathrm{h}, 1), (\mathrm{h}, 2), (\mathrm{h}, 3), (\mathrm{h}, 4), (\mathrm{h}, 5), (\mathrm{h}, 6)\}$$

represents the event that the first experiment lands on heads. It has probability $P(E) = \frac{6}{12} = \frac{1}{2}$, which coincides with the one in the single-round experiment of tossing a fair coin.

We make a direct generalization of the example above: For each $i = 1, ..., n$, perform an experiment with outcomes in a finite probability space S_i equipped with a probability P_i. Consider the sample space

$$S = \{(s_1, ..., s_n) : s_1 \in S_1, ..., s_n \in S_n\} = S_1 \times \cdots \times S_n.$$

We say that the experiments are independent if S is equipped with a probability P that is defined by

$$P(\{s_1, ..., s_n\}) = P_1(\{s_1\}) \cdots P_n(\{s_n\}) \quad \text{for all } s_1 \in S_1, ..., s_n \in S_n.$$

If S_i and P_i are identical for all $i = 1, ..., n$, then the experiments are said to be *independent trials*.

PROPOSITION 3.4. *Suppose that n independent experiments are performed with respect to probability spaces S_i equipped with probabilities P_i, $i = 1, ..., n$. Assume that E_i is an event that only depends on the i-th experiment, $i = 1, ..., n$. Then $E_1, ..., E_n$ are independent.*

Example (Independent tosses of a coin). Consider n independent tosses of a coin, which lands on heads with probability p. Then the sample space

$$S = \{(s_1, ..., s_n) : s_1, ..., s_n \in \{\mathrm{h}, \mathrm{t}\}\}$$

is equipped with a probability P such that

$$P(\{s_1, ..., s_n\}) = P(\{s_1\}) \cdots P(\{s_n\}) \quad \text{for all } s_1, ..., s_n \in \{\mathrm{h}, \mathrm{t}\}.$$

Here, $P(\{s_i\}) = p$ if $s_i = \mathrm{h}$ and $P(\{s_i\}) = 1 - p$ if $s_i = \mathrm{t}$. We can then derive the probabilities of any event in these experiments. For example, let E be the event that exactly r heads are obtained. Then

$$P(E) = \binom{n}{r} p^r (1-p)^{n-r}.$$

Indeed, each outcome in E corresponds to a selection of r members out of $s_1, ..., s_n$, in which heads are obtained so the remaining ones obtain tails. There are $\binom{n}{r}$ such outcomes, each of which has a chance of $p^r(1-p)^{n-r}$, since there are r heads and $n - r$ tails.

[i] Strictly speaking, we would need to verify that the function here indeed defines a probability, that is, it satisfies the three axioms in Chapter 2.

Remark (Bernoulli trials[i]). The coin toss above is an example of Bernoulli trials: Repeated independent trials such that there are only two possible outcomes for each trial and their probability remain the same throughout the trials. Such examples, in addition to the coin toss, include success or failure of a medicine trial, boy or girl of a newborn, etc.

Homework Assignment.

Question 3.1. The probability that a new car battery functions for more than 10,000 miles is 0.8, the probability that it functions for more than 20,000 miles is 0.4, and the probability that it functions for more than 30,000 miles is 0.1. If a new car battery is still working after 10,000 miles, what is the probability that

(a). its total life will exceed 20,000 miles?
(b). its additional life will exceed 20,000 miles?

Answer. Let E denote the event that a new car battery functions for more than 10,000 miles, F denote the event that a new car battery functions for more than 20,000 miles, and G denote the event that a new car battery functions for more than 30,000 miles. Then $P(E) = 0.8$, $P(F) = 0.4$, and $P(G) = 0.1$.

(a). The required is the conditional probability

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{P(F)}{P(E)} = \frac{0.4}{0.8} = \frac{1}{2},$$

in which we used the fact that $F \subset E$.

(b). The required is the conditional probability

$$P(G|E) = \frac{P(G \cap E)}{P(E)} = \frac{P(G)}{P(E)} = \frac{0.1}{0.8} = \frac{1}{8},$$

in which we used the fact that $G \subset E$.

Question 3.2. As a simplified model for weather forecasting, suppose that the weather (either wet or dry) tomorrow will be the same as the weather today with probability p. Suppose that is dry on January 1.

(a). Show that P_n, the probability that it will be dry n days later, satisfies $p_0 = 1$ and

$$P_n = (2p-1)P_{n-1} + (1-p) \quad \text{for } n \geq 1.$$

(b). Using Part (a) and by mathematical induction, show that

$$P_n = \frac{1}{2} + \frac{1}{2}(2p-1)^n \quad \text{for } n \geq 0.$$

Answer. (a). On the condition that $n-1$ days later is dry (with probability P_{n-1}), the conditional probability that the n days later is dry is p; on the condition that $n-1$ days later is wet (with probability $1 - P_{n-1}$), the conditional probability that the n days later is dry is $1-p$. Hence, by Bayes' formula,

$$P_n = p \cdot P_{n-1} + (1-p) \cdot (1 - P_{n-1}) = (2p-1)P_{n-1} + (1-p).$$

(b). To prove the formula for P_n by mathematical induction, first notice that when $n = 0$,

$$P_0 = \frac{1}{2} + \frac{1}{2}(2p-1)^0 = \frac{1}{2} + \frac{1}{2} = 1.$$

Assume the formula for $n-1$. Then

$$\begin{aligned} P_n &= (2p-1)P_{n-1} + (1-p) \\ &= (2p-1)\left(\frac{1}{2} + \frac{1}{2}(2p-1)^{n-1}\right) + (1-p) \end{aligned}$$

[i]Jacob Bernoulli, *Ars Conjectandi*. (1713).

$$\begin{aligned} &= \frac{1}{2}(2p-1) + \frac{1}{2}(2p-1)^n + (1-p) \\ &= \frac{1}{2} + \frac{1}{2}(2p-1)^n . \end{aligned}$$

Question 3.3. Consider n independent tosses of a coin, which lands on heads with probability p. How large need n be so that the probability of obtaining at least one head is at least $\frac{1}{2}$?

Answer. The probability that there is no heads in n independent tosses is $(1-p)^n$. Therefore, the probability of obtaining at least one head is $1-(1-p)^n$. It is at least 0.5, if

$$1-(1-p)^n \geq \frac{1}{2},$$

which solves to

$$n \geq -\frac{\log 2}{\log(1-p)}.$$

Question 3.4. Independent trials that result in a success with probability p are successively performed for n times. Let $0 \leq r \leq n$. Find the probability that exactly r successes are obtained.

Answer. There are r successes and $n-k$ failures. Therefore, the required probability is

$$\binom{n}{r} p^r (1-p)^{n-r}.$$

Question 3.5. Independent trials that result in a success with probability p are successively performed until a total of r successes is obtained.

(a). Find the probability that exactly n trails are required.
(b). Find the probability that r successes occur before m failures.

Answer.

(a). In order for it to take n trails to obtain r successes, the n-th trial must be a success, and there must be $r-1$ successes and $(n-1)-(r-1) = n-r$ failures in the first $n-1$ trials. Therefore, the probability is

$$p \cdot \binom{n-1}{r-1} p^{r-1}(1-p)^{n-r} = \binom{n-1}{r-1} p^r (1-p)^{n-r}.$$

(b). To have r successes before m failures, the minimal number of trials to run is r (i.e., there are no failures), while the maximal number of trials is $m+r-1$ (i.e., there are r successes and $m-1$ failures). Therefore, the required probability is

$$\sum_{n=r}^{m+r-1} \binom{n-1}{r-1} p^r (1-p)^{n-r}.$$

CHAPTER 4

Discrete random variables

La vie n'est bonne qu'à deux choses: découvrir les mathématiques et enseigner les mathématiques. [Life is good for only two things: discovering mathematics and teaching mathematics.]

Siméon Denis Poisson (1781–1840)

In this chapter, we define the random variables, the basic object in the probability theory. Through the discrete random variables we introduce the most fundamental concepts in the probability theory: expectation (i.e., mean), median, variance, standard deviation, and moments. These quantities provide the information on the statistical analysis of a random variable, in terms of the distribution of the values.

Three important random variables are discussed: binomial, Poisson, and geometric. These random models have a great variety of applications to a diverse fields. Provided certain parameters, the expectation, median, variance, standard deviation, moments, and their value distribution can be completely determined, which then supply critical information in the applications.

Definition (Random variables). A random variable is a function on a probability space.

Let S be a discrete probability space with a probability P throughout this chapter. Then any random variable on S can only take a discrete set of possible values, that is, it is a discrete random variable.

4.1. Examples

Example. Suppose that the class of *Introduction to Probability* is consisted of 10 students. Equip this sample space S with a probability P such that each sample point has the same probability (of $\frac{1}{10}$). Let X denote the score a student obtained in Test 1. Then X is a random variable taking one of the values in $\{0, ..., 40\}$. (The full score is 40.) Assume that the scores of Test 1 are $33, 17, 32, 38, 32, 23, 32, 38, 40, 23$. Then

$$P(X = 17) = \frac{1}{10} = 0.1,$$
$$P(X = 23) = \frac{2}{10} = 0.2,$$
$$P(X = 32) = \frac{3}{10} = 0.3,$$
$$P(X = 33) = \frac{1}{10} = 0.1,$$
$$P(X = 38) = \frac{2}{10} = 0.2,$$
$$P(X = 40) = \frac{1}{10} = 0.1,$$

and $P(X = i) = 0$ for $i \neq 17, 23, 32, 33, 38, 40$. Notice that

$$\sum_{i=0}^{40} P(X = i) = 1,$$

in which $X = i$, $i \in \mathbb{N}$, provide a disjoint partition of the sample space S.

Remark. Recall that for a function $X : S \to \mathbb{R}$, $X = i$ denotes the inverse image of $\{x\}$, i.e., $X^{-1}(\{i\}) = \{s \in S : X(s) = i\}$, see Chapter 1.

Definition (Probability mass function). Let X be a random variable on S. The probability mass function of X is defined as

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

Example. Consider three independent tosses of a fair coin. Then the sample space

$$S = \{(s_1, s_2, s_3) : s_1, s_2, s_3 \in \{\mathrm{h}, \mathrm{t}\}\}$$

has size eight. Let X denote the number of heads that appear. Then X is a random variable taking one of the values $0, 1, 2, 3$ with probabilities

$$\begin{aligned}
p(0) = P(X = 0) &= P\left(\{(\mathrm{t}, \mathrm{t}, \mathrm{t})\}\right) = \frac{1}{8}, \\
p(1) = P(X = 1) &= P\left(\{(\mathrm{h}, \mathrm{t}, \mathrm{t}), (\mathrm{t}, \mathrm{h}, \mathrm{t}), (\mathrm{t}, \mathrm{t}, \mathrm{h})\}\right) = \frac{3}{8}, \\
p(2) = P(X = 2) &= P\left(\{(\mathrm{t}, \mathrm{h}, \mathrm{h}), (\mathrm{h}, \mathrm{t}, \mathrm{h}), (\mathrm{h}, \mathrm{h}, \mathrm{t})\}\right) = \frac{3}{8}, \\
p(3) = P(X = 3) &= P\left(\{(\mathrm{h}, \mathrm{h}, \mathrm{h})\}\right) = \frac{1}{8}.
\end{aligned}$$

Notice that

$$p(0) + p(1) + p(2) + p(3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1,$$

in which $X = i$, $i = 0, 1, 2, 3$, provide a disjoint partition of the sample space S.

Example (Binomial random variable). Consider n independent tosses of a coin, which lands on heads with probability p. Then the sample space

$$S = \{(s_1, ..., s_n) : s_1, ..., s_n \in \{\mathrm{h}, \mathrm{t}\}\}.$$

Let X denote the number of heads that appear. Then X is a random variable taking one of the values $0, 1, ..., n$ with probabilities

$$p(i) = P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, ..., n.$$

Notice that

$$\sum_{i=0}^{n} p(i) = \sum_{i=0}^{n} P(X = i) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = (p + (1-p))^n = 1,$$

in which $X = i$, $i = 0, 1, ..., n$, provide a disjoint partition of the sample space S. Here, we used the binomial formula in Theorem 1.4 that

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

in which we set $x = p$ and $y = 1 - p$ in the formula.

Example (Poisson random variable[i]). Set $\lambda > 0$. We say that X is a Poisson random variable with parameter λ, if

$$p(i) = P(X = i) = \frac{\lambda^i e^{-\lambda}}{i!} \quad \text{for all } i = 0, 1, ...$$

[i] Siméon Denis Poisson, *Recherches sur la probabilité des jugements en matière criminelle et en matière civile.* [Research on the probability of judgments in criminal and civil matters]. (1837).

Notice that

$$\sum_{i=0}^{\infty} p(i) = \sum_{i=0}^{\infty} P(X=i) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Here, we used the Taylor expansion of e^{λ} that

$$e^{\lambda} = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}.$$

Example (Geometric random variable). Consider independent tosses of a coin, which lands on heads with probability p. Let X denote the number of times the coin is tossed until the a head occurs. Then X is a random variable on the sample space

$$S = \{(s_1), (s_1, s_2), (s_1, s_2, s_3), ... : s_1, s_2, ... \in \{\mathrm{h}, \mathrm{t}\}\}.$$

For $i = 1, ..., n$, X takes value i with probabilities

$$p(i) = P(X=i) = (1-p)^{i-1} p.$$

Notice that

$$\sum_{i=1}^{\infty} p(i) = \sum_{i=1}^{\infty} P(X=i) = p \sum_{i=1}^{\infty} (1-p)^{i-1} = p \cdot \frac{1}{1-(1-p)} = 1,$$

in which $X = i$, $i \in \mathbb{N}$, provide a disjoint partition of the sample space S. Here, we used the summation formula for a geometric series that

$$\sum_{i=1}^{\infty} r^i = \frac{1}{1-r},$$

in which we set $r = 1 - p$.

4.2. Expectation

Definition (Expectation). Let X be a random variable on S. Then the expectation (or expected value, mean) of X is defined as

$$\mathbb{E}[X] = \sum_{s \in S} X(s) \cdot P(\{s\}) = \sum_{x \in X(S)} x \cdot p(x).$$

Remark. The first summation above is with respect to all sample points $\{s\} \subset S$, while the second one all possible values of the random variable $x \in X(S)$. Notice that $\{s\}$, $s \in S$, provide a disjoint partition of S, and $X = x$, $x \in X(S)$, also provides a disjoint partition of S. Hence,

$$\sum_{s \in S} P(\{s\}) = \sum_{x \in X(S)} p(x) = P(S) = 1.$$

Moreover, one only needs to include the terms in the summations which have non-zero probabilities:

$$\mathbb{E}[X] = \sum_{P(\{s\})>0} X(s) \cdot P(\{s\}) = \sum_{p(x)>0} y \cdot p(x).$$

Proposition 4.1. *Let X be a random variable on a finite probability space S. Then*

$$\min_{s \in S}\{X(s)\} \leq \mathbb{E}[X] \leq \max_{s \in S}\{X(s)\}.$$

Proof. Write $M = \max_{s \in S}\{X(s)\}$. Then

$$\mathbb{E}[X] = \sum_{x \in X(S)} x \cdot p(x) \leq \sum_{x \in X(S)} M \cdot p(x) = M \sum_{x \in X(S)} x \cdot p(x) = M.$$

Similarly, write $m = \min_{s \in S}\{X(s)\}$. Then

$$\mathbb{E}[X] = \sum_{x \in X(S)} x \cdot p(x) \geq \sum_{x \in X(S)} m \cdot p(x) = m \sum_{x \in X(S)} x \cdot p(x) = m.$$

□

Remark. We point out that $\mathbb{E}[X] = \min$ iff $X = \min$ almost surely, i.e., $X = \min$ except possibly on a subset with probability 0. Similarly, $\mathbb{E}[X] = \max$ iff $X = \max$ almost surely.

Definition (Median). Let X be a random variable on a probability space S. We say that $m \in \mathbb{R}$ is a median of X, if

$$P(X \geq m) \geq \frac{1}{2} \quad \text{and} \quad P(X \leq m) \geq \frac{1}{2}.$$

Notice that a median of a random variable is also bounded between the maximal and minimal values. It equals the maximal or minimal value iff the random variable is almost surely constant. However, it may not be unique. See below.

Example. Suppose that the class of *Introduction to Probability* is consisted of 10 students. Equip this sample space S with a probability P such that each sample point has the same probability (of $\frac{1}{10}$). Assume that the scores of Test 1 are $33, 17, 32, 38, 32, 23, 32, 38, 40, 23$. Let X denote the score a student obtained in Test 1. Then X is a random variable whose expectation

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{s \in S} X(s) \cdot P(\{s\}) \\
&= 33 \cdot \frac{1}{10} + 17 \cdot \frac{1}{10} + 32 \cdot \frac{1}{10} + 38 \cdot \frac{1}{10} + 32 \cdot \frac{1}{10} \\
&\quad +23 \cdot \frac{1}{10} + 32 \cdot \frac{1}{10} + 38 \cdot \frac{1}{10} + 40 \cdot \frac{1}{10} + 23 \cdot \frac{1}{10} \\
&= \frac{33 + 17 + 32 + 38 + 32 + 23 + 32 + 38 + 40 + 23}{10} \\
&= 30.8.
\end{aligned}$$

On the other hand, X takes values in $17, 23, 32, 33, 38, 40$. So

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in X(S)} x \cdot p(x) \\
&= 17 \cdot \frac{1}{10} + 23 \cdot \frac{2}{10} + 32 \cdot \frac{3}{10} + 33 \cdot \frac{1}{10} + 38 \cdot \frac{2}{10} + 40 \cdot \frac{1}{10} \\
&= 30.8.
\end{aligned}$$

In this model, the expectation is simply the mean. To find a median, we arrange the values of X in the increasing order: $17, 23, 23, 32, 32, 32, 33, 38, 38, 40$. Then the median is 32 (and is unique).

Example (Constant random variable). Suppose that the class of *Introduction to Probability* is consisted of 10 students. Equip this sample space S with a probability P such that each sample point has the same probability (of $\frac{1}{10}$). Assume each student attempted all eight questions in Test 1. Let X denote the score a student obtained in Test 1. Then X is a random variable whose expectation $\mathbb{E}[X] = 8$, and so is the median. They equal the maximal value 8 of X and $X(s) = 8$ for all $s \in S$. This is an example of a constant random variable that it takes only one value.

Example. Consider three independent tosses of a fair coin. Then the sample space

$$S = \{(s_1, s_2, s_3) : s_1, s_2, s_3 \in \{\mathrm{h}, \mathrm{t}\}\}$$

has size eight. Let X denote the number of heads that appear. Then X is a random variable whose expectation

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{s \in S} X(s) \cdot P(\{s\}) \\
&= 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8}
\end{aligned}$$

$$\begin{aligned} &= \frac{0+1+1+1+2+2+2+3}{8} \\ &= \frac{3}{2}. \end{aligned}$$

On the other hand, X takes values in $0, 1, 2, 3$. So

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in X(S)} x \cdot p(x) \\ &= 0 \cdot P(X=0) + 1 \cdot P(X=1) + 2 \cdot P(X=2) + 3 \cdot P(X=3) \\ &= 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} \\ &= \frac{3}{2}. \end{aligned}$$

To find a median, we arrange the values of X in the increasing order: $0, 1, 1, 1, 2, 2, 2, 3$. Then any number $m \in [1, 2]$ is a median (and is not unique), e.g., $1, 1.4, 1.5, 1.6, 2$.

Example. Casino offers a game of three independent tosses of a fair coin. It takes one dollar[i] to play a round, after which the player gets a reward of six dollars if three heads appear and nothing otherwise. Let X denote the number of dollars the player gets. Then X is a random variable taking one of values $0, 6$ with probabilities $p(0) = \frac{7}{8}, p(6) = \frac{1}{8}$. The expectation

$$\mathbb{E}[X] = 0 \cdot p(0) + 6 \cdot p(6) = \frac{6}{8} = 0.75.$$

Therefore, if the game is played for a great many rounds, say n rounds, then the casino makes a profit which tends to $n - 0.75n = 0.25n$ dollars.

Example. Consider n independent tosses of a fair coin. Let X denote the number of heads that appear. Then X is a random variable whose expectation

$$\begin{aligned} \mathbb{E}[X] &= \sum_{i=0}^{n} i \cdot p(i) \\ &= \sum_{i=0}^{n} i \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} \\ &= \left(\frac{1}{2}\right)^n \sum_{i=1}^{n} \frac{n!}{(i-1)!(n-i)!} \\ &= n \left(\frac{1}{2}\right)^n \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-1-j)!} \\ &= n \left(\frac{1}{2}\right)^n \sum_{j=0}^{n-1} \binom{n-1}{j} \\ &= n \left(\frac{1}{2}\right)^n \cdot 2^{n-1} \\ &= \frac{n}{2}. \end{aligned}$$

PROPOSITION 4.2. *Let X be a random variable on S. Suppose that $a, b \in \mathbb{R}$. Then $aX + b$ is also a random variable on S, whose expectation*

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

[i]We can interpret the number of dollar the player pays as a constant random variable which take only one value of 1.

PROOF. Compute that

$$\begin{aligned}\mathbb{E}[aX+b] &= \sum_{s\in S}(aX+b)(x)\cdot P(\{s\}) \\ &= a\sum_{s\in S}X(s)\cdot P(\{s\}) + b\sum_{s\in S}P(\{s\}) \\ &= a\mathbb{E}[X]+b.\end{aligned}$$

□

Example. Suppose that the class of *Introduction to Probability* is consisted of 10 students. Equip this sample space S with a probability P such that each sample point has the same probability (of $\frac{1}{10}$). Assume that the scores of Test 1 are $33, 17, 32, 38, 32, 23, 32, 38, 40, 23$. Let X denote the score a student obtained in Test 1. Then X is a random variable with expectation $\mathbb{E}[X] = 30.8$.

(a). $0.9X + 4$ is a random variable whose expectation

$$\begin{aligned}\mathbb{E}[0.9X+4] &= (0.9\cdot 33+4)\cdot\frac{1}{10} + (0.9\cdot 17+4)\cdot\frac{1}{10} + (0.9\cdot 32+4)\cdot\frac{1}{10} \\ &\quad +(0.9\cdot 38+4)\cdot\frac{1}{10} + (0.9\cdot 32+4)\cdot\frac{1}{10} + (0.9\cdot 23+4)\cdot\frac{1}{10} \\ &\quad +(0.9\cdot 32+4)\cdot\frac{1}{10} + (0.9\cdot 38+4)\cdot\frac{1}{10} + (0.9\cdot 40+4)\cdot\frac{1}{10} \\ &\quad +(0.9\cdot 23+4)\cdot\frac{1}{10} \\ &= \frac{33.7+19.3+32.8+38.2+32.8+24.7+32.8+38.2+40+24.7}{10} \\ &= 31.72,\end{aligned}$$

which coincides with

$$0.9\mathbb{E}[X]+4 = 0.9\cdot 30.8+4 = 31.72.$$

(b). $\sqrt{40X}$ is a random variable whose expectation

$$\begin{aligned}&\mathbb{E}\left[\sqrt{40X}\right] \\ =\ & \sqrt{40\cdot 33}\cdot\frac{1}{10} + \sqrt{40\cdot 17}\cdot\frac{1}{10} + \sqrt{40\cdot 32}\cdot\frac{1}{10} + \sqrt{40\cdot 38}\cdot\frac{1}{10} + \sqrt{40\cdot 32}\cdot\frac{1}{10} \\ & +\sqrt{40\cdot 23}\cdot\frac{1}{10} + \sqrt{40\cdot 32}\cdot\frac{1}{10} + \sqrt{40\cdot 38}\cdot\frac{1}{10} + \sqrt{40\cdot 40}\cdot\frac{1}{10} + \sqrt{40\cdot 23}\cdot\frac{1}{10} \\ =\ & \frac{36.3+26.1+35.8+39.0+35.8+30.3+35.8+39.0+40+30.3}{10} \\ =\ & 34.8.\end{aligned}$$

But

$$\sqrt{40\cdot\mathbb{E}[X]} = \sqrt{40\cdot 30.8} \approx 35.1.$$

Example. Consider three independent tosses of a fair coin. Then the sample space

$$S = \{(s_1, s_2, s_3) : s_1, s_2, s_3 \in \{\mathrm{h}, \mathrm{t}\}\}$$

has size eight. Let X denote the number of heads that appear. Then X is a random variable with expectation $\frac{3}{2}$. Now X^2 is a random variable whose expectation

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= 0^2\cdot p(0) + 1^2\cdot p(1) + 2^2\cdot p(2) + 3^2\cdot p(3) \\ &= 0\cdot\frac{1}{8} + 1\cdot\frac{3}{8} + 4\cdot\frac{3}{8} + 9\cdot\frac{1}{8} \\ &= 3.\end{aligned}$$

But

$$\mathbb{E}[X]^2 = \left(\frac{3}{2}\right)^2 = \frac{9}{4}.$$

Remark. The proposition and examples above indicate that taking the expectation $\mathbb{E}$ commutes with linear operators $X \to aX + b$. However, it does not commute with nonlinear operators such as $X \to \sqrt{X}$ and $X \to X^2$.

4.3. Variance

Definition (Variance). Let X be a random variable. Write $\mu = \mathbb{E}[X]$. The variance of X is defined as

$$\mathrm{Var}[X] = \mathbb{E}\left[(X-\mu)^2\right],$$

which is usually denoted by σ^2. The standard derivation is defined as $\sigma = \sqrt{\mathrm{Var}[X]}$.

Remark. The variance is the expectation of the square of the distance between values of a random variable and its expectation. Therefore, the variance, as well as the standard deviation, measure how much "on average" the values deviates from the expectation. In particular, $\mathrm{Var}[X] \geq 0$ and $\mathrm{Var}[X] = 0$ iff $X = \mu$ almost surely.

Notice that taking expectation does not commute with taking squares:

$$\mathbb{E}\left[(X-\mu)^2\right] \neq \mathbb{E}[X-\mu]^2.$$

Indeed, compute that

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}\left[(X-\mu)^2\right] \\
&= \mathbb{E}\left[X^2 - 2\mu X + \mu^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mu\mathbb{E}[X] + \mathbb{E}\left[\mu^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mu^2 + \mu^2 \\
&= \mathbb{E}\left[X^2\right] - \mu^2.
\end{aligned}$$

Definition (Moments). Let X be a random variable. For $n \in \mathbb{N}$, the n-th moment of X is defined as $\mathbb{E}[X^n]$. In particular, the first moment $\mathbb{E}[X] = \mu$ is the expectation and the second moment $\mathbb{E}[X^2]$ is related to the variance $\mathrm{Var}[X] = \sigma^2$ via

$$\mathbb{E}\left[X^2\right] = \sigma^2 + \mu^2.$$

Remark. Similar to the variance, the n-th moment of a random variable measures how much "on average" its values deviate from the origin 0, weighted according to the n-th power of the distance with 0. These moments provide quantitative estimate of the deviation, in particular, higher the moments, more weighted toward the values which are further deviated from 0.

Proposition 4.3. *Let X be a random variable. Suppose that $a, b \in \mathbb{R}$. Then $aX + b$ is also a random variable, whose variance is given by*

$$\mathrm{Var}[aX+b] = a^2\mathrm{Var}[X].$$

Proof. We know that $\mathbb{E}[aX+b] = a\mathbb{E}[X] + b$. Compute that

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

□

Example. Suppose that the class of *Introduction to Probability* is consisted of 10 students. Equip this sample space S with a probability P such that each sample point has the same probability (of $\frac{1}{10}$). Assume that the scores of Test 1 are $33, 17, 32, 38, 32, 23, 32, 38, 40, 23$. Let X denote the score a student obtained in Test 1. Then X is a random variable whose expectation $\mu = \mathbb{E}[X] = 30.8$. On one hand,

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}\left[(X-\mu)^2\right] \\ &= (17-30.8)^2 \cdot \frac{1}{10} + (23-30.8)^2 \cdot \frac{2}{10} + (32-30.8)^2 \cdot \frac{3}{10} \\ &\quad +(33-30.8)^2 \cdot \frac{1}{10} + (38-30.8)^2 \cdot \frac{2}{10} + (40-30.8)^2 \cdot \frac{1}{10} \\ &= 50.96.\end{aligned}$$

On other other hand,

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= 17^2 \cdot \frac{1}{10} + 23^2 \cdot \frac{2}{10} + 32^2 \cdot \frac{3}{10} + 33^2 \cdot \frac{1}{10} + 38^2 \cdot \frac{2}{10} + 40^2 \cdot \frac{1}{10} \\ &= 999.6.\end{aligned}$$

So the variance of X

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mu^2 = 999.6 - 30.8^2 = 50.96.$$

The standard deviation is $\sqrt{\mathrm{Var}[X]} \approx 7.139$. Moreover, the probability that the random variable X is in the range $\mathbb{E}[X] \pm \sqrt{\mathrm{Var}[X]}$, 30.8 ± 7.13, is 0.8.

Example. Consider three independent tosses of a fair coin. Let X denote the number of heads that appear. Then X is a random variable with expectation $\mu = \mathbb{E}[X] = \frac{3}{2}$.

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}\left[(X-\mu)^2\right] \\ &= \left(0-\frac{3}{2}\right)^2 \cdot \frac{1}{8} + \left(1-\frac{3}{2}\right)^2 \cdot \frac{3}{8} + \left(2-\frac{3}{2}\right)^2 \cdot \frac{3}{8} + \left(3-\frac{3}{2}\right) \cdot \frac{1}{8} \\ &= \frac{3}{4}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= 0^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{3}{8} + 2^2 \cdot \frac{3}{8} + 3^2 \cdot \frac{1}{8} \\ &= 3.\end{aligned}$$

So the variance of X

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mu^2 = 3 - \left(\frac{3}{2}\right)^2 = \frac{3}{4}.$$

The standard deviation is $\sqrt{\mathrm{Var}[X]} \approx 0.866$. Moreover, the probability that the random variable X is in the range $\mathbb{E}[X] \pm \sqrt{\mathrm{Var}[X]}$, 1.5 ± 0.866, is $\frac{6}{8} = 0.75$.

Example. Consider n independent tosses of a fair coin. Let X denote the number of heads that appear. Then X is a random variable with expectation $\mu = \mathbb{E}[X] = \frac{n}{2}$. While

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \sum_{i=0}^{n} i^2 \cdot p(i) \\ &= \sum_{i=0}^{n} i^2 \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1-\frac{1}{2}\right)^{n-i} \\ &= \sum_{i=1}^{n} i \cdot \frac{n!}{(i-1)!(n-i)!} \left(\frac{1}{2}\right)^i \left(1-\frac{1}{2}\right)^{n-i}\end{aligned}$$

$$\begin{aligned}
&= n\sum_{j=0}^{n-1}(j+1)\cdot\frac{(n-1)!}{j!(n-1-j)!}\left(\frac{1}{2}\right)^{j+1}\left(1-\frac{1}{2}\right)^{n-1-j}\\
&= \frac{n}{2}\sum_{j=0}^{n-1}(j+1)\cdot\binom{n-1}{j}\left(\frac{1}{2}\right)^{j}\left(1-\frac{1}{2}\right)^{n-1-j}\\
&= \frac{n}{2}\sum_{j=0}^{n-1}(j+1)\cdot P_{n-1}(X=j)\\
&= \frac{n}{2}\mathbb{E}_{n-1}[X+1]\\
&= \frac{n}{2}\left(\mathbb{E}_{n-1}[X]+1\right)\\
&= \frac{n}{2}\left(\frac{n-1}{2}+1\right)\\
&= \frac{n(n+1)}{4}.
\end{aligned}$$

Hence, the variance of X

$$\mathrm{Var}[X]=\mathbb{E}\left[X^2\right]-\mu^2=\frac{n(n+1)}{4}-\left(\frac{n}{2}\right)^2=\frac{n}{4}.$$

4.4. Three random variables: binomial, Poisson, and geometric

In this section, we derive the expectations and variances of three important random variables:

TABLE 4.1. Table of expectations and variances of three discrete random variables

Random variable	Probability mass function	Expectation	Variance
Binomial with parameter (n,p)	$p(i)=\binom{n}{i}p^i(1-p)^{n-i}$ for $i=0,...,n$	np	$np(1-p)$
Poisson with parameter λ	$p(i)=\frac{\lambda^i e^{-\lambda}}{i!}$ for $i=0,1,...$	λ	λ
Geometric with parameter p	$p(i)=p(1-p)^{i-1}$ for $i=1,2,...$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

In general, the variance of a random variable is more difficult to calculate than the expectation. – It often requires an inductive argument on the parameters or on the moments. But the argument is not unique. Indeed, in Chapter 6, we use a different (and much more simpler) argument involving multiple (independent) random variables to derive expectation and variance of the binomial random variables, see Propositions 6.7 and 6.13.

To simplify the notations, we denote

$$q=1-p.$$

4.4.1. Binomial random variable. Consider n independent tosses of a coin, which lands on heads with probability p. Let X denote the number of heads that appear. Then X is a binomial random variable that

$$p(i)=\binom{n}{i}p^i(1-p)^{n-i}=\binom{n}{i}p^iq^{n-i}\quad\text{for } i=0,...,n.$$

Compute the expectation of X that

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=0}^{n} i\cdot p(i)\\
&= \sum_{i=0}^{n} i\binom{n}{i}p^iq^{n-i}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^{n} \frac{n!}{(i-1)!(n-i)!} p^i q^{n-i} \\
&= np \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!(n-i)!} p^{i-1} q^{n-i} \\
&= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-1-j} \\
&= np(p+q)^{n-1} \\
&= np.
\end{aligned}$$

Compute the second moment of X that

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \sum_{i=0}^{n} i^2 \cdot p(i) \\
&= \sum_{i=0}^{n} i^2 \binom{n}{i} p^i q^{n-i} \\
&= \sum_{i=1}^{n} i \cdot \frac{n!}{(i-1)!(n-i)!} \cdot p^i q^{n-i} \\
&= n \sum_{j=0}^{n-1} (j+1) \cdot \frac{(n-1)!}{j!(n-1-j)!} p^{j+1} q^{n-1-j} \\
&= np \sum_{j=0}^{n-1} (j+1) \cdot \binom{n-1}{j} p^j q^{n-1-j} \\
&= np \sum_{j=0}^{n-1} (j+1) \cdot P_{n-1}(X=j) \\
&= np \mathbb{E}_{n-1}[X+1] \\
&= np \left(\mathbb{E}_{n-1}[X] + 1\right) \\
&= np \left((n-1)p + 1\right) \\
&= np(np - p + 1).
\end{aligned}$$

Here, P_{n-1} and $\mathbb{E}_{n-1}$ refer to the probability and the expectation, respectively, of the binomial random variable with parameter $(n-1, p)$.

Hence, the variance of X is given by

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = np(np - p + 1) - (np)^2 = np(1-p).$$

4.4.2. Poisson random variable. Set $\lambda > 0$. Consider the Poisson random variable X with parameter λ, that is,

$$p(i) = \frac{\lambda^i e^{-\lambda}}{i!} \quad \text{for } i = 0, 1, ...$$

Compute the expectation of X that

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=0}^{\infty} i \cdot p(i) \\
&= \sum_{i=1}^{\infty} i \cdot \frac{\lambda^i e^{-\lambda}}{i!}
\end{aligned}$$

$$\begin{aligned} &= e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^i}{(i-1)!} \\ &= e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^{j+1}}{j!} \\ &= \lambda e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} \\ &= \lambda e^{-\lambda}\cdot e^{\lambda} \\ &= \lambda. \end{aligned}$$

Compute the second moment of X that

$$\begin{aligned} \mathbb{E}\left[X^2\right] &= \sum_{i=0}^{\infty} i^2\cdot p(i) \\ &= \sum_{i=1}^{\infty} i^2\cdot\frac{\lambda^i e^{-\lambda}}{i!} \\ &= e^{-\lambda}\sum_{i=1}^{\infty}\frac{i\lambda^i}{(i-1)!} \\ &= e^{-\lambda}\sum_{j=0}^{\infty}\frac{(j+1)\lambda^{j+1}}{j!} \\ &= \lambda\left(\sum_{j=0}^{\infty} j\cdot\frac{\lambda^j e^{-\lambda}}{j!}+\sum_{j=0}^{\infty}\frac{\lambda^j e^{-\lambda}}{j!}\right) \\ &= \lambda\left(\mathbb{E}[X]+1\right) \\ &= \lambda(\lambda+1). \end{aligned}$$

Hence, the variance of X is given by

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right]-\mathbb{E}[X]^2 = \lambda(\lambda+1)-\lambda^2 = \lambda.$$

The Poisson random variable has a tremendous range of applications in diverse areas, because it can be used as an approximation for a binomial random variable X with parameter (n,p) when is n large and p is small enough so that $\lambda = np$ is of moderate size. To see this,

$$P(X=0) = \binom{n}{0}p^n = \left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda} = P(X=0).$$

Since $p\approx 0$ and $q = 1-p\approx 1$,

$$\frac{P(X=i)}{P(X=i-1)} = \frac{\binom{n}{i}p^i q^{n-i}}{\binom{n}{i-1}p^{i-1}q^{n-i+1}} = \frac{np-(i-1)p}{iq}\approx\frac{\lambda}{i}.$$

Hence,

$$P(X=i)\approx P(X=0)\cdot\frac{\lambda}{1}\cdot\frac{\lambda}{2}\cdots\frac{\lambda}{i} = \frac{\lambda^i e^{-\lambda}}{i!}.$$

We mention some examples of random variable that generally obey the Poisson probability law.

(a). The number of persons with the same birthday in a large group.
(b). The number of misprints on a page of a book.
(c). The number of people in a community who survive to an advanced age, say, 100.
(d). The number of customers entering a post office on a day.
(e). The number of α-particles discharged in a period of time from some radioactive material.

We discuss the example of birthdays in more details.

Example. Suppose that the College of Mathematics and Computer Science is consisted of 500 students. Assume that there are 365 possible birthdays. We find the probability that exactly k students have birthdays on February 26, using the binomial random variable with parameter $(500, \frac{1}{365})$. Then we find the approximation by the Poisson random variable with parameter $\frac{500}{365}$.

If the 500 people are chosen at random, we may apply the scheme of 500 Bernoulli trials with probability of success $p = \frac{1}{365} \approx 0.00274$ and probability of failure $q = 1 - p = \frac{364}{365} \approx 0.997$, i.e., whether a person has birthday on February 26 or not.

On one hand, using the binomial random variable with parameters $n = 500$ and p, we derive the required probability that exactly k have birthdays on February 26:

$$\binom{n}{k} p^k q^{n-k}.$$

On the other hand, using the Poisson approximation with parameter $\lambda = np \approx 1.370$, we derive the approximation:

$$\frac{\lambda^k e^{-\lambda}}{k!}.$$

The following table describes the probability and its Poisson approximation that exactly k have birthdays on February 26 for small k.

k	0	1	2	3	4	5	6
Binomial	0.2537	0.3484	0.2388	0.1089	0.0372	0.0101	0.0023
Poisson	0.2541	0.3481	0.2385	0.1089	0.0373	0.0102	0.0023

FIGURE 4.1. The binomial random variable with parameter $(500, \frac{1}{365})$ and its Poisson approximation with parameter $\lambda = \frac{500}{365}$. Notice that the probability achieves the maximal value at $k = 1$, which is the closest to the expectation $\lambda = np \approx 1.370$.

4.4.3. Geometric random variable*. Consider independent tosses of a coin, which lands on heads with probability p. Let X denote the number of times the coin is tossed until the a head occurs. Then X is a geometric random variable that

$$p(i) = p(1-p)^{i-1} = pq^{i-1} \quad \text{for } i = 1, 2, ...$$

Compute the expectation of X that

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{\infty} i \cdot p(i) \\
&= \sum_{i=1}^{\infty} i \cdot pq^{i-1} \\
&= \sum_{i=1}^{\infty} (i-1+1) \cdot pq^{i-1} \\
&= \sum_{i=1}^{\infty} (i-1) \cdot pq^{i-1} + \sum_{i=1}^{\infty} pq^{i-1} \\
&= \sum_{i=0}^{\infty} (i-1) \cdot pq^{i-1} + 1
\end{aligned}$$

$$\begin{aligned} &= \sum_{j=1}^{\infty} j \cdot pq^j + 1 \\ &= q\sum_{j=1}^{\infty} j \cdot pq^{j-1} + 1 \\ &= q\mathbb{E}[X] + 1, \end{aligned}$$

which implies that

$$\mathbb{E}[X] = \frac{1}{1-q} = \frac{1}{p}.$$

Compute the second moment of X that

$$\begin{aligned} \mathbb{E}\left[X^2\right] &= \sum_{i=1}^{\infty} i^2 \cdot p(i) \\ &= \sum_{i=1}^{\infty} (i-1+1)^2 \cdot pq^{i-1} \\ &= \sum_{i=1}^{\infty} (i-1)^2 \cdot pq^{i-1} + 2\sum_{i=1}^{\infty} (i-1) \cdot pq^{i-1} + \sum_{i=1}^{\infty} pq^{i-1} \\ &= \sum_{i=0}^{\infty} (i-1)^2 \cdot pq^{i-1} + 2\sum_{i=0}^{\infty} (i-1) \cdot pq^{i-1} + 1 \\ &= \sum_{j=1}^{\infty} j^2 \cdot pq^j + 2\sum_{j=1}^{\infty} j \cdot pq^j + 1 \\ &= q\sum_{j=1}^{\infty} j^2 \cdot pq^{j-1} + 2q\sum_{j=1}^{\infty} j \cdot pq^{j-1} + 1 \\ &= q\mathbb{E}\left[X^2\right] + 2q\mathbb{E}[X] + 1 \\ &= q\mathbb{E}\left[X^2\right] + \frac{2q}{p} + 1, \end{aligned}$$

which implies that

$$\mathbb{E}\left[X^2\right] = \frac{2q+p}{p^2}.$$

Hence, the variance of X is given by

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{2q+p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{2(1-p)+p-1}{p^2} = \frac{1-p}{p^2}.$$

Homework Assignment.

Question 4.1. Find the minimal number of people such that the probability that at least one of them has the same birthday as you is greater than 0.1. (Assume that there are 365 possible birthdays.)

Answer. Suppose that there are n people. Then the probability that all of them have birthdays different from me, i.e., in the remaining 364 days, is

$$\left(\frac{364}{365}\right)^n.$$

The probability that at least one of them has the same birthday as me is greater than 0.1, if

$$1 - \left(\frac{364}{365}\right)^n > \frac{1}{10},$$

which implies that

$$\left(\frac{364}{365}\right)^n < \frac{9}{10}.$$

Hence,

$$n > \frac{\log \frac{9}{10}}{\log \frac{364}{365}} \approx 38.4.$$

Therefore, the minimal number required is 39.

Question 4.2. Consider independent tosses of two coins, the first of which lands on heads with probability 0.9 and the second with probability 0.3. Let X be the total number of heads that appear.
(a). Find $P(X = 1)$.
(b). Find $\mathbb{E}[X]$.

Answer. The sample space is

$$S = \{(\mathrm{h}, \mathrm{h}), (\mathrm{h}, \mathrm{t}), (\mathrm{t}, \mathrm{h}), (\mathrm{h}, \mathrm{h})\},$$

in which

$$P(\{(\mathrm{h}, \mathrm{h})\}) = 0.27, \quad P(\{(\mathrm{h}, \mathrm{t})\}) = 0.63, \quad P(\{(\mathrm{t}, \mathrm{h})\}) = 0.03, \quad P(\{(\mathrm{t}, \mathrm{t})\}) = 0.07.$$

(a).

$$\begin{aligned} p(0) &= P(X = 0) = P(\{(\mathrm{t}, \mathrm{t})\}) = 0.07, \\ p(1) &= P(X = 1) = P(\{(\mathrm{h}, \mathrm{t}), (\mathrm{t}, \mathrm{h})\}) = 0.66, \\ p(2) &= P(X = 2) = P(\{(\mathrm{h}, \mathrm{h})\}) = 0.27. \end{aligned}$$

(b).

$$\mathbb{E}[X] = 0 \cdot p(0) + 1 \cdot p(1) + 2 \cdot p(2) = 0 \cdot 0.07 + 1 \cdot 0.66 + 2 \cdot 0.27 = 1.2.$$

Question 4.3. Casino offers a game of three independent tosses of a fair coin. It takes one dollar to play a round, after which the player gets a reward of six dollars if three heads appear, r dollars if two heads appear, and nothing otherwise. Find the largest reward r so that the casino would make a profit in a long run.

Answer. Let X denote the number of dollars the player gets. Then X is a random variable taking one of values $0, r, 6$ with probabilities $p(0) = \frac{4}{8}, p(r) = \frac{3}{8}, p(6) = \frac{1}{8}$. The expectation

$$\mathbb{E}[X] = 0 \cdot p(0) + r \cdot p(r) + 6 \cdot p(6) = \frac{3}{8}r + \frac{6}{8} < 1,$$

if $r < \frac{2}{3}$. Therefore, the largest reward required is $\frac{2}{3} \approx 0.667$ dollar.

Question 4.4. Let X be a random variable on a probability space. Suppose that $\mathbb{E}[X] = 1$ and $\mathrm{Var}[X] = 4$.
(a). Find $\mathbb{E}[(5 + X)^2]$.
(b). Find $\mathrm{Var}[2 + 3X]$.

Answer. The second moment

$$\mathbb{E}\left[X^2\right] = \mathrm{Var}[X] + \mathbb{E}[X]^2 = 5.$$

(a).

$$\mathbb{E}\left[(5 + X)^2\right] = \mathbb{E}\left[25 + 10X + X^2\right] = 25 + 10 \cdot \mathbb{E}[X] + \mathbb{E}\left[X^2\right] = 25 + 10 + 5 = 40.$$

(b). Firstly,

$$\mathbb{E}[2 + 3X] = 2 + 3 \cdot \mathbb{E}[X] = 5.$$

Then

$$\mathbb{E}\left[(2 + 3X)^2\right] = \mathbb{E}\left[4 + 12X + 9X^2\right] = 4 + 12 \cdot \mathbb{E}[X] + 9 \cdot \mathbb{E}\left[X^2\right] = 4 + 12 + 45 = 61.$$

Hence,

$$\mathrm{Var}[2+3X] = \mathbb{E}\left[(2+3X)^2\right] - \mathbb{E}[2+3X]^2 = 61 - 25 = 36.$$

Question 4.5. Let X be a random variable on a probability space with expectation μ and variance σ^2. Find the expectation and variance of $X = \frac{X-\mu}{\sigma}$.

Answer. Compute that

$$\mathbb{E}[X] = \mathbb{E}\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma}\mathbb{E}[X] - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0.$$

Since the second moment $\mathbb{E}[X^2] = \mathrm{Var}[X] + \mathbb{E}[X]^2 = \sigma^2 + \mu^2$,

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^2\right] \\
&= \frac{1}{\sigma^2}\mathbb{E}\left[X^2 - 2\mu X - \mu^2\right] \\
&= \frac{1}{\sigma^2}\left(\mathbb{E}\left[X^2\right] - 2\mu\mathbb{E}[X] - \mu^2\right) \\
&= \frac{1}{\sigma^2}\left(\sigma^2 + \mu^2 - 2\mu\mu - \mu^2\right) \\
&= 1.
\end{aligned}$$

Hence,

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = 1.$$

CHAPTER 5

Continuous random variables

There is a story about two friends, who were classmates in high school, talking about their jobs. One of them became a statistician and was working on population trends. He showed a reprint to his former classmate. The reprint started, as usual, with the Gaussian distribution and the statistician explained to his former classmate the meaning of the symbols for the actual population, for the average population, and so on. His classmate was a bit incredulous and was not quite sure whether the statistician was pulling his leg. "How can you know that?" was his query. "And what is this symbol here?" "Oh," said the statistician, "this is π." "What is that?" "The ratio of the circumference of the circle to its diameter" "Well, now you are pushing your joke too far," said the classmate, "surely the population has nothing to do with the circumference of the circle."

Eugene Wigner, 1960[i]

In this chapter, we introduce the random variables whose sets of possible values are uncountable, such as the length of calls received by a call center that is considered in Chapter 2. We study the expectation (i.e., mean), median, variance, standard deviation, moments, in the continuous setting, which draw clear comparison with their counterparts in the discrete setting of Chapter 4. Three important random variables are discussed: uniform, exponential, and normal (i.e., Gaussian).

Let S be a probability space with a probability P throughout this chapter.

Calculating the expectations and variances for continuous random variables require evaluation of integrals (of the probability density functions). We therefore recall the basic limit and integration formulas in Calculus.

5.1. Review of Calculus

5.1.1. Limits.

- For any $n \in \mathbb{N}$,
$$\lim_{x\to-\infty} \frac{1}{x^n} = 0 \quad \text{and} \quad \lim_{x\to\infty} \frac{1}{x^n} = 0.$$
-
$$\lim_{x\to-\infty} e^x = 0 \quad \text{and} \quad \lim_{x\to\infty} e^x = \infty.$$

5.1.2. Indefinite integrals.

-
$$\int x^n \, dx = \frac{x^{n+1}}{n+1} + c.$$
-
$$\int e^x \, dx = e^x + c \quad \text{and} \quad \int \frac{1}{x} \, dx = \ln|x| + c.$$
- Integration by parts:
$$\int u \, dv = uv - \int v \, du.$$

[i]Eugene Wigner, *The unreasonable effectiveness of mathematics in the natural sciences.* (1960).

5.1.3. Definite integrals. The definite integral of a function f on $[a, b]$ is given by

$$\int_a^b f(x)\,dx$$

which represents the (signed) area of the region between the graph of $f(x)$ and the interval $[a, b]$ on the x-axis.

Example. Consider a constant function $f(x) = L$ for some $L \in \mathbb{R}$. Then

$$\int_a^b L\,dx = L(b - a).$$

See Figure 5.1 below.

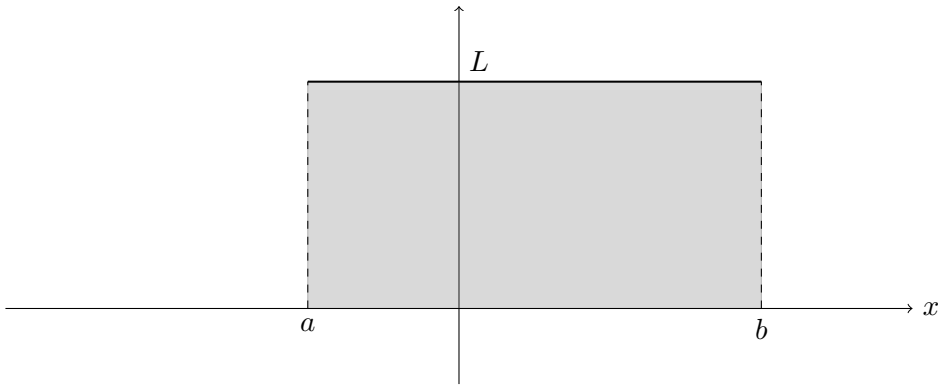


FIGURE 5.1. The integral $\int_a^b L\,dx$ equals the (signed) area of a rectangle with base $b - a$ and height L.

- Fundamental theory of calculus: Suppose that $F'(x) = f(x)$ on $[a, b]$. Then
$$\int_a^b f(x)\,dx = F(b) - F(a).$$
- The integral of an (integrable) odd function is 0: Suppose that $f(-x) = -f(x)$ on $[-a, a]$. Then $\int_{-a}^a f = 0$. For example, assume that n is odd. Then
$$\int_{-a}^a x^n\,dx = 0 \quad \text{for each } a > 0.$$
Moreover, noticing that $x^n e^{-x^2}$ is odd (and is integrable),
$$\int_{-\infty}^{\infty} x^n e^{-x^2}\,dx = 0.$$
- Integration by parts:
$$\int_a^b u\,dv = uv\Big|_a^b - \int_a^b v\,du.$$

LEMMA 5.1.

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx = \sqrt{2\pi}.$$

PROOF. On $\mathbb{R}^2 = \{(x, y) : x, y \in \mathbb{R}\}$, change the variables such that $(x, y) = (r\cos\theta, r\sin\theta)$ with $r > 0$ and $\theta \in [0, 2\pi)$. Then

$$x^2 + y^2 = r^2 \quad \text{and} \quad \frac{\partial(x, y)}{\partial(r, \theta)} = r.$$

Hence,

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}e^{-\frac{1}{2}y^2}\,dxdy\\
=\;&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(x^2+y^2\right)}\,dxdy\\
=\;&\int_{0}^{\infty}\int_{0}^{2\pi} e^{-\frac{1}{2}r^2}\,r\,drd\theta\\
=\;&2\pi\int_{0}^{\infty} e^{-\frac{1}{2}r^2}\,r\,dr\\
=\;&2\pi\int_{0}^{\infty} e^{-\frac{1}{2}r^2}\,d\left(\frac{1}{2}r^2\right)\\
=\;&2\pi\int_{0}^{\infty} e^{-u}\,du\\
=\;&2\pi\cdot\left(-e^{-u}\right)\Big|_0^{\infty}\\
=\;&2\pi.
\end{aligned}$$

This means that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}e^{-\frac{1}{2}y^2}\,dxdy=\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx\cdot\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\,dy=\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx\right)^2=2\pi,$$

which implies that

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx=\sqrt{2\pi}.$$

□

5.2. Distribution and density functions

Definition (Continuous random variables). A continuous random variable $X : S \to \mathbb{R}$ is defined by a non-negative integrable function f on $\mathbb{R}$, called a probability density function, such that $\int_{-\infty}^{\infty} f = 1$ and

$$P(X \in I)=\int_I f(x)\,dx \quad \text{for each interval } I \subset \mathbb{R}.$$

In particular,

$$F(a)=P\left(X\in(-\infty,a)\right)=P\left(X<a\right)=\int_{-\infty}^{a} f(x)\,dx$$

is called the (cumulative) distribution function.

Remark. Let X be a random variable with probability density function f.

- The probability that X takes values in an interval I is given by $\int_I f(x)\,dx$, i.e., the area between the graph of $f(x)$ and the interval I on the x-axis.
- Since the random variable X always takes values in $\mathbb{R}$,

  $$P(-\infty<X<\infty)=P\left((-\infty,\infty)\right)=\int_{-\infty}^{\infty} f(x)\,dx=1,$$

  which is the condition in the definition. This also means that the total area below the graph of $f(x)$ on $\mathbb{R}$ is 1.
- For any $a<b$,

  $$P(a<X<b)=P(a\le X<b)=P(a<X\le b)=P(a\le X\le b)=\int_a^b f(x)\,dx.$$

  That is, the probability does not see the difference between open and closed intervals.

- For any $a \in \mathbb{R}$,
$$P(X = a) = P(X \le a) - P(X < a) = 0.$$
That is, the probability that X takes any one value is always zero.

Example. Suppose that a passenger arrives at a specified stop at a time that is uniformly distributed between 7 and 7:30 AM. Let X be the number of minutes past 7 that she arrives at the stop. This means that X is a random variable with probability density function
$$f(x) = \begin{cases} \frac{1}{30} & \text{if } 0 < x < 30, \\ 0 & \text{otherwise.} \end{cases}$$
For example,

(a). The probability that she arrives at the stop between 7:10 and 7:20 (i.e., X is between 10 and 20) is
$$P(10 < X < 20) = \int_{10}^{20} f(x)\,dx = \int_{10}^{20} \frac{1}{30}\,dx = \frac{10}{30} = \frac{1}{3}.$$

(b). The probability that she arrives at the stop between 7:25 and 7:35 (i.e., X is between 25 and 35) is
$$P(25 < X < 35) = \int_{25}^{35} f(x)\,dx = \int_{25}^{30} \frac{1}{30}\,dx = \frac{5}{30} = \frac{1}{6}.$$

(c). The probability that she arrives at the stop between 7:40 and 7:50 (i.e., X is between 40 and 50) is
$$P(40 < X < 50) = \int_{40}^{50} f(x)\,dx = 0.$$

Definition (Uniform random variables). Let $\alpha < \beta$. We say that X is a uniform random variable on (α, β), if the probability density function of X is given by
$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if } \alpha < x < \beta, \\ 0 & \text{otherwise.} \end{cases}$$

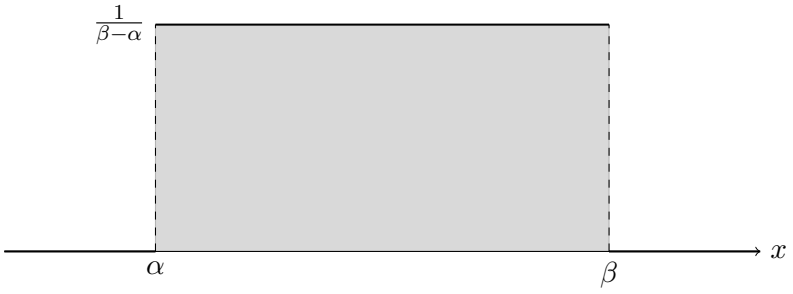


FIGURE 5.2. The probability density function of the uniform random variable on (α, β)

Notice that f is integrable, which satisfies that
$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{\alpha}^{\beta} \frac{1}{\beta-\alpha}\,dx = (\beta-\alpha)\cdot\frac{1}{\beta-\alpha} = 1.$$
It indicates that the area of the gray area in Figure 5.2 is 1.

Remark. Given any interval $I \subset (\alpha, \beta)$, the probability
$$P(X \in I) = \int_I \frac{1}{\beta-\alpha}\,dx = \frac{|I|}{\beta-\alpha},$$

which is the ratio of the length of I in the interval (α, β). In particular,

$$P(\alpha < X < \beta) = 1.$$

In general, for any interval $I \subset \mathbb{R}$,

$$P(X \in I) = \frac{|I \cap (\alpha, \beta)|}{\beta - \alpha}.$$

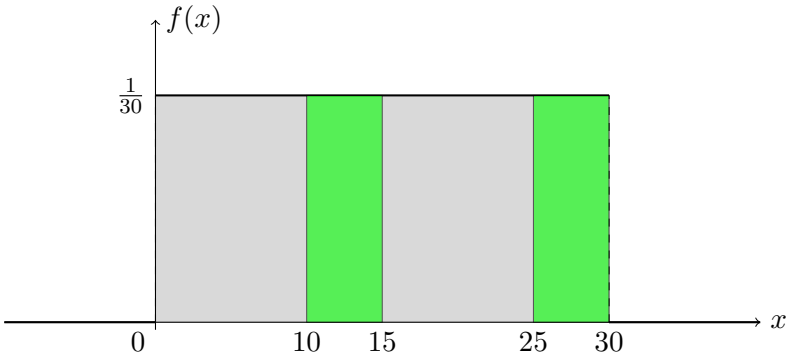
Example. Buses arrive at a specified stop at 15-minute intervals staring at 7 AM. That is, they arrive at 7, 7:15, 7:30, 7:45, and so on. Suppose that a passenger arrives at the stop at a time that is uniformly distributed between 7 and 7:30. Let X be the number of minutes past 7 that she arrives at the stop. Then X is a random variable on $(0, 30)$ with probability density function

$$f(x) = \begin{cases} \frac{1}{30} & \text{if } 0 < x < 30, \\ 0 & \text{otherwise.} \end{cases}$$

(a). We find the probability that she waits less than five minutes for a bus. To this end, she must arrive at the stop between 7:10 and 7:15 (i.e., X is between 10 and 15), or between 7:25 and 7:30 (i.e., X is between 25 and 30). Hence, the desired probability is

$$P(10 < X < 15) + P(25 < X < 30) = \int_{10}^{15} \frac{1}{30}\, dx + \int_{25}^{30} \frac{1}{30}\, dx = \frac{1}{3}.$$

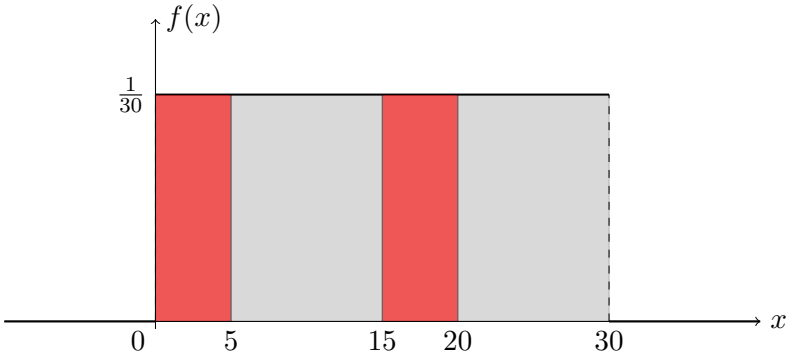
It is represented by the area of the green region below.



(b). We find the probability that she waits more than 10 minutes for a bus. To this end, she must arrive at the stop between 7 and 7:05 (i.e., X is between 0 and 5), or between 7:15 and 7:20 (i.e., X is between 15 and 20). Hence, the desired probability is

$$P(0 < X < 5) + P(15 < X < 20) = \int_{0}^{5} \frac{1}{30}\, dx + \int_{15}^{20} \frac{1}{30}\, dx = \frac{1}{3}.$$

It is represented by the area of the red region below.

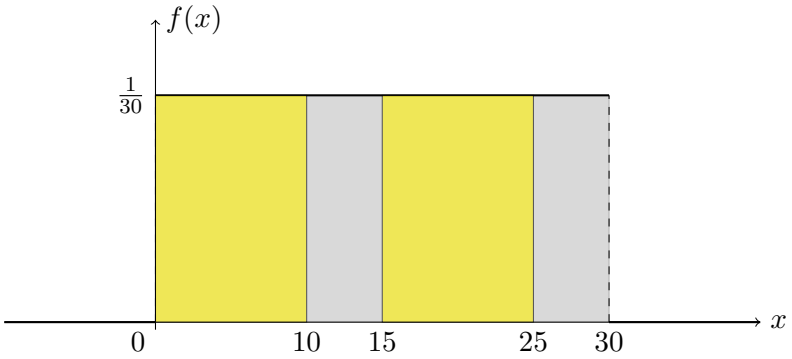


Remark. Notice that the probabilities of the events of "she waits less than five minutes for a bus" and of "she waits more than 10 minutes for a bus" are the same. A similar argument shows that, for any given $0 < a < 15$, the probabilities of the events of "she waits less than a minutes for a bus" and of "she waits more than $15 - a$ minutes for a bus" are also the same. For example, there is an equal chance for her to get on a bus within one minute or to wait for the bus more than 14 minutes. This may explain the perception that one has equal chance to get on a bus shortly after arriving at a stop or to wait for a bus at the stop for a long time.

(c). We find the probability that she waits more than five minutes for a bus. To this end, she must arrive at the stop between 7 and 7:10 (i.e., X is between 0 and 10), or between 7:15 and 7:25 (i.e., X is between 15 and 25). Hence, the desired probability is

$$P(0 < X < 10) + P(15 < X < 25) = \int_0^{10} \frac{1}{30}\, dx + \int_{15}^{25} \frac{1}{30}\, dx = \frac{2}{3}.$$

It is represented by the area of the yellow region below. It is exactly the complement of the green region in (1) under the graph of $f(x)$, which indicates that the events of "she waits less than (or equal to) five minutes for a bus" and of "she waits more than five minutes for a bus" are complement to each other. Therefore, their probabilities sum to 1. We generalize this phenomenon in the following proposition.



PROPOSITION 5.2. *Let X be a random variable. For any (measurable) $E \subset \mathbb{R}$,*

$$P\left(X \in E^c\right) = P(X \notin E) = 1 - P(X \in E).$$

Definition (Exponential random variables). Let $\lambda > 0$. We say that X is an exponential random variable with parameter λ, if the probability density function of X is given by

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

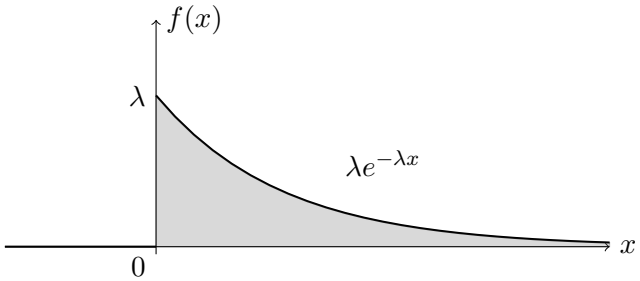


FIGURE 5.3. The probability density function of an exponential random variable with parameter λ

Notice that f is integrable, which satisfies that

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{\infty} \lambda e^{-\lambda x} = -e^{-\lambda x}\Big|_0^{\infty} = -\lim_{x\to\infty} e^{-\lambda x} - \left(-e^{-\lambda\cdot 0}\right) = 1.$$

It indicates that the area of the gray area in Figure 5.3 is 1.

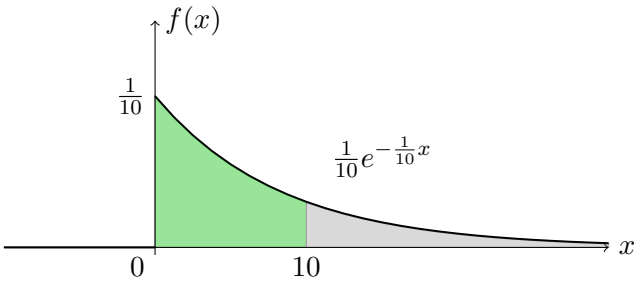
Example. A customer calls a service line. Suppose that the waiting time in minutes is an exponential random variable with parameter $\lambda = \frac{1}{10}$. Let X be the number of minutes she has to wait. Then X is a random variable with the probability density function

$$f(x) = \begin{cases} \frac{1}{10}e^{-\frac{x}{10}} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

(a). The probability that she has to wait less than 10 minutes is

$$P(X < 10) = \int_{-\infty}^{10} f(x)\,dx = \int_{0}^{10} \frac{1}{10}e^{-\frac{x}{10}}\,dx \approx 0.632.$$

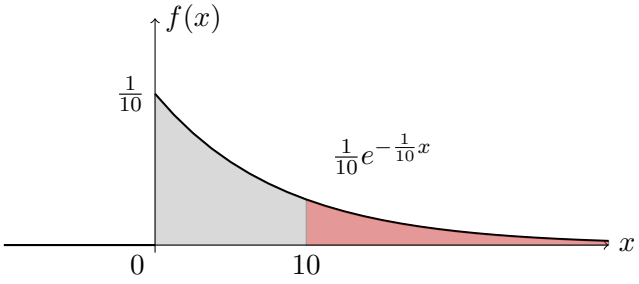
It is represented by the area of the green region below. Notice that $P(X < 10)$ is the distribution function F at 10.



(b). The probability that she has to wait more than 10 minutes is

$$F(10) = P(X > 10) = \int_{10}^{\infty} f(x)\,dx = \int_{10}^{\infty} \frac{1}{10}e^{-\frac{x}{10}}\,dx \approx 0.368.$$

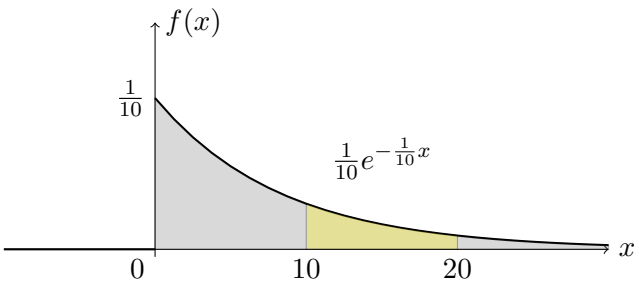
It is represented by the area of the red region below. Notice that that the probabilities of $X > 10$ and of $X < 10$ sum to 1. This follows from the fact that the events of $X \geq 10$ and of $X < 10$ are complement to each other, so $P(X \geq 10) + P(X < 10) = 1$. But $P(X \geq 10) = P(X > 10)$.



(c). The probability that she has to wait between 10 and 20 minutes is

$$P(10 < X < 20) = \int_{10}^{20} f(x)\,dx = \int_{10}^{20} \frac{1}{10}e^{-\frac{x}{10}}\,dx \approx 0.233.$$

It is represented by the area of the yellow region below.



5.3. Expectation and variance

Definition (Expectation and variance). Let X be a random variable with probability density function f. Then the expectation of X is defined as

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx,$$

and the variance of X is defined as

$$\sigma^2 = \mathrm{Var}[X] = \mathbb{E}\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\, dx.$$

The standard derivation is defined as $\sigma = \sqrt{\mathrm{Var}[X]}$.

Similar to the case in the discrete setting, taking expectation commutes with linear operations: Let X be a random variable. Suppose that $a, b \in \mathbb{R}$. Then $aX + b$ is also a random variable, whose expectation $\mathbb{E}[aX+b] = a\mathbb{E}[X] + b$ and variance $\mathrm{Var}[aX+b] = a^2\mathrm{Var}[X]$.

Remark (Moments). Let X be a random variable with probability density function f. Then

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}\left[(X-\mu)^2\right] \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\, dx \\
&= \int_{-\infty}^{\infty} \left(x^2 - 2\mu x + \mu^2\right) f(x)\, dx \\
&= \int_{-\infty}^{\infty} x^2 f(x)\, dx - 2\mu \int_{-\infty}^{\infty} x f(x)\, dx + \mu^2 \int_{-\infty}^{\infty} f(x)\, dx \\
&= E\left[X^2\right] - \mu^2.
\end{aligned}$$

Hence,

$$E\left[X^2\right] = \sigma^2 + \mu^2,$$

which is called the second moment of X. More generally, for $n \in \mathbb{N}$, we call $\mathbb{E}[X^n]$ the n-th moment of X.

Proposition 5.3. *Let $\alpha < \beta$ and X be a uniform random variable on (α, β). Then*

$$\mathbb{E}[X] = \frac{\beta+\alpha}{2} \quad \textit{and} \quad \mathrm{Var}[X] = \frac{(\beta-\alpha)^2}{12}.$$

Proof. See Question 5.4 for the expectation. Compute that

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \int_{\alpha}^{\beta} x^2 \cdot \frac{1}{\beta-\alpha}\, dx \\
&= \frac{1}{\beta-\alpha}\left(\frac{1}{3}\beta^3 - \frac{1}{3}\alpha^3\right) \\
&= \frac{\beta^3-\alpha^3}{3(\beta-\alpha)}
\end{aligned}$$

$$= \frac{\beta^2 + \beta\alpha + \alpha^2}{3}.$$

Hence, the variance

$$\begin{aligned}
\mathrm{Var}\left[X^2\right] &= \mathbb{E}\left[X^2\right] - \mu^2 \\
&= \frac{\beta^2 + \beta\alpha + \alpha^2}{3} - \left(\frac{\beta+\alpha}{2}\right)^2 \\
&= \frac{\beta^2 + \beta\alpha + \alpha^2}{3} - \frac{\beta^2 + 2\beta\alpha + \alpha^2}{4} \\
&= \frac{\beta^2 - 2\beta\alpha + \alpha^2}{12} \\
&= \frac{(\beta-\alpha)^2}{12}.
\end{aligned}$$

□

PROPOSITION 5.4. *Let $\lambda > 0$ and X be an exponential random variable with parameter λ. Then*

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad \textit{and} \quad \mathrm{Var}[X] = \frac{1}{\lambda^2}.$$

PROOF. See Question 5.6 for the expectation. Using integration by parts,

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \int_0^\infty x^2 \cdot \lambda e^{-\lambda x}\, dx \\
&= \int_0^\infty x^2\, d\left(-e^{-\lambda x}\right) \\
&= -x^2 e^{-\lambda x}\Big|_0^\infty - \int_0^\infty \left(-e^{-\lambda x}\right)\, d\left(x^2\right) \\
&= 2\int_0^\infty x \cdot e^{-\lambda x}\, dx \\
&= \frac{2}{\lambda}\int_0^\infty x \cdot \lambda e^{-\lambda x}\, dx \\
&= \frac{2}{\lambda}\mathbb{E}[X] \\
&= \frac{2}{\lambda^2}.
\end{aligned}$$

Hence, the variance

$$\begin{aligned}
\mathrm{Var}\left[X^2\right] &= \mathbb{E}\left[X^2\right] - \mu^2 \\
&= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\
&= \frac{1}{\lambda^2}.
\end{aligned}$$

□

Remark (Infinite expectation and variance). We shall point out that expectation and variance of a random variable may be infinite. For example, let $r > 1$, consider the function

$$f(x) = \begin{cases} \frac{r-1}{x^r} & \text{if } x > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Since

$$\int_{-\infty}^\infty f(x)\, dx = (r-1)\int_1^\infty x^{-r}\, dx = (r-1)x^{1-r}\Big|_1^\infty = 1,$$

f defines a probability density function of a random variable, denoted by X.

Example. Set $r \in (1,2)$. Then the expectation

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = (r-1)\int_{1}^{\infty} x^{1-r}\,dx = (r-1)x^{2-r}\Big|_{1}^{\infty} = \infty.$$

Example. Set $r \in (2,3)$. Then the expectation

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = (r-1)\int_{1}^{\infty} x^{1-r}\,dx = (r-1)x^{2-r}\Big|_{1}^{\infty} = r-1.$$

However, the second moment

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx = (r-1)\int_{1}^{\infty} x^{2-r}\,dx = (r-1)x^{3-r}\Big|_{1}^{\infty} = \infty.$$

Hence, the variance

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \infty.$$

We summarize the expectations and variances of the continuous random variables, in which the normal random variable is discussed in the next section.

TABLE 5.1. Table of expectations and variances of three continuous random variables

Random variable	Probability density function	Expectation	Variance
Uniform on (α,β)	$f(x) = \frac{1}{\beta-\alpha}$ on (α,β) and $=0$ otherwise	$\frac{\beta+\alpha}{2}$	$\frac{(\beta-\alpha)^2}{12}$
Exponential with parameter λ	$f(x) = \lambda e^{-\lambda x}$ on $(0,\infty)$ and $=0$ otherwise	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Standard normal	$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$	0	1

5.4. Normal distribution

Gauss[i] coined the term "normal" in the normal random variable and distribution, which are now also commonly called "Gaussian".

Definition (Standard normal random variable). We say that X is a standard normal random variable, if the probability density function of X is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \text{for all } x \in \mathbb{R}.$$

The standard normal distribution is

$$\Phi(a) = P(X < a) = \int_{-\infty}^{a} \phi(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{1}{2}x^2}\,dx.$$

[i]Carl Friedrich Gauss, *Theoria combinationis observationum erroribus minimis obnoxiae.* [Theory of the combination of observations least subject to errors]. (1823).

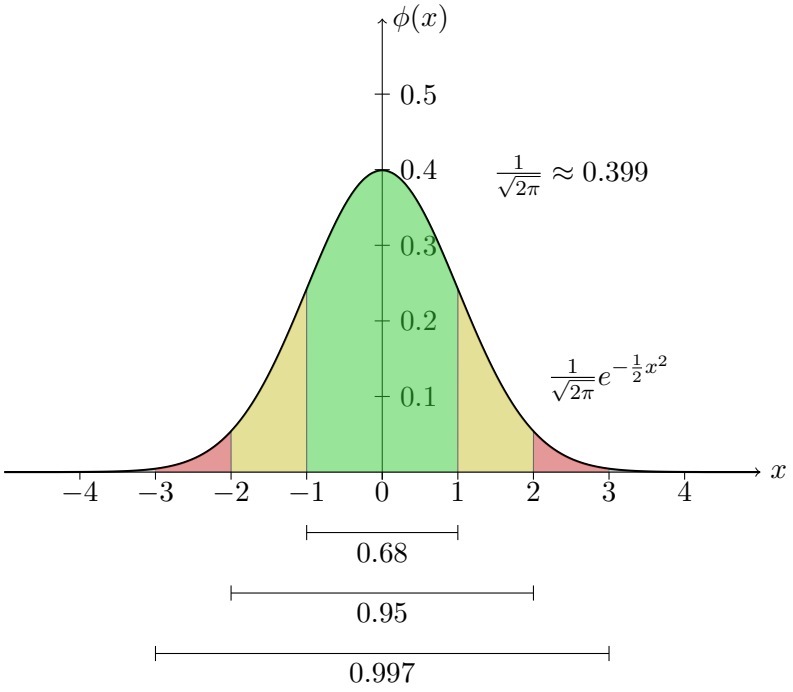


FIGURE 5.4. The probability density function of the standard normal random variable and the 68-95-99.7 rule

Notice that f is integrable, which satisfies that

$$\int_{-\infty}^{\infty} \phi(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, dx = 1.$$

PROPOSITION 5.5 (Symmetry of the standard normal random variable). *Let X be a standard normal random variable.*

(i). The probability density function of the standard normal random variable is even, that is, $\phi(-x) = \phi(x)$ for all $x \in \mathbb{R}$.
(ii). For each $a \in \mathbb{R}$, $P(X > a) = P(X < -a)$. In particular, $P(X > 0) = P(X < 0) = \frac{1}{2}$.
(iii). For each $a \in \mathbb{R}$, $\Phi(-a) = 1 - \Phi(a)$.
(iv). For each $a > 0$, $P(-a < X < a) = 2\Phi(a) - 1$.

PROOF. (i) and (ii) are obvious since $f(x)$ is even.

(iii).

$$\Phi(-a) = P(X < -a) = 1 - P(X > -a) = 1 - P(X < a) = 1 - \Phi(a).$$

(iv).

$$P(-a < X < a) = P(X < a) - P(X < -a) = \Phi(a) - \Phi(-a) = \Phi(a) - (1 - \Phi(a)) = 2\Phi(a) - 1.$$

□

Remark. In Table 5.2, we provide the approximation of $\Phi(a)$ for certain values of $a > 0$. Because of the symmetry in (iii) above, the distribution $\Phi(a)$ for $a < 0$ follows.

Example. Let X be the standard normal random variable.

(a).

$$P\left(-\frac{1}{3} < X < \frac{2}{3}\right) = P\left(X < \frac{2}{3}\right) - P\left(X < -\frac{1}{3}\right)$$

$$\begin{aligned} &= \Phi\left(\frac{2}{3}\right) - \left[1 - \Phi\left(\frac{1}{3}\right)\right] \\ &\approx 0.7486 - (1 - 0.6293) \\ &= 0.3779. \end{aligned}$$

(b).

$$P(X > -1) = P(X < 1) = \Phi(1) \approx 0.8413.$$

(c).

$$P(|X| > 2) = 1 - P(|X| < 2) = 1 - (2\Phi(2) - 1) = 2 - 2\Phi(2) \approx 2 - 2 \cdot 0.9772 = 0.0456.$$

THEOREM 5.6. *Let X be a standard normal random variable. Then $\mathbb{E}[X] = 0$ and $\mathrm{Var}[X] = 1$.*

PROOF. Notice that ϕ is even on $\mathbb{R}$. Then $x\phi(x)$ is odd on $\mathbb{R}$, so

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot \phi(x)\, dx = 0.$$

Using integration by parts,

$$\begin{aligned} \mathbb{E}\left[X^2\right] &= \int_{-\infty}^{\infty} x^2 \cdot \phi(x)\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \cdot e^{-\frac{1}{2}x^2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x d\left(-e^{-\frac{1}{2}x^2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \left(-xe^{-\frac{1}{2}x^2}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \left(-e^{-\frac{1}{2}x^2}\right)\, dx\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, dx \\ &= 1. \end{aligned}$$

Hence, $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mu^2 = 1$. □

TABLE 5.2. Table of the standard normal distribution $\Phi(a)$ for $0 \leq a \leq 3.09$

a	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Remark (The 68-95-99.7 rule). Using the table in Table 5.2, we compute the probabilities of X taking values within 1, 2, and 3 (standard deviations).

- $P(-1 < X < 1) = 2\Phi(1) - 1 \approx 2 \cdot 0.8413 - 1 \approx 0.68$.
- $P(-2 < X < 2) = 2\Phi(2) - 1 \approx 2 \cdot 0.9772 - 1 \approx 0.95$.
- $P(-3 < X < 3) = 2\Phi(3) - 1 \approx 2 \cdot 0.9987 - 1 \approx 0.997$.

Definition (Normal random variables). Let $\mu, \sigma \in \mathbb{R}$. We say that X is an exponential random variable with parameters μ and σ^2, if the probability density function of X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for all } x \in \mathbb{R}.$$

The normal distribution is

$$F(a) = P(X < a) = \int_{-\infty}^{a} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

Suppose that X is a normal random variable with expectation μ and variance σ^2. Let $Y = \frac{X-\mu}{\sigma}$, i.e, $X = \sigma Y + \mu$. Then X is a standard random variable, that is, the expectation is 1 and variance is 1. See Question 4.5.

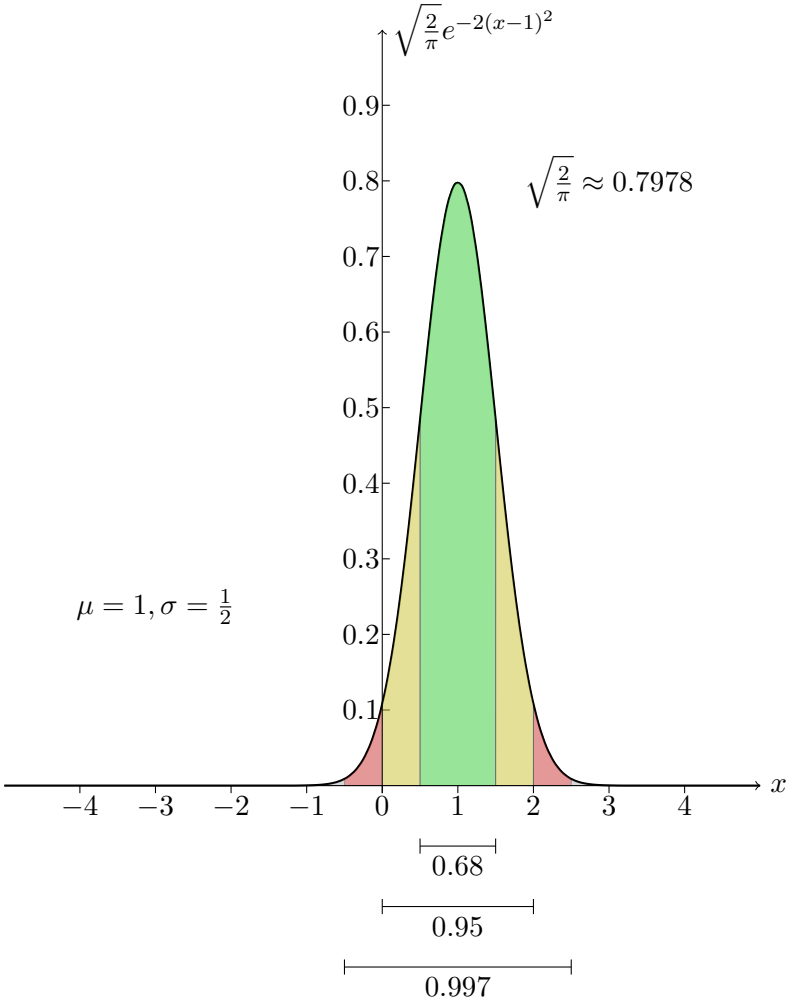


FIGURE 5.5. The probability density function of the standard normal random variable and the 68-95-99.7 rule

Remark (The 68-95-99.7 rule). Let X be the normal random variable with expectation μ and variance σ^2. Then $Y = \frac{X-\mu}{\sigma}$ is the standard normal random variable with expectation 1 and variance 1, i.e., $X = \sigma Y + \mu$. Hence,

$$P(|X - \mu| < k\sigma) = P(|(\sigma Y + \mu) - \mu| < k\sigma) = P(|Y| < k).$$

Letting $k = 1, 2, 3$, we have the following 68-95-99.7 rule for X:

- $P(|X - \mu| < \sigma) = P(|Y| < 1) \approx 0.68$.
- $P(|X - \mu| < 2\sigma) = P(|Y| < 2) \approx 0.95$.
- $P(|X - \mu| < 3\sigma) = P(|Y| < 3) \approx 0.997$.

Example. Let X be the normal random variable with expectation $\mu = 3$ and variance $\sigma^2 = 9$. Then $Y = \frac{X-3}{3}$ is the standard normal random variable, that is, $X = 3Y + 3$.
(a).

$$P(2 < X < 5) \quad = \quad P(2 < 3Y + 3 < 5)$$

$$\begin{aligned} &= P\left(-\frac{1}{3} < Y < \frac{2}{3}\right) \\ &= \Phi\left(Y < \frac{2}{3}\right) - \Phi\left(-\frac{1}{3}\right) \\ &= \Phi\left(\frac{2}{3}\right) - \left[1 - \Phi\left(\frac{1}{3}\right)\right] \\ &\approx 0.7486 - (1 - 0.6293) \\ &= 0.3779. \end{aligned}$$

(b).

$$P(X > 0) = P(3Y + 3 > 0) = P(Y > -1) = P(Y < 1) = \Phi(1) \approx 0.8413.$$

(c).

$$\begin{aligned} P(|X - 3| > 6) &= P(|(3Y + 3) - 3| > 6) \\ &= P(|Y| > 2) \\ &= 1 - P(|Y| < 2) \\ &= 1 - (2\Phi(2) - 1) \\ &= 2 - 2\Phi(2) \\ &\approx 2 - 2 \cdot 0.9772 \\ &= 0.0456. \end{aligned}$$

Example. Suppose that the height of the adult men in the world is distributed as a normal distribution with an expectation of 171 cm and a standard deviation of 7 cm. Let X be the normal random variable with expectation $\mu = 171$ and standard derivation 7. Then $Y = \frac{X-171}{7}$ is the standard normal random variable, that is, $X = 7Y + 171$. Compute that

$$P(X < 176) = P(7Y + 171 < 176) \approx P(Y < 0.71) \approx 0.7611,$$

that is, a man with a height of 176 cm (i.e., 5 feet and 9 inches) is taller than 0.7611 of the adult men.

Example. Suppose that the height of the adult men in the Netherlands is distributed as a normal distribution with an expectation of 183 cm and a standard deviation of 10.3 cm. Let X be the normal random variable with expectation $\mu = 3$ and standard deviation $\sigma = 10.3$. Then $Y = \frac{X-183}{10.3}$ is the standard normal random variable, that is, $X = 10.3Y + 183$. Compute that

$$P(X > 176) = P(10.3Y + 183 > 176) \approx P(Y > -0.68) = P(Y < 0.68) \approx 0.7517,$$

that is, a man with a height of 176 cm is shorter than 0.7517 of the adult men in the Netherlands, i.e., he is taller than 0.2483 of them.

Example. Suppose that the height of the adult men in the US is distributed as a normal distribution with an expectation of 70 inches and a standard deviation of 3 inches. Let X be the normal random variable with expectation $\mu = 70$ and standard deviation $\sigma = 3$. Then $Y = \frac{X-70}{3}$ is the standard normal random variable, that is, $X = 3Y + 70$. Note that Michael Jordan has a height of 6 feet and 6 inches, i.e., 78 inches. Compute that

$$P(X < 78) = P(3Y + 70 < 78) \approx P(Y < 2.67) \approx 0.9962,$$

that is, he is taller than 0.9962 of the adult men in the US, i.e., he is shorter than 0.0038 of them. Given that the population of adult men in the US is approximately 130 million, approximately 494,000 adult men are taller than Michael Jordan.

5.4.1. The normal approximation to the binomial distribution. The normal random variables and distributions were first introduced by de Moivre[i], who used it to approximate probabilities associated with binomial random variables when the parameter n is large and $p = \frac{1}{2}$. This result was later extended by Laplace[ii] to general p. The following DeMoivre-Laplace limit theorem states that, the binomial random variable X with parameters n and p, can be approximated by the normal distribution with expectation np and variance $np(1-p)$ as $n \to \infty$. That is,

$$\frac{X - np}{\sqrt{np(1-p)}}$$

can be approximated by the standard normal random variable.

THEOREM 5.7 (The De Moivre-Laplace limit theorem). *Let X be the binomial random variable with parameters n and p. Then for each interval $I \subset \mathbb{R}$,*

$$\lim_{n\to\infty} P\left(\frac{X - np}{\sqrt{np(1-p)}} \in I\right) = \int_I \phi(x)\,dx.$$

This is a special case of the central limit theorem, which is proved later in the book.

Example. Consider 100 independent tosses of a fair coin. Let X denote the number of heads that appear. Then X is a binomial random variable with parameters $n = 100$ and $p = 0.5$. It can be approximated by the normal random variable with expectation $np = 50$ and variance $np(1-p) = 25$. Hence,

$$Y = \frac{X - 50}{5}$$

can be approximated by the standard normal random variable. That is,

$$X = 5Y + 50$$

(a). The probability that more than 60 heads appear is

$$\begin{aligned} P(X > 60) &= P\left(5Y + 50 > 60\right) \\ &= P(Y > 2) \\ &= 1 - P(Y < 2) \\ &= 1 - \Phi(2) \\ &\approx 1 - 0.9772 \\ &= 0.0228. \end{aligned}$$

(b). The probability that fewer than 35 heads appear is

$$\begin{aligned} P(X < 35) &= P\left(5Y + 50 < 35\right) \\ &= P(Y < -3) \\ &= 1 - P(Y > -3) \\ &= 1 - \Phi(3) \\ &\approx 1 - 0.9987 \\ &= 0.0013. \end{aligned}$$

(c). According to the 68-95-99.7 rule, the probability that the number of heads that appear is between 45 and 55 (i.e., 50 ± 5) is approximately 0.68, is between 40 and 60 (i.e., $50 \pm 2 \cdot 5$) is approximately 0.95, and is between 35 and 65 (i.e., $50 \pm 3 \cdot 5$) is approximately 0.997.

Homework Assignment.

[i] Abraham de Moivre, *The doctrine of chances.* (1738).

[ii] Pierre-Simon Laplace, *Mémoire sur la probabilité des causes par les événements.* [Memoir on the probability of the causes of events]. (1774).

Question 5.1. Suppose that the lifetime of a certain type of electronic device (measured in hours) is a random variable X whose probability density function

$$f(x) = \begin{cases} \frac{1}{x^2} & \text{if } x > 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a). Find $P(X > 4)$.
(b). Find the distribution function $F(a)$.
(c). Find the probability that four independent such type devices, at least two function more than four hours.

Answer.
(a). Compute that

$$P(X > 4) = \int_4^\infty f(x)\,dx = \int_4^\infty \frac{1}{x^2}\,dx = \frac{1}{4}.$$

(b). The distribution function $F(a) = \int_{-\infty}^a f(x)\,dx = 0$ if $a \leq 1$, and if $a > 1$, then

$$F(a) = P(X < a) = \int_{-\infty}^a f(x)\,dx = \int_1^a \frac{1}{x^2}\,dx = 1 - \frac{1}{a}.$$

(c). The probability that one such device functions at least four hours is given by $P(X > 4) = \frac{1}{4}$, while the probability that one such device functions less than four hours is given by $P(X < 4) = \frac{3}{4}$. Hence, the required probability is given by

$$\binom{4}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^2 + \binom{4}{3}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^1 + \binom{4}{4}\left(\frac{1}{4}\right)^4 = \frac{31}{64}.$$

Question 5.2. A Bus arrives at a specified stop at a time that is uniformly distributed between 7 and 7:30 AM. Suppose that a passenger arrives at the stop at 7.
(a). Find the probability that she will have to wait more than 14 minutes?
(b). Suppose that at 7:15 the bus has not yet arrived. Find the probability that she will have to wait more than an additional seven minutes?

Answer. Let X be the number of minutes past 7 that the bus arrives at the stop. This means that X is a random variable with probability density function

$$f(x) = \begin{cases} \frac{1}{30} & \text{if } 0 < x < 30, \\ 0 & \text{otherwise.} \end{cases}$$

(a). We find the probability that she waits more than 14 minutes for a bus. To this end, the bus must arrive at the stop between 7:14 and 7:30 (i.e., X is between 14 and 30). Hence, the desired probability is

$$P(14 < X < 30) = \int_{14}^{30} \frac{1}{30}\,dx = \frac{8}{15}.$$

(b). Let E be the event that at 7:15 the bus has not yet arrived, which means that the bus arrives at the stop between 7:15 and 7:30 (i.e., X is between 15 and 30). Then

$$P(E) = P(15 < X < 30) = \int_{15}^{30} \frac{1}{30}\,dx = \frac{15}{30}.$$

Let F be the event that she will have to wait more than an additional 7 minutes, which means that the bus arrives at the stop between 7:22 and 7:30 (i.e., X is between 22 and 30). Then

$$P(F) = P(0 < X < 22) = \int_{22}^{30} \frac{1}{30}\,dx = \frac{8}{30}.$$

Notice that $F \subset E$. Hence, the desired conditional probability is

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{P(F)}{P(E)} = \frac{8}{15}.$$

Question 5.3. Let X be a random variable X whose probability density function

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

(a). Find the expectation $\mu = \mathbb{E}[X]$.
(b). Find the variance $\mathrm{Var}[X]$.

Answer.
(a). Compute that

$$\begin{aligned} \mu &= \mathbb{E}[X] \\ &= \int_{-\infty}^{\infty} x \cdot f(x)\, dx \\ &= \int_{1}^{\infty} x \cdot \frac{3}{x^4}\, dx \\ &= \int_{1}^{\infty} \frac{3}{x^3}\, dx \\ &= \left(-\frac{3}{2}x^{-2}\right)\Big|_1^{\infty} \\ &= \frac{3}{2}. \end{aligned}$$

(b). Compute that

$$\begin{aligned} \mathbb{E}\left[X^2\right] &= \int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx \\ &= \int_{1}^{\infty} x^2 \cdot \frac{3}{x^4}\, dx \\ &= \int_{1}^{\infty} \frac{3}{x^2}\, dx \\ &= \left(-3x^{-1}\right)\Big|_1^{\infty} \\ &= 3. \end{aligned}$$

Hence, the variance

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mu^2 = 3 - \left(\frac{3}{2}\right)^2 = \frac{27}{4}.$$

Question 5.4. Let $\alpha < \beta$ and X be a uniform random variable on (α, β). Find the expectation $\mu = \mathbb{E}[X]$.

Answer. Compute that

$$\begin{aligned} \mu &= \mathbb{E}[X] \\ &= \int_{\alpha}^{\beta} x \cdot \frac{1}{\beta - \alpha}\, dx \\ &= \frac{1}{\beta - \alpha} \cdot \left(\frac{1}{2}\beta^2 - \frac{1}{2}\alpha^2\right) \end{aligned}$$

$$\begin{aligned} &= \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} \\ &= \frac{\beta + \alpha}{2}. \end{aligned}$$

Question 5.5. Let $\alpha < \beta$ and X be a uniform random variable on (α, β) such that the expectation $\mathbb{E}[X] = 10$ and the variance $\mathrm{Var}[X] = 48$. Find α and β.

Answer. From

$$\frac{\beta + \alpha}{2} = 10 \quad \text{and} \quad \frac{(\beta - \alpha)^2}{12} = 48,$$

we have that

$$\beta + \alpha = 20 \quad \text{and} \quad \beta - \alpha = 24.$$

Hence,

$$\beta = 22 \quad \text{and} \quad \alpha = -2.$$

Question 5.6. Let $\lambda > 0$ and X be an exponential random variable with parameter λ. Find the expectation $\mu = \mathbb{E}[X]$.

Answer. Using integration by parts,

$$\begin{aligned} \mu &= \mathbb{E}[X] \\ &= \int_0^\infty x \cdot \lambda e^{-\lambda x}\, dx \\ &= \int_0^\infty x\, d\left(-e^{-\lambda x}\right) \\ &= -xe^{-\lambda x}\Big|_0^\infty - \int_0^\infty \left(-e^{-\lambda x}\right)\, dx \\ &= \int_0^\infty e^{-\lambda x}\, dx \\ &= \frac{1}{\lambda}. \end{aligned}$$

Question 5.7. Suppose that the probability that an offer sent by a university is accepted by the student is 0.8. Assume that 10,000 offers have been sent in 2023 and each student's decision of acceptance is independent of others.

(a). Find the approximate probability that the new enrollment is larger than 8,100.

(b). Find the approximate probability that the new enrollment is smaller than 7,950.

Answer. Let X be the new enrollment in 2023, i.e., the number of students who accepted the offer. Then X is a binomial random variable with parameters $n = 10000$ and $p = 0.8$. It can be approximated by the normal random variable with expectation $np = 8000$ and variance $np(1-p) = 1600$. Hence,

$$Y = \frac{X - 8000}{40}$$

can be approximated by the standard normal random variable. That is,

$$X = 40Y + 8000.$$

(a). The probability that the new enrollment is larger than 8,100 is

$$\begin{aligned} P(X > 8100) &= P\left(40Y + 8000 > 8100\right) \\ &= P(Y > 2.5) \\ &= 1 - P(Y < 2.5) \\ &\approx 1 - \Phi(2.5) \\ &\approx 1 - 0.9938 \end{aligned}$$

$$= 0.0062.$$

(b). The probability that the new enrollment is smaller than 7,950 is

$$\begin{aligned} P(X < 7950) &= P\left(40Y + 8000 < 7950\right) \\ &\approx P(Y < -1.25) \\ &= 1 - P(Y > -1.25) \\ &\approx 1 - \Phi(1.25) \\ &\approx 1 - 0.8944 \\ &= 0.1056. \end{aligned}$$

In fact, according to the 68-95-99.7 rule, the probability that the new enrollment is between 8,040 and 7,960 (i.e., 8000 ± 40) is approximately 0.68, is between 8,080 and 7,920 (i.e., $8000 \pm 2 \cdot 40$) is approximately 0.95, and is between $8,120$ and $7,880$ (i.e., $8000 \pm 3 \cdot 40$) is approximately 0.997.

CHAPTER 6

Jointly distributed random variables

In everyday life we feel, and justifiably feel, irritated with the man who is perpetually asking us to define the words we use.

Karl Pearson, 1941[i]

Thus far, we have concerned ourselves only with probability distributions for single random variables (discrete in Chapter 4 and continuous in Chapter 5). However, we are often interested in probability statements concerning two or more random variables. For example, the grade a student obtained in the prerequisite and the score she obtains in a course can be both regarded as random variables (on the sample space of the class). How do we determine whether they are independent, i.e., the performance a student in a course is independent of her prerequisite?

In this chapter, we introduce the jointly distributed random variables and then the important concept of independent random variables. For simplicity, we only discuss the discrete case. Let S be a discrete probability space with a probability P throughout the chapter.

6.1. Examples

Suppose that X is a random variable with probability mass function p_X. That is, $p_X(x) = P(X = x)$ provides the probability that X takes the value x. Suppose that Y be a random variable with probability mass function p_Y. That is, $p_Y(y) = P(Y = y)$ provides the probability that Y takes the value y.

Then (X, Y) defines a joint discrete random variable, for which the joint probability mass function of X and Y is defined as

$$p(x, y) = P(X = x, Y = y).$$

Next, we define conditional probability mass function and independence of random variables. To this end, we recall these concepts for events: Let $E, F \subset S$ and $P(F) > 0$. Then the conditional probability of E for given F (also called the conditional probability that E occurs given that F has occurred) is defined as

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

We say that E and F are independent, if $P(E|F) = P(E)$, that is, $P(E \cap F) = P(E)P(F)$.

Switching to random variables X and Y, we replace the events E and F above by $X = x$ and $Y = y$ for values of x and y:

Definition (Conditional probability mass function and independence). Let X and Y be two random variables with probability mass functions p_X and p_Y, respectively. Suppose that p is the joint probability density function of X and Y. Then the conditional probability mass function $p_{X|Y}$ is defined as

$$p_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x, y)}{p_Y(y)} \quad \text{for } p_Y(y) > 0.$$

We say that X and Y are independent, if

$$p_{X|Y}(x|y) = p_X(x) \quad \text{for all } x, y,$$

[i]Karl Pearson, *The laws of chance, in relation to thought and conduct.* (1941).

that is, $p(x,y) = p_X(x)p_Y(y)$ for all x, y.

Remark. If two random variables X and Y are independent, then knowing that X takes a value does not change the probability distribution of Y, vice versa.

Example. Suppose that the class of *Introduction to Probability* is consisted of 10 students. Equip this sample space S with a probability P such that each sample point has the same probability (of $\frac{1}{10}$). Let X denote the grade (out of A, B) a student obtained in the prerequisite *Calculus* and Y denote the score (out of $1, 0.8, 0.6, 0.4, 0.2, 0$) a student obtained in *Introduction to Probability*. Then X and Y are two random variables. Assume that the grades and scores of the students are $(A,1)$, $(A,1)$, $(A,0.8)$, $(A,0.4)$, $(B,1)$, $(B,0.8)$, $(B,0.6)$, $(B,0.6)$, $(B,0.6)$, $(B,0.2)$. The joint probability mass function p of X and Y is given by

(X,Y)	1	0.8	0.6	0.4	0.2	0
A	$\frac{2}{10}$	$\frac{1}{10}$	0	$\frac{1}{10}$	0	0
B	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{3}{10}$	0	$\frac{1}{10}$	0

(a). The sum in each row provides the probability mass function p_X:

$$p_X(A) = \frac{4}{10} \quad \text{and} \quad p_X(B) = \frac{6}{10},$$

which then sum to 1.

The sum in each column provides the probability mass function p_Y:

$$p_Y(1) = \frac{3}{10}, \quad p_Y(0.8) = \frac{2}{10}, \quad p_Y(0.6) = \frac{3}{10}, \quad p_Y(0.4) = \frac{1}{10}, \quad p_Y(0.2) = \frac{1}{10}, \quad p_Y(0) = 0,$$

which also sum to 1.

(b). Compute that

$$p_{X|Y}(A|1) = \frac{2}{3} \quad \text{and} \quad p_{X|Y}(B|0.4) = 0.$$

Compute also that

$$p_{Y|X}(1|A) = \frac{1}{2} \quad \text{and} \quad p_{Y|X}(0.4|B) = 0.$$

(c). Notice that

$$p(A,1) \neq p_X(A)p_Y(1).$$

Therefore, X and Y are dependent. Indeed, knowing that the random variable X takes the value of A changes the probability that the random variable Y takes the value of 1 to $P_{Y|X}(1|A) = \frac{1}{2}$, from $p_Y(1) = \frac{3}{10}$ in the case when X is not known.

PROPOSITION 6.1. *Let X and Y be two random variables with probability mass functions p_X and p_Y, respectively. Suppose that p is the joint probability density function of X and Y. Then*

$$p_X(x) = \sum_y p(x,y) \quad \textit{and} \quad p_Y(y) = \sum_x p(x,y).$$

Moreover,

$$\sum_{x,y} p(x,y) = \sum_x p_X(x) = \sum_y p_Y(y) = 1.$$

Example. Under the same setup as the example as before, assume that the grades and scores of the students are $(A,1)$, $(A,1)$, $(A,0.6)$, $(A,0.6)$, $(B,1)$, $(B,1)$, $(B,1)$, $(B,0.6)$, $(B,0.6)$, $(B,0.6)$. The joint probability mass function p of X and Y is given by

(X,Y)	1	0.8	0.6	0.4	0.2	0
A	$\frac{2}{10}$	0	$\frac{2}{10}$	0	0	0
B	$\frac{3}{10}$	0	$\frac{3}{10}$	0	0	0

(a). The sum in each row provides the probability mass function p_X:

$$p_X(A) = \frac{4}{10} \quad \text{and} \quad p_X(B) = \frac{6}{10},$$

which then sum to 1.

The sum in each column provides the probability mass function p_Y:

$$p_Y(1) = \frac{5}{10}, \quad p_Y(0.8) = 0, \quad p_Y(0.6) = \frac{5}{10}, \quad p_Y(0.4) = 0, \quad p_Y(0.2) = 0 \quad p_Y(0) = 0,$$

which also sum to 1.

(b). Compute that

$$p_{X|Y}(A|1) = \frac{2}{5}, \quad p_{X|Y}(B|1) = \frac{3}{5}, \quad p_{X|Y}(A|0.6) = \frac{2}{5}, \quad p_{X|Y}(B|0.6) = \frac{3}{5}.$$

Compute also that

$$p_{Y|X}(1|A) = \frac{1}{2}, \quad p_{Y|X}(0.6|A) = \frac{1}{2}, \quad p_{Y|X}(1|B) = \frac{1}{2}, \quad p_{Y|X}(0.6|B) = \frac{1}{2}.$$

(c). Compute that

$$p(A,1) = p_X(A)p_Y(1), \quad p(A,0.6) = p_X(A)p_Y(0.6),$$
$$p(B,1) = p_X(B)p_Y(1), \quad p(B,0.8) = p_X(B)p_Y(0.8).$$

Therefore, X and Y are independent.

Example. Consider three independent tosses of a fair coin. Then the sample space

$$S = \{(s_1, s_2, s_3) : s_1, s_2, s_3 \in \{\mathrm{h}, \mathrm{t}\}\}$$

has size eight. Let X denote the number of heads that appear and Y denote the number of tails that appear. Then X and Y are two random variables. The joint probability mass function p of X and Y is given by

(X,Y)	0	1	2	3
0	0	0	0	$\frac{1}{8}$
1	0	0	$\frac{3}{8}$	0
2	0	$\frac{3}{8}$	0	0
3	$\frac{1}{8}$	0	0	0

(a). The sum in each row provides the probability mass function p_X:

$$p_X(0) = \frac{1}{8}, \quad p_X(1) = \frac{3}{8}, \quad p_X(2) = \frac{3}{8}, \quad p_X(3) = \frac{1}{8},$$

which then sum to 1. The sum in each column provides the probability mass function p_Y:

$$p_Y(0) = \frac{1}{8}, \quad p_Y(1) = \frac{3}{8}, \quad p_Y(2) = \frac{3}{8}, \quad p_Y(3) = \frac{1}{8},$$

which also sum to 1.

(b). The conditional probability mass function

$$p_{X|Y}(x|y) = \begin{cases} 1 & \text{if } x + y = 3, \\ 0 & \text{otherwise.} \end{cases}$$

(c). Notice that

$$p(0,0) \neq p_X(0)p_Y(0).$$

Therefore, X and Y are dependent. Indeed, knowing that the random variable X takes a value x (i.e., the number of heads that appear) changes the probability that Y takes a value y (i.e., the number of tails that appear), in fact, Y can only take a unique value $3 - x$, from all four possible values in the case when X is not known.

Example. Consider two independent rolls of a fair die. Then the sample space

$$S = \{(s_1, s_2) : s_1, s_2 \in \{1, 2, 3, 4, 5, 6\}\},$$

which has size 36. Let $X = s_1 + s_2$ (i.e., the sum of the two numbers) and $Y = s_1 - s_2$ (i.e., the difference of the second number from the first one). Then X and Y are two random variables. To find the joint probability mass function p of X and Y, notice that

$$s_1 = \frac{X+Y}{2} \quad \text{and} \quad s_2 = \frac{X-Y}{2}.$$

They must have integer solutions in $\{1, 2, 3, 4, 5, 6\}$ for the $p(s_1, s_2)$ to be non-zero, and if so, s_1 and s_2 is uniquely determined by X and Y. Hence,

(X, Y)	-5	-4	-3	-2	-1	0	1	2	3	4	5
2	0	0	0	0	0	$\frac{1}{36}$	0	0	0	0	0
3	0	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0	0
4	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0
5	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0
6	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0
7	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$
8	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0
9	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0
10	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0
11	0	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0	0
12	0	0	0	0	0	$\frac{1}{36}$	0	0	0	0	0

(a). The sum in each row provides the probability mass function p_X:

$$p_X(2) = \frac{1}{36},\ p_X(3) = \frac{2}{36},\ p_X(4) = \frac{3}{36},\ p_X(5) = \frac{4}{36},\ p_X(6) = \frac{5}{36},\ p_X(7) = \frac{6}{36},$$
$$p_X(12) = \frac{1}{36},\ p_X(11) = \frac{2}{36},\ p_X(10) = \frac{3}{36},\ p_X(9) = \frac{4}{36},\ p_X(8) = \frac{5}{36},$$

which then sum to 1.

The sum in each column provides the probability mass function p_Y:

$$p_Y(-5) = \frac{1}{36},\ p_Y(-4) = \frac{2}{36},\ p_Y(-3) = \frac{3}{36},\ p_Y(-2) = \frac{4}{36},\ p_Y(-1) = \frac{5}{36},\ p_Y(0) = \frac{6}{36},$$
$$p_Y(5) = \frac{1}{36},\ p_Y(4) = \frac{2}{36},\ p_Y(3) = \frac{3}{36},\ p_Y(2) = \frac{4}{36},\ p_Y(1) = \frac{5}{36},$$

which also sum to 1.

(b). Compute that

$$p_{X|Y}(2|0) = \frac{1}{6} \quad \text{and} \quad p_{X|Y}(5|1) = \frac{1}{4}.$$

(c). Compute that

$$p_{Y|X}(0|2) = 1 \quad \text{and} \quad p_{Y|X}(1|5) = \frac{1}{4}.$$

(d). Notice that

$$p(2, -5) \neq p_X(2)p_Y(-5).$$

Therefore, X and Y are dependent. Indeed, knowing that the random variable X (i.e., the sum of two numbers) take a value of 2 determines that they must be the two numbers are $1, 1$, which then changes the probability that the random variable Y (i.e., their difference) takes a value of -5 to 0, from $p_Y(-5) = \frac{1}{36}$ in the case when X is not known.

The concepts discussed in this section can be generalized to any collection of random variables:

Definition. Let $X_1, ..., X_n$ be random variables with probability mass functions $p_1, ..., p_n$, respectively. Then $(X_1, ..., X_n)$ defines a joint discrete random variable, for which the joint probability mass function of $X_1, ..., X_n$ is defined as

$$p(x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n).$$

- We say that $X_1, ..., X_n$ are pairwise independent if X_i and X_j are independent for all $i, j = 1, ..., n$, that is,

$$P(X_i = x_i, X_j = x_j) = P(X_i = x_i) P(X_j = x_j) \quad \text{for all } x_i, x_j.$$

  As a consequence,

$$P(X_i \in A_i, X_j \in A_j) = P(X_i \in A_i) P(X_j \in A_j) \quad \text{for all } A_i, A_j.$$

- We say that $X_1, ..., X_n$ are independent if $X_{i_1}, ..., X_{i_k}$ are independent for all possible collections $i_1, ..., i_k = 1, ..., n$, that is,

$$P(X_{i_1} = x_{i_1}, ..., X_{i_k} = x_{i_k}) = P(X_{i_1} = x_{i_1}) \cdots P(X_{i_k} = x_{i_k}) \quad \text{for all } x_{i_1}, ..., x_{i_k}.$$

  As a consequence,

$$P(X_{i_1} \in A_{i_1}, ..., X_{i_k} \in A_{i_k}) = P(X_{i_1} \in A_{i_1}) \cdots P(X_{i_k} \in A_{i_k}) \quad \text{for all } A_{i_1}, ..., A_{i_k}.$$

6.2. Independent random variables

From the discussion in the previous section, multiple random variables being independent is a rather restrictive condition. However, it can be guaranteed in the following situation.

THEOREM 6.2. *Consider n independent trials. Suppose that the random variable X_i only depends on the i-th trial for $i = 1, ..., n$. Then $X_1, ..., X_n$ are independent.*

Example. Consider four independent tosses of a fair coin. Then the sample space

$$S = \{(s_1, s_2, s_3, s_4) : s_1, s_2, s_3, s_4 \in \{\mathrm{h}, \mathrm{t}\}\}$$

has size 16. Let X denote the number of heads that appear in the first two tosses and Y denote the number of heads that appear in the last two tosses. Then X and Y are two binomial random variables with parameters $(2, \frac{1}{2})$ and $(2, \frac{1}{2})$. The joint probability mass function p of X and Y is given by

(X, Y)	0	1	2
0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$
1	$\frac{2}{16}$	$\frac{4}{16}$	$\frac{2}{16}$
2	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

(a). The sum in each row provides the probability mass function p_X:

$$p_X(0) = \frac{1}{4}, \quad p_X(1) = \frac{1}{2}, \quad p_X(2) = \frac{1}{4},$$

which then sum to 1.

The sum in each column provides the probability mass function p_Y:

$$p_Y(0) = \frac{1}{4}, \quad p_Y(1) = \frac{1}{2}, \quad p_Y(2) = \frac{1}{4},$$

which also sum to 1.

(b). We verify that X and Y are independent:

$$\begin{aligned}
&p(0,0) = p_X(0)p_Y(0), \quad p(0,1) = p_X(0)p_Y(1), \quad p(0,2) = p_X(0)p_Y(2),\\
&p(1,0) = p_X(1)p_Y(0), \quad p(1,1) = p_X(1)p_Y(1), \quad p(1,2) = p_X(1)p_Y(2),\\
&p(2,0) = p_X(2)p_Y(0), \quad p(2,1) = p_X(2)p_Y(1), \quad p(2,2) = p_X(2)p_Y(2).
\end{aligned}$$

(c). Let $Z = X + Y$. Then Z is a random variable with probability mass function p_Z as follows.

$$\begin{aligned}
p_Z(0) &= P(X+Y=0) = P(X=0,Y=0) \\
&= p(0,0) = \frac{1}{16}, \\
p_Z(1) &= P(X+Y=1) = P(X=0,Y=1) + P(X=1,Y=0) \\
&= p(0,1) + p(1,0) = \frac{4}{16}, \\
p_Z(2) &= P(X+Y=2) = P(X=0,Y=2) + P(X=1,Y=1) + P(X=2,Y=0) \\
&= p(0,2) + p(1,1) + p(2,0) = \frac{6}{16}, \\
p_Z(3) &= P(X+Y=3) = P(X=1,Y=2) + P(X=2,Y=1) \\
&= p(1,2) + p(2,1) = \frac{4}{16}, \\
p_Z(4) &= P(X+Y=4) = P(X=2,Y=2) \\
&= p(2,2) = \frac{1}{16}.
\end{aligned}$$

Observe that Z denotes the number of heads that appear in the four independent tosses of a fair coin, which is therefore a binomial random variable with parameter $(4, \frac{1}{2})$. The above computation verifies this fact.

More generally, consider $n+m$ independent tosses of a coin, which lands on heads with probability p. Let X denote the number of heads in the first n tosses and Y denote the number of heads in the last m tosses. Then X and Y are independent binomial random variables with parameters (n,p) and (m,p), respectively. Observe that $X+Y$ counts the number of heads in the $n+m$ independent tosses, which is a binomial random variable with parameter $(n+m,p)$.

PROPOSITION 6.3. *Let X be a binomial random variable with parameters (n,p). Let Y be a binomial random variable with parameters (m,p). Suppose that X and Y are independent. Then $Z = X+Y$ is a binomial random variable with parameter $(n+m,p)$.*

PROOF. The probability mass functions p_X and p_Y of X and Y, respectively, are given by

$$p_X(i) = \binom{n}{i} p^i (1-p)^i \quad \text{for } 0 \le i \le n$$

and

$$p_Y(j) = \binom{m}{j} p^j (1-p)^{m-j} \quad \text{for } 0 \le j \le m.$$

Compute the probability mass function p_Z of Z for $0 \le r \le n+m$:

$$\begin{aligned}
p_Z(r) &= P(X+Y=r) \\
&= \sum_{i=0}^{r} P(X=i, Y=r-i) \\
&= \sum_{i=0}^{r} P(X=i)P(Y=r-i) \\
&= \sum_{i=0}^{r} p_X(i) p_Y(r-i) \\
&= \sum_{i=0}^{r} \binom{n}{i} p^i (1-p)^i \binom{m}{r-i} p^{r-i} (1-p)^{m-(r-i)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=0}^{r} \binom{n}{i}\binom{m}{r-i} p^r (1-p)^{(n+m)-r} \\
&= p^r (1-p)^{(n+m)-r} \sum_{i=0}^{r} \binom{n}{i}\binom{m}{k} \\
&= \binom{n+m}{r} p^r (1-p)^{(n+m)-r},
\end{aligned}
$$

in which we used a combinatorial identity in Question 1.5 that

$$\sum_{i=0}^{r} \binom{n}{i}\binom{m}{r-i} = \binom{n+m}{r}.$$

□

Remark. We saw that the sum of two independent binomial random variables with parameters (n, p) and (m, p) is still binomial. In the following proposition, we prove that the sum of two independent Poisson random variables is still Poisson.

However, we shall point out that this phenomenon is rather uncommon, that is, the sum of two random variables of the same type is in general not of the same type. For example, the sum of two independent geometric random variables is no longer a geometric random variable.

PROPOSITION 6.4. *Let X be a Poisson random variable with parameters λ_1. Let Y be a Poisson random variable with parameters λ_2. Suppose that X and Y are independent. Then $Z = X + Y$ is a Poisson random variable with parameter $\lambda_1 + \lambda_2$.*

The probability mass functions p_X and p_Y of X and Y, respectively, are given by

$$p_X(i) = e^{-\lambda_1} \frac{\lambda_1^i}{i!} \quad \text{for } i = 0, 1, 2, ...$$

and

$$p_Y(j) = e^{-\lambda_2} \frac{\lambda_2^j}{j!} \quad \text{for } j = 0, 1, 2, ...$$

Then we establish the probability mass function p_Z of $Z = X + Y$ similarly as above. See Question 6.3.

Remark. Recall that a Poisson random variable with parameter λ is an approximation for a binomial random variable X with parameter (n, p) when is n large and p is small enough so that $\lambda = np$ is of moderate size, see Section 4.4.2. We use this intuition to interpret the proposition above.

The probability that a person has birthday on a given day, say, February 26, is $p = \frac{1}{365} \approx 0.00274$. (Assume that there are 365 possible birthdays.)

Consider a group of n people. Then the number of people with birthday at February 26 is a binomial random variable with parameter (n, p). It can be approximated by a Poisson random variable X with parameter $\lambda_1 = np$.

Consider another group of m people. Then the number of people with birthday at February 26 is a binomial random variable with parameter (m, p). It can be approximated by a Poisson random variable X with parameter $\lambda_2 = mp$.

Now the Poisson random variable $Z = X + Y$ approximates the total number of people with birthday at February 26 among the $n + m$ ones. Hence, the parameter of Z must be

$$(n + m)p = np + mp = \lambda_1 + \lambda_2.$$

Definition (Independent and identically distributed (iid) random variables). We say that a collection of random variables are independent and identically distributed if they are independent and have the same probability distribution.

Example. Consider n independent tosses of a coin, which lands on heads with probability p. Let X_i, $i = 1, ..., n$, denote the number of heads in the i-th toss, i.e., $X_i(1) = p$ and $X_i(0) = 1 - p$. Then X_i are iid random variables.

6.3. Expectation

In Section 4.2, we defined the expectation of a random variable $X : S \to \mathbb{R}$ with probability mass function p_X:

$$\mathbb{E}[X] = \sum_{s \in S} X(s) \cdot P(\{s\}) = \sum_{x \in X(S)} x \cdot p_X(x).$$

We established several properties of expectation, for example, it is always bounded between the maximal value and minimal value of X. Moreover, taking expectation commutes with linear operations:

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b,$$

but it does not commute with nonlinear operations such as $X \to X^2$ and $X \to \sqrt{X}$. In particular, if $\mu = \mathbb{E}[X]$, then

$$\mathbb{E}[X - \mu] = 0.$$

In this section, we investigate the expectations involving multiple random variables.

PROPOSITION 6.5. *Let $X, Y : S \to \mathbb{R}$ be random variables. Suppose that $X \geq Y$, i.e., $X(s) \geq Y(s)$ for all $s \in S$. Then $\mathbb{E}[X] \geq \mathbb{E}[Y]$.*

PROOF. Compute that

$$\mathbb{E}[X] = \sum_{s \in S} X(s) \cdot P(\{s\}) \leq \sum_{s \in S} Y(s) \cdot P(\{s\}) = \mathbb{E}[Y].$$

□

THEOREM 6.6. *Let $X_1, ..., X_n : S \to \mathbb{R}$ be random variables and $X = X_1 + \cdots + X_n$. Then*

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

PROOF. Compute that

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X_1 + \cdots + X_n] \\ &= \sum_{s \in S} (X_1(s) + \cdots + X_n(s)) \cdot P(\{s\}) \\ &= \sum_{s \in S} X_1(s) \cdot P(\{s\}) + \cdots + \sum_{s \in S} X_n(s) \cdot P(\{s\}) \\ &= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]. \end{aligned}$$

□

Example. Consider three independent tosses of a fair coin. Let X denote the number of heads that appear and Y denote the number of tails that appear. Then X and Y are two random variables, both of which have expectations $\mathbb{E}[X] = \mathbb{E}[Y] = \frac{3}{2}$.

Consider $Z = X + Y$, that is, the sum of the number of heads and the number of tails. Hence,

$$\mathbb{E}[Z] = \mathbb{E}[X] + \mathbb{E}[Y] = 3.$$

This also follows from the fact that Z is the number of tosses, i.e., a constant random variable of 3. Notice that the random variables X and Y are dependent as shown in Section 6.1.

Example. Suppose that n people throw their hat and each person randomly selects one. We find the expectation of the number of the people who select their own hats. To this end, for $i = 1, ..., n$, define

$$X_i = \begin{cases} 1 & \text{if } i \text{ selects her own hat,} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\mathbb{E}[X_i] = 0 \cdot P(X_i = 0) + 1 \cdot P(X_i = 1) = \frac{1}{n}.$$

Consider $X = X_1 + \cdots + X_n$. Then X is the number of people who select their own hats. Hence,

$$\mathbb{E}[X] = \mathbb{E}\left[X_1 + \cdots + X_n\right] = n \cdot \frac{1}{n} = 1.$$

Notice that the random variables $X_1, ..., X_n$ are dependent. Indeed, knowing that the random variable $X_1 = \cdots = X_{n-1} = 1$ determines that all people except n select their own hats, which then changes the probability that the random variable X_n takes a value of 1 (i.e., n selects her own hat) to 1, from $\frac{1}{n}$ in the case when $X_1, ..., X_{n-1}$ are not known.

In Section 4.4.1, we calculated the expectation of a binomial random variable X with parameter (n, p): $\mathbb{E}[X] = np$. In the following, we establish this fact by considering X as a sum of iid binomial random variables with parameter $(1, p)$.

PROPOSITION 6.7 (Expectation of binomial random variables). *Let X be a binomial random variable X with parameter (n, p). Then the expectation $\mathbb{E}[X] = np$.*

PROOF. Let $X_1, ..., X_n$ be iid binomial random variables with parameter $(1, p)$. That is, X_i denotes the number of heads that appear in one toss of a coin, which lands on heads with probability p. Then

$$\mathbb{E}\left[X_i\right] = 0 \cdot P\left(X_i = 0\right) + 1 \cdot P\left(X_i = 1\right) = p.$$

Then $X = X_1 + \cdots + X_n$, i.e., the binomial random variable with parameter (n, p). Hence,

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

□

Example. We say that a changeover occurs in coin tosses if an outcome differs with the one preceding it. For instance, there are three changeovers in $\mathrm{h}, \mathrm{h}, \mathrm{t}, \mathrm{h}, \mathrm{t}$. We find the expectation of the number of changeovers in n independent tosses of a coin, which lands on heads with probability p. To this end, notice that sample space is

$$S = \left\{(s_1, ..., s_n) : s_1, ..., s_n \in \{\mathrm{h}, \mathrm{t}\}\right\}.$$

For $i = 1, ..., n-1$, define

$$X_i = \begin{cases} 1 & \text{if } s_i \neq s_{i+1}, \\ 0 & \text{if } s_i = s_{i+1}. \end{cases}$$

That is, $X_i = 1$ if a changeover occurs from the i-th toss to the following one and $X_i = 0$ otherwise. Then

$$\begin{aligned} \mathbb{E}\left[X_i\right] &= 0 \cdot P\left(s_i = \mathrm{h}, s_{i+1} = \mathrm{h}\right) + 0 \cdot P\left(s_i = \mathrm{t}, s_{i+1} = \mathrm{t}\right) \\ &\quad + 1 \cdot P\left(s_i = \mathrm{h}, s_{i+1} = \mathrm{t}\right) + 1 \cdot P\left(s_i = \mathrm{t}, s_{i+1} = \mathrm{h}\right) \\ &= 2p(1-p). \end{aligned}$$

Consider $X = X_1 + \cdots + X_{n-1}$. Then X denotes the number of changeovers. Hence,

$$\mathbb{E}[X] = \mathbb{E}\left[X_1 + \cdots + X_{n-1}\right] = 2(n-1)p(1-p).$$

Remark. We shall point out that that

$$\mathbb{E}\left[X_1 + \cdots + X_n\right] = \mathbb{E}\left[X_1\right] + \cdots + \mathbb{E}\left[X_n\right]$$

is always valid, and does *not* require that the random variables $X_1, ..., X_n$ are independent. This is essentially because taking expectation commutes with taking sum (a linear operation). Since taking expectation does not commute with nonlinear operations, $\mathbb{E}[XY]$ may be different from $\mathbb{E}[X]\mathbb{E}[Y]$.

For example, consider one toss of a fair coin. Let $X = 1$ if it lands on heads and $X = -1$ if it lands on tails. Let $Y = -1$ if it lands on heads and $Y = 1$ if it lands on tails. Clearly, X and Y are dependent random variables. Compute that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. But $\mathbb{E}[XY] = 1$ since $XY = 1$ is a constant random variable.

However, the following proposition states that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ holds for independent random variables X and Y.

PROPOSITION 6.8. *Let $X, Y : S \to \mathbb{R}$ be independent random variables. Then*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

More generally, if $X_1, ..., X_n : S \to \mathbb{R}$ are independent random variables, then

$$\mathbb{E}\left[X_1 \cdots X_n\right] = \mathbb{E}\left[X_1\right] \cdots \mathbb{E}\left[X_n\right].$$

PROOF. Since X, Y are independent, the joint probability mass function $p(X, Y) = p_X(x)p_Y(y)$ for all x, y. Here, p_X and p_Y are the probability mass functions of X and of Y, respectively. Hence,

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x,y} xy \cdot p(x,y) \\
&= \sum_{x,y} xy \cdot p_X(x)p_Y(y) \\
&= \left(\sum_x x \cdot p_X(x)\right)\left(\sum_y y \cdot p_Y(y)\right) \\
&= \mathbb{E}[X]\mathbb{E}[Y].
\end{aligned}$$

□

6.4. Variance and covariance

In Section 4.3, we defined the variance of a random variable X: Write $\mathbb{E}[X] = \mu$. Then

$$\mathrm{Var}[X] = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left[X^2\right] - \mu^2,$$

in which $\mathbb{E}[X^2]$ is the second moment.

We established several properties of expectation, for example, $\mathrm{Var}[X] \geq 0$ and equals 0 iff $X = \mu$ almost surely. Moreover, under the linear operations, we have that

$$\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X].$$

In particular, if the standard deviation $\sigma = \sqrt{\mathrm{Var}[X]} > 0$, then

$$\mathrm{Var}\left[\frac{X}{\sigma}\right] = 1.$$

In this section, we investigate the (co-)variance involving multiple random variables.

Definition (Covariance). Let $X, Y : S \to \mathbb{R}$ be random variables. The covariance of X and Y is defined as

$$\mathrm{Cov}[X, Y] = \mathbb{E}[(X-\mu)(Y-\nu)],$$

in which $\mu = \mathbb{E}[X]$ and $\nu = \mathbb{E}[Y]$.

Remark. Compute that

$$\begin{aligned}
\mathrm{Cov}[X,Y] &= \mathbb{E}[(X-\mu)(Y-\nu)] \\
&= \mathbb{E}[XY - \mu Y - \nu Y + \mu\nu] \\
&= \mathbb{E}[XY] - \mu\mathbb{E}[Y] - \nu\mathbb{E}[X] + \mu\nu \\
&= \mathbb{E}[XY] - \mu\nu - \nu\mu + \mu\nu \\
&= \mathbb{E}[XY] - \mu\nu \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].
\end{aligned}$$

PROPOSITION 6.9. *Let $X, Y : S \to \mathbb{R}$ be random variables.*

(i). $\mathrm{Cov}[Y, X] = \mathrm{Cov}[X, Y]$.

(ii). $\mathrm{Cov}[X, X] = \mathrm{Var}[X]$.
(iii). *If X and Y are independent, then* $\mathrm{Cov}[X, Y] = 0$.
(iv). *Let $a \in \mathbb{R}$. Then* $\mathrm{Cov}[aX, Y] = a\mathrm{Cov}[X, Y]$. *In particular,* $\mathrm{Cov}[-X, X] = -\mathrm{Var}[X]$.

Proof.

(i). Obvious.
(ii). Denote $\mu = \mathbb{E}[X]$. Then

$$\mathrm{Cov}[X, X] = \mathbb{E}[(X-\mu)(X-\mu)] = \mathbb{E}\left[(X-\mu)^2\right] = \mathrm{Var}[X].$$

(iii). $\mathrm{Cov}[Y, X] = \mathrm{Cov}[X, Y]$.
(iv). If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ by Proposition 6.8. Hence,

$$\mathrm{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

(v). Compute that

$$\mathrm{Cov}[aX, Y] = \mathbb{E}[aXY] - \mathbb{E}[aX]\mathbb{E}[Y] = a(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) = a\mathrm{Cov}[X, Y].$$

□

Definition (Correlated and uncorrelated random variables). Let $X, Y : S \to \mathbb{R}$ be random variables.

- If $\mathrm{Cov}[X, Y] > 0$, then X and Y are said to be positively correlated.
- If $\mathrm{Cov}[X, Y] = 0$, then X and Y are said to be uncorrelated.
- If $\mathrm{Cov}[X, Y] < 0$, then X and Y are said to be negatively correlated.

Remark. If $\mathrm{Cov}[X, Y] \neq 0$, then the random variables X and Y are said to be correlated. The sign of $\mathrm{Cov}[X, Y]$ then indicates the direction of their correlation.

In the case when $\mathrm{Cov}[X, Y] > 0$, if X increases, then Y is likely to increase.

In the case when $\mathrm{Cov}[X, Y] < 0$, if X increases, then Y is likely to decrease.

Remark (Correlation and independence). As a consequence of (iii) above, if X and Y are independent, then $\mathrm{Cov}[X, Y] = 0$, i.e., X and Y are uncorrelated. However, $\mathrm{Cov}[X, Y] = 0$ does *not* imply that X and Y are independent. For example, let X be a random variable that

$$P(X = 0) = P(X = -1) = P(X = 1) = \frac{1}{3}.$$

Let $Y = 1$ if $X \neq 0$ and $Y = 1$ if $X = 0$. Since $XY = 0$ and $\mathbb{E}[X] = 0$,

$$\mathrm{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

But clearly, X and Y are dependent.

Definition (Correlation coefficients). Let $X, Y : S \to \mathbb{R}$ be random variables. The correlation coefficient of X and Y is defined as

$$\rho[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \sigma_Y},$$

if the standard deviations $\sigma_X = \sqrt{\mathrm{Var}[X]} > 0$ and $\sigma_Y = \sqrt{\mathrm{Var}[Y]} > 0$.

Remark. The correlation coefficient of two random variables X and Y, as the correlation after normalization by the standard deviations of X and of Y, is always bounded between -1 and 1, see Theorem 6.11. It quantifies the degree of the correlation of X and Y. For example,

$$\mathrm{Cov}[X, X] = \frac{\mathrm{Cov}[X, X]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[X]}} = 1 \quad \text{and} \quad \mathrm{Cov}[-X, X] = \frac{\mathrm{Cov}[-X, X]}{\sqrt{\mathrm{Var}[-X]\mathrm{Var}[X]}} = -1.$$

This indicates that X is positively correlated with itself and is negatively correlated with $-X$ (unless X is constant almost surely so $\mathrm{Var}[X] = 0$), both of which are at the maximal degree.

Example. Consider three independent tosses of a fair coin. Let X denote the number of heads that appear and Y denote the number of tails that appear. Then X and Y are two random variables. The joint probability mass function p of X and Y is given by

(X,Y)	0	1	2	3
0	0	0	0	$\frac{1}{8}$
1	0	0	$\frac{3}{8}$	0
2	0	$\frac{3}{8}$	0	0
3	$\frac{1}{8}$	0	0	0

(a). Compute that

$$\mathbb{E}[XY] = \sum_{xy\neq 0} xy \cdot P(X=x, Y=y) = 2\cdot p(1,2) + 2\cdot p(2,1) = \frac{3}{2}.$$

(b). Compute that

$$\mathbb{E}[X] = \frac{3}{2},\ \mathrm{Var}[X] = \frac{3}{4} \quad \text{and} \quad \mathbb{E}[Y] = \frac{3}{2},\ \mathrm{Var}[Y] = \frac{3}{4}.$$

(c). Compute that

$$\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = -\frac{3}{4},$$

and

$$\rho[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} = -1.$$

This indicates that X are Y are negatively correlated to the maximal degree. Indeed, since $Y = 3 - X$, if X increases, then Y must decrease.

Example. Consider three independent tosses of a fair coin. Let X denote the number of heads that appear in the first two tosses and Y denote the number of heads that appear in three tosses. Then X and Y are two random variables. The joint probability mass function p of X and Y is given by

(X,Y)	0	1	2	3
0	$\frac{1}{8}$	$\frac{1}{8}$	0	0
1	0	$\frac{2}{8}$	$\frac{2}{8}$	0
2	0	0	$\frac{1}{8}$	$\frac{1}{8}$

(a). Compute that

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{xy\neq 0} xy \cdot P(X=x, Y=y) \\ &= 1\cdot p(1,1) + 2\cdot p(1,2) + 4\cdot p(2,2) + 6\cdot p(2,3) \\ &= 2.\end{aligned}$$

(b). Compute that

$$\mathbb{E}[X] = 1,\ \mathrm{Var}[X] = \frac{1}{2} \quad \text{and} \quad \mathbb{E}[Y] = \frac{3}{2},\ \mathrm{Var}[Y] = \frac{3}{4}.$$

(c). Compute that

$$\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{1}{2},$$

and

$$\rho[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} = \sqrt{\frac{2}{3}} \approx 0.816.$$

This indicates that X are Y are positively correlated, which means that if X increases, then Y is likely to increase.

We establish an important lemma, which leads several important consequences.

LEMMA 6.10. *Let $X_1, ..., X_n, Y_1, ..., Y_m : S \to \mathbb{R}$ be random variables. Set $X = X_1 + \cdots + X_n$ and $Y = Y_1 + \cdots + Y_m$. Then*

$$\mathrm{Cov}[X, Y] = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{Cov}\left[X_i, Y_j\right].$$

PROOF. Write $\mu_i = \mathbb{E}[X_i]$, $i = 1, ..., n$, and $\mathbb{E}[Y_j]$, $j = 1, ..., m$. Then by Theorem 6.6,

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mu_i \quad \text{and} \quad \mathbb{E}\left[\sum_{j=1}^{m} Y_j\right] = \sum_{j=1}^{m} \nu_j.$$

Compute that

$$\begin{aligned}
\mathrm{Cov}[X, Y] &= \mathrm{Cov}\left[\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)\left(\sum_{j=1}^{m} Y_j - \mathbb{E}\left[\sum_{j=1}^{m} Y_j\right]\right)\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} (X_i - \mu_i)\right)\left(\sum_{j=1}^{m} (Y_j - \nu_j)\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - \mu_i)(Y_j - \nu_j)\right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{E}\left[(X_i - \mu_i)(Y_j - \nu_j)\right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{Cov}\left[X_i, Y_j\right].
\end{aligned}$$

□

THEOREM 6.11. *Let $X, Y : S \to \mathbb{R}$ be random variables. Suppose that the standard deviations of X and of Y, respectively, $\sigma_X, \sigma_Y > 0$. Then*

$$-\sigma_X \sigma_Y \leq \mathrm{Cov}[X, Y] \leq \sigma_X \sigma_Y.$$

As a consequence,

$$-1 \leq \rho[X, Y] \leq 1.$$

PROOF. By Lemma 6.10, we compute that

$$\begin{aligned}
0 &\leq \mathrm{Var}\left[\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right] \\
&= \mathrm{Cov}\left[\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}, \frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right] \\
&= \mathrm{Cov}\left[\frac{X}{\sigma_X}, \frac{X}{\sigma_X}\right] + \mathrm{Cov}\left[\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right] + \mathrm{Cov}\left[\frac{Y}{\sigma_Y}, \frac{X}{\sigma_X}\right] + \mathrm{Cov}\left[\frac{Y}{\sigma_Y}, \frac{Y}{\sigma_Y}\right] \\
&= \mathrm{Var}\left[\frac{X}{\sigma_X}\right] + 2\mathrm{Cov}\left[\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right] + \mathrm{Var}\left[\frac{Y}{\sigma_Y}\right] \\
&= 2 + 2 \cdot \frac{\mathrm{Cov}[X, Y]}{\sigma_X \sigma_Y},
\end{aligned}$$

which implies that

$$\mathrm{Cov}[X,Y] \geq -\sigma_X \sigma_Y.$$

See Question 6.7 for a similar proof that

$$\mathrm{Cov}[X,Y] \leq \sigma_X \sigma_Y.$$

□

Example. Consider 100 independent tosses of a fair coin. Let X_i be the number of heads that appear in the i-th toss, $i = 1, ..., 100$. Then X_i are independent, which implies that

$$\mathrm{Cov}\left[X_i, X_j\right] = \begin{cases} 0 & \text{if } i \neq j, \\ \mathrm{Var}\left[X_i\right] = \frac{1}{4} & \text{if } i = j. \end{cases}$$

Let X be the number of heads that appear in the first 60 tosses and Y be the number of heads that appear in the last 60 tosses. Then

$$\mathbb{E}[X] = 30,\ \mathrm{Var}[X] = 15 \quad \text{and} \quad \mathbb{E}[Y] = 30,\ \mathrm{Var}[Y] = 15.$$

Compute that

$$\mathrm{Cov}[X,Y] = \sum_{i=1}^{60} \sum_{j=41}^{100} \mathrm{Cov}\left[X_i, X_j\right] = \sum_{i=41}^{60} \mathrm{Cov}\left[X_i, X_i\right] = 20 \cdot \frac{1}{4} = 5,$$

and

$$\rho[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} = \frac{1}{3}.$$

Example. Consider 100 independent tosses of a fair coin. Let X be the number of heads that appear in the first 51 tosses and Y be the number of heads that appear in the last 51 tosses. Then

$$\mathbb{E}[X] = \frac{51}{2},\ \mathrm{Var}[X] = \frac{51}{4} \quad \text{and} \quad \mathbb{E}[Y] = \frac{51}{2},\ \mathrm{Var}[Y] = \frac{51}{4}.$$

Using the same notation as the previous example, compute that

$$\mathrm{Cov}[X,Y] = \sum_{i=1}^{51} \sum_{j=49}^{100} \mathrm{Cov}\left[X_i, X_j\right] = \sum_{i=49}^{51} \mathrm{Cov}\left[X_i, X_i\right] = 3 \cdot \frac{1}{4} = \frac{3}{4},$$

and

$$\rho[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} = \frac{3}{51} \approx 0.059.$$

We now derive the variance of the sum $X = X_1 + \cdots + X_n$ through the covariances among the random variables $X_1, ..., X_n$.

THEOREM 6.12. *Let $X_1, ..., X_n : S \to \mathbb{R}$ be random variables and $X = X_1 + \cdots + X_n$. Then*

$$\mathrm{Var}[X] = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}\left[X_i, X_j\right].$$

As a consequence, if $X_1, ..., X_n$ are pairwise uncorrelated, then

$$\mathrm{Var}[X] = \sum_{i=j} \mathrm{Cov}\left[X_i, X_j\right] = \sum_{i=1}^{n} \mathrm{Cov}\left[X_i, X_i\right] = \sum_{i=1}^{n} \mathrm{Var}\left[X_i\right].$$

PROOF. By Lemma 6.10, we compute that

$$\mathrm{Var}[X] = \mathrm{Cov}[X,X] = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}\left[X_i, X_j\right].$$

□

In Section 4.4.1, we calculated the variance of a binomial random variable X with parameter (n, p): $\mathrm{Var}[X] = np(1-p)$. In the following, we establish this fact by considering X as a sum of iid binomial random variables with parameter $(1, p)$. See Question 6.8.

PROPOSITION 6.13 (Variance of binomial random variables). *Let X be a binomial random variable X with parameter (n, p). Then the variance* $\mathrm{Var}[X] = np(1-p)$.

Homework Assignment.

Question 6.1. Consider two independent rolls of a fair die. Let X denote the larger number and Y denote the smaller number.

(a). Find the joint probability mass function of X and Y.
(b). Find $p_{X|Y}(1|1)$ and $p_{X|Y}(6|6)$.
(c). Find $p_{Y|X}(1|1)$ and $p_{Y|X}(6|6)$.
(d). Determine whether X and Y are independent.

Answer. Notice that X and Y are random variables on the sample space

$$S = \{(x_1, x_2) : x_1, x_2 \in \{1, 2, 3, 4, 5, 6\}\},$$

which has size 36.

(a). The joint probability mass function of X and Y is given by

(X,Y)	1	2	3	4	5	6
1	$\frac{1}{36}$	0	0	0	0	0
2	$\frac{2}{36}$	$\frac{1}{36}$	0	0	0	0
3	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	0	0	0
4	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	0	0
5	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	0
6	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The sum in each row provides the probability mass function p_X:

$$p_X(1) = \frac{1}{36},\ p_X(2) = \frac{3}{36},\ p_X(3) = \frac{5}{36},\ p_X(4) = \frac{7}{36},\ p_X(5) = \frac{9}{36},\ p_X(6) = \frac{11}{36},$$

which then sum to 1.

The sum in each column provides the probability mass function p_Y:

$$p_Y(1) = \frac{11}{36},\ p_Y(2) = \frac{9}{36},\ p_Y(3) = \frac{7}{36},\ p_Y(4) = \frac{5}{36},\ p_Y(5) = \frac{3}{36},\ p_Y(6) = \frac{1}{36}$$

which also sum to 1.

(b). Compute that

$$p_{X|Y}(1|1) = \frac{1}{11} \quad \text{and} \quad p_{X|Y}(6|6) = 1.$$

(c). Compute that

$$p_{Y|X}(1|1) = 1 \quad \text{and} \quad p_{Y|X}(6|6) = \frac{1}{11}.$$

(d). Notice that

$$p(1,1) \neq p_X(1)p_Y(1).$$

Therefore, X and Y are dependent. Indeed, knowing that the random variable X takes a value x (i.e., the larger number) changes the probability that Y takes a value (i.e., the smaller number), in fact, Y can only take a value smaller than x, among all six possible values in the case when X is not known.

Question 6.2. Suppose that the joint probability mass function p of X and Y is given by

$$p(1,1) = \frac{1}{2}, \quad p(1,2) = \frac{1}{4}, \quad p(2,1) = \frac{1}{8}, \quad p(2,2) = \frac{1}{8}.$$

(a). Find the conditional mass function of X given $Y = i$, $i = 1, 2$.
(b). Determine whether X and Y are independent.
(c). Find $P(XY \leq 3)$, $P(X + Y > 2)$, $P(\frac{X}{Y} > 1)$.

Answer. Compute the probability mass functions p_X and p_Y, respectively:

$$p_X(1) = p(1,1) + p(1,2) = \frac{3}{4} \quad \text{and} \quad p_X(2) = p(2,1) + p(2,2) = \frac{1}{4}.$$
$$p_Y(1) = p(1,1) + p(2,1) = \frac{5}{8} \quad \text{and} \quad p_Y(2) = p(1,2) + p(2,2) = \frac{3}{8}.$$

(a). Compute that

$$P_{X|Y}(1|1) = \frac{p(1,1)}{p_Y(1)} = \frac{4}{5} \quad \text{and} \quad P_{X|Y}(2|1) = \frac{p(2,1)}{p_Y(1)} = \frac{1}{5}.$$
$$P_{X|Y}(1|2) = \frac{p(1,2)}{p_Y(2)} = \frac{2}{3} \quad \text{and} \quad P_{X|Y}(2|1) = \frac{p(2,1)}{p_Y(2)} = \frac{1}{3}.$$

(b). Since

$$p(1,1) \neq p_X(1)p_Y(1),$$

X and Y are dependent.
(c). Compute that

$$P(XY \leq 3) = p(1,1) + p(1,2) + p(2,1) = \frac{7}{8},$$
$$P(X + Y > 2) = p(1,2) + p(2,1) + p(2,2) = \frac{1}{2},$$
$$P\left(\frac{X}{Y} > 1\right) = p(2,1) = \frac{1}{8}.$$

Question 6.3. Let X be a Poisson random variable with parameters λ_1. Let Y be a Poisson random variable with parameters λ_2. Suppose that X and Y are independent. Show that $Z = X + Y$ is a Poisson random variable with parameter $\lambda_1 + \lambda_2$.

Answer. Compute the probability mass function p_Z of Z for $0 \leq r \leq n + m$:

$$\begin{aligned}
p_Z(r) &= P(X + Y = r) \\
&= \sum_{i=0}^{r} P(X = i, Y = r - i) \\
&= \sum_{i=0}^{r} P(X = i)P(Y = r - i) \\
&= \sum_{i=0}^{r} p_X(i)p_Y(r - i) \\
&= \sum_{i=0}^{r} e^{-\lambda_1} \frac{\lambda_1^i}{i!} e^{-\lambda_2} \frac{\lambda_2^{r-i}}{(r-i)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{r!} \sum_{i=0}^{r} \frac{r!}{i!(r-i)!} \lambda_1^i \lambda_2^{r-i} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^r}{r!},
\end{aligned}$$

in which we used the binomial theorem in Theorem 1.4 that

$$\sum_{i=0}^{r} \frac{r!}{i!(r-i)!} \lambda_1^i \lambda_2^{r-i} = (\lambda_1 + \lambda_2)^r.$$

Question 6.4. Consider n independent rolls of a fair die. Find the expectation of the sum of the numbers that appear.

Answer. Let X_i denote the number that appear in the i-th roll. Then X_i is a random variable whose expectation

$$\begin{aligned}\mathbb{E}[X_i] &= 1\cdot P(X_i=1)+2\cdot P(X_i=2)+3\cdot P(X_i=3)\\ &\quad +4\cdot P(X_i=4)+5\cdot P(X_i=5)+6\cdot P(X_i=6)\\ &= \frac{7}{2}.\end{aligned}$$

Consider $X = X_1 + \cdots + X_n$. Then X is the sum of the numbers appear. Hence,

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = \frac{7n}{2}.$$

Question 6.5. Suppose that n people throw their hat and each person randomly selects one. We say that i and j for $i \neq j$ are a matched pair if i selects the hat belonging to j and j selects the hat belonging to i. Find the expectation of the number of the matched pairs.

Answer. The number of pairs of people is given by

$$\binom{n}{2} = \frac{n(n-1)}{2}.$$

For each pair of i and j, define

$$X_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ is a matched pair,}\\ 0 & \text{otherwise.}\end{cases}$$

Then

$$\mathbb{E}[X_{ij}] = 0\cdot P(X_{ij}=0) + 1\cdot P(X_{ij}=1) = \frac{1}{n(n-1)}.$$

Consider

$$X = \sum_{\text{pairs } i \text{ and } j} X_{ij}.$$

Then X is the number of matched pairs. Hence,

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{\text{pairs } i \text{ and } j} X_{ij}\right] = \frac{n(n-1)}{2}\cdot\frac{1}{n(n-1)} = \frac{1}{2}.$$

Question 6.6. Consider three independent tosses of a fair coin. Let X denote the number of heads that appear in the first two tosses and Y denote the number of heads that appear in the last two tosses. Find the correlation coefficient of X and Y.

Answer. The joint probability mass function p of X and Y is given by

(X,Y)	0	1	2
0	$\frac{1}{8}$	$\frac{1}{8}$	0
1	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$
2	0	$\frac{1}{8}$	$\frac{1}{8}$

(a). Compute that

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{xy\neq 0} xy\cdot P(X=x, Y=y)\\ &= 1\cdot p(1,1)+2\cdot p(1,2)+2\cdot p(2,1)+4\cdot p(2,2)\\ &= \frac{5}{4}.\end{aligned}$$

(b). Compute that

$$\mathbb{E}[X] = 1,\ \mathrm{Var}[X] = \frac{1}{2} \quad \text{and} \quad \mathbb{E}[Y] = 1,\ \mathrm{Var}[Y] = \frac{1}{2}.$$

(c). Compute that

$$\mathrm{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{1}{4} > 0,$$

and

$$\rho[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} = \frac{1}{2}.$$

Question 6.7. Let $X, Y : S \to \mathbb{R}$ be random variables. Suppose that the standard deviations of X and of Y, respectively, $\sigma_X, \sigma_Y > 0$. Show that

$$\mathrm{Cov}[X, Y] \leq \sigma_X \sigma_Y.$$

Proof. Compute that

$$\begin{aligned}
0 &\leq \mathrm{Var}\left[\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right] \\
&= \mathrm{Cov}\left[\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}, \frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right] \\
&= \mathrm{Cov}\left[\frac{X}{\sigma_X}, \frac{X}{\sigma_X}\right] - \mathrm{Cov}\left[\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right] - \mathrm{Cov}\left[\frac{Y}{\sigma_Y}, \frac{X}{\sigma_X}\right] + \mathrm{Cov}\left[\frac{Y}{\sigma_Y}, \frac{Y}{\sigma_Y}\right] \\
&= \mathrm{Var}\left[\frac{X}{\sigma_X}\right] - 2\mathrm{Cov}\left[\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right] + \mathrm{Var}\left[\frac{Y}{\sigma_Y}\right] \\
&= 2 - 2 \cdot \frac{\mathrm{Cov}[X, Y]}{\sigma_X \sigma_Y},
\end{aligned}$$

which implies that

$$\mathrm{Cov}[X, Y] \leq \sigma_X \sigma_Y.$$

□

Question 6.8. Let X be a binomial random variable X with parameter (n, p). Show that the variance $\mathrm{Var}[X] = np(1-p)$.

Answer. Let $X_1, ..., X_n$ be iid binomial random variables with parameter $(1, p)$. That is, X_i denotes the number of heads that appear in one toss of a coin, which lands on heads with probability p. Then $\mathbb{E}[X_i] = p$ amd

$$\mathbb{E}\left[X_i^2\right] = 0^2 \cdot P\left(X_i = 0\right) + 1^2 \cdot P\left(X_i = 1\right) = p.$$

Hence,

$$\mathrm{Var}\left[X_i\right] = \mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[X_i\right]^2 = p - p^2 = p(1-p).$$

Then $X = X_1 + \cdots + X_n$, i.e., the binomial random variable with parameter (n, p). Hence,

$$\mathrm{Var}[X] = \mathrm{Var}\left[X_1\right] + \cdots + \mathrm{Var}\left[X_1\right] = np(1-p).$$

Question 6.9. Let X_1, X_2, X_3, X_4 be iid random variables with expectation 1 and variance 4.
(a). Find $\rho[X_1 + X_2, X_2 + X_3]$.
(b). Find $\rho[X_1 + X_2, X_3 + X_4]$.

Answer. Since X_1, X_2, X_3, X_4 are independent, if $i \neq j$, then

$$\mathrm{Cov}\left[X_i, X_j\right] = 0 \quad \text{and} \quad \mathrm{Var}\left[X_i + X_j\right] = \mathrm{Var}\left[X_i\right] + \mathrm{Var}\left[X_j\right] = 8.$$

(a). Compute that

$$\begin{aligned} & \operatorname{Cov}[X_1 + X_2, X_2 + X_3] \\ = \; & \operatorname{Cov}[X_1, X_2] + \operatorname{Cov}[X_1, X_3] + \operatorname{Cov}[X_2, X_2] + \operatorname{Cov}[X_2, X_3] \\ = \; & \operatorname{Cov}[X_2, X_2] \\ = \; & \operatorname{Var}[X_2] \\ = \; & 4. \end{aligned}$$

Hence,

$$\rho[X_1 + X_2, X_2 + X_3] = \frac{\operatorname{Cov}[X_1 + X_2, X_2 + X_3]}{\sqrt{\operatorname{Var}[X_1 + X_2]}\sqrt{\operatorname{Var}[X_2 + X_3]}} = \frac{1}{2}.$$

(b). Compute that

$$\begin{aligned} & \operatorname{Cov}[X_1 + X_2, X_3 + X_4] \\ = \; & \operatorname{Cov}[X_1, X_3] + \operatorname{Cov}[X_1, X_4] + \operatorname{Cov}[X_2, X_3] + \operatorname{Cov}[X_2, X_4] \\ = \; & 0. \end{aligned}$$

Hence,

$$\rho[X_1 + X_2, X_3 + X_4] = 0.$$

CHAPTER 7

Limiting theorems

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

Francis Galton, 1889[i]

In this chapter, we present the most important theoretical results in probability: the laws of large numbers (LLN) and the central limit theorem (CLT). They are concerned with the phenomenon that certain uniformity appears in a sum (or average, i.e., the sum divided by the number of variables) of a large number of independent random variables.

That is, each individual random variable behaves on its own and takes values close to or deviated from its expectation. However, if we take their sum or average, then this new random variable behaves much more uniformly than the individual ones. In particular, the laws of large numbers assert that the average stays close to the expectation with overwhelming probability, that is, it does *not* deviate much (whereas each individual random variable certainly can). The central limit theorem, on the other hand, asserts that the sum converges to a normal random variable (whereas each individual random variable may not be normal at all.)

In fact, we have already encountered this phenomenon in previous chapters and we recall them in the first section. Our study in this chapter involves discrete random variables and their continuous counterparts, as well as multiple random variables. We therefore also review the basics on these topics in the probability theory.

Let S be a (discrete or continuous) probability space with a probability P throughout the chapter.

7.1. Review of Probability

7.1.1. Discrete random variables. Let $X : S \to \mathbb{R}$ be a discrete random variable with probability mass function p.

- For each value x, $p(x) = P(X = x)$ provides the probability that X takes the value x.
- The expectation of X is given by
$$\mu = \mathbb{E}[X] = \sum_{s \in S} X(s) \cdot P(\{s\}) = \sum_{x \in X(S)} x \cdot p(x).$$
- The variance of X is given by
$$\mathrm{Var}[X] = \mathbb{E}\left[(X - \mu)^2\right] = \mathbb{E}\left[X^2\right] - \mu^2,$$
in which $\sigma = \sqrt{\mathrm{Var}[X]}$ is the standard deviation of X.

[i]Francis Galton, *Natural inheritance*. (1889).

Example. Let X be a binomial random variable with parameter (n, p), that is,

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i} \quad \text{for } i = 0, ..., n.$$

Then $\mathbb{E}[X] = np$ and $\mathrm{Var}[X] = np(1-p)$.

7.1.2. Continuous random variables. Let $X : S \to \mathbb{R}$ be a continuous random variable with probability density function $f(x)$.

- For each interval $I \subset \mathbb{R}$,
$$P(X \in I) = \int_I f(x)\, dx.$$
- The probability distribution function of X is given by
$$F(a) = P\left(X \in (-\infty, a)\right) = P\left(X < a\right) = \int_{-\infty}^{a} f(x)\, dx.$$
- The expectation of X is given by
$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx.$$
- The variance of X is given by
$$\mathrm{Var}[X] = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left[X^2\right] - \mu^2.$$

Example. Let X be the standard normal random variable, that is, the probability density function is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \text{for all } x \in \mathbb{R},$$

and the probability distribution function is

$$\Phi(a) = \int_{-\infty}^{a} \phi(x)\, dx \quad \text{for all } a \in \mathbb{R}.$$

Then

$$\mathbb{E}[X] = 1 \quad \text{and} \quad \mathrm{Var}[X] = 1.$$

The important 68-95-99.7 rule states that

$$P(|X| < 1) \approx 0.68, \quad P(|X| < 2) \approx 0.95, \quad P(|X| < 3) \approx 0.997.$$

7.1.3. Multiple random variables. Let $X_1, ..., X_n : S \to \mathbb{R}$ be (discrete or continuous) random variables.

- The expectation
$$\mathbb{E}\left[X_1 + \cdots + X_n\right] = \mathbb{E}\left[X_1\right] + \cdots + \mathbb{E}\left[X_n\right].$$
- If $X_1, ..., X_n$ are independent, then the variance
$$\mathrm{Var}\left[X_1 + \cdots + X_n\right] = \mathrm{Var}\left[X_1\right] + \cdots + \mathrm{Var}\left[X_n\right].$$

7.1.4. Laws of large numbers. In Chapter 2, we defined the basic concepts of sample space and outcomes in probability. For example, consider the experiment of tossing a coin. Then the sample space $S = \{\mathrm{h}, \mathrm{t}\}$ is the set of two possible outcomes of heads (h) and tails (t). Suppose that the coin lands on heads with probability p. The concept of probability here should be understood as follows: If we toss the coin for a great many times (independently), then the percentage of tosses which land on heads in these experiments tends to p. That is, in, say, 100 tosses, it is much more likely to have, say, 50 ± 10 heads than the case out of such a range. This phenomenon reflects the law of large numbers.

Indeed, $X_1, ..., X_n$ are iid (independent and identically distributed) random variables with expectation $\mathbb{E}[X_i] = \mu$. Then the random variable

$$Y = \frac{X_1 + \cdots + X_n}{n},$$

that is, the average of $X_1, ..., X_n$, has expectation

$$\mathbb{E}[Y] = \mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n}\left(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]\right) = \mu.$$

The laws of large numbers state that as $n \to \infty$, Y converges to its expectation μ with overwhelming probability, in certain senses, thus the weak law and the strong law. Hence, as $n \to \infty$, it becomes more unlikely that the average value Y of $X_1, ..., X_n$ deviates from the expectation, though each individual random variable X_i certainly can.

Set X_i, $i = 1, ..., n$, as iid binomial random variables with parameter $(1, p)$. That is, X_i is the number of heads in the i-th toss in n independent toss of a coin, which lands on heads with probability p. Then the expectation of X_i is $\mathbb{E}[X_i] = p$. In this case,

$$Y = \frac{X_1 + \cdots + X_n}{n}$$

is exactly the percentage of tosses which land on heads. The expectation $\mathbb{E}[Y] = p$. According to the laws of large numbers, Y converges to p as $n \to \infty$ with overwhelming probability, which justifies our understanding of the concept of probability.

7.1.5. Central limit theorems. In Section 5.4, we mentioned the De Moivre-Laplace limit theorem without a proof, that is, let X be the binomial random variable with parameters (n, p). As $n \to \infty$, the distribution of X converges to the one of the normal distribution. This theorem belongs to the realm of central limit theorem.

Indeed, let X_i, $i = 1, ..., n$, be iid random variables with expectation μ and variance σ^2. Then

$$X = X_1 + \cdots + X_n$$

is a random variable with expectation $n\mu$ and variance $n\sigma^2$ (so the standard deviation is $\sigma\sqrt{n}$). The central limit theorem states that as $n \to \infty$, the distribution of X converges to one of the normal distribution with expectation $n\mu$ and variance $n\sigma^2$. Therefore, the normalized random variable

$$\frac{X - n\mu}{\sigma\sqrt{n}}$$

converges to the standard normal distribution: For each interval $I \subset \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\frac{X - n\mu}{\sigma\sqrt{n}} \in I\right) = \int_I \phi(x)\, dx.$$

The central limit theorem may appear surprising as it applies as long as the random variables $X_1, ..., X_n$ are independent and identical, regardless which type they are and whether they are discrete or continuous. So if one repeats any experiments independently for a great many times, the overall result must tend to be normal.

Set X_i, $i = 1, ..., n$, as iid binomial random variables with parameter $(1, p)$, which has expectation $\mu = p$ and variance $\sigma^2 = p(1 - p)$. Then $X = X_1 + \cdots + X_n$ is the binomial random variable with

parameter (n, p). In this case, the central limit theorem recovers the one of De Moivre-Laplace: For each interval $I \subset \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\frac{X - np}{\sqrt{np(1-p)}} \in I\right) = \int_I \phi(x)\, dx.$$

7.2. The weak law of large numbers

In this section, we prove the law of large numbers in its weak form. It follows from Chebyshev's inequality, which in turn is a consequence of Markov's inequality. These inequalities have their own independent interests. That is, consider a random variable X, discrete or continuous. Suppose that we know its expectation $\mathbb{E}[X] = \mu$ and variance $\mathrm{Var}[X] = \sigma^2$. If, say, we also know that X is a normal random variable, then the probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

from which all information concerning the distribution of the random variables can be derived.

However, and more commonly, when we do not know what the random variable is, and we want to extract information of the distribution of its values from the expectation and variance, then Markov's inequality and Chebyshev's inequality provide valuable (albeit can be imprecise) information.

THEOREM 7.1 (Markov). *Let X be a non-negative random variable, that is, $X(s) \geq 0$ for all $s \in S$. Then for all $a > 0$,*

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

PROOF. Suppose that X is a discrete random variable. Since $X(s) \geq 0$ for all $s \in S$,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{s\in S} X(s) \cdot P(\{s\}) \\
&\geq \sum_{X(s)\geq a} X(s) \cdot P(\{s\}) \\
&\geq \sum_{X(s)\geq a} a \cdot P(\{s\}) \\
&= a \sum_{X(s)\geq a} P(\{s\}) \\
&= aP(X \geq a),
\end{aligned}$$

which implies that

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Suppose that X is a continuous random variable with probability density function $f(x)$. Then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx \geq \int_a^{\infty} x \cdot f(x)\, dx \geq \int_a^{\infty} a \cdot f(x)\, dx = a\int_a^{\infty} f(x)\, dx = aP(X \geq a),$$

which implies Markov's inequality. □

Example. Suppose that the class of *Introduction to Probability* is consisted of 10 students. Equip this sample space S with a probability P such that each sample point has the same probability (of $\frac{1}{10}$). Assume that the scores of Test 1 are $33, 17, 32, 38, 32, 23, 32, 38, 40, 23$. Let X denote the score a student obtained in Test 1. Then X is a random variable whose expectation $\mathbb{E}[X] = 30.8$.

(a). Set $a = 22$. Then

$$\frac{\mathbb{E}[X]}{a} = 1.4, \quad \text{while} \quad P(X \geq 30) = \frac{9}{10}.$$

(b). Set $a = 30.8$. Then

$$\frac{\mathbb{E}[X]}{a} = 1, \quad \text{while} \quad P(X \geq 30.8) = \frac{7}{10}.$$

(c). Set $a = 38$. Then

$$\frac{\mathbb{E}[X]}{a} \approx 0.811, \quad \text{while} \quad P(X \geq 38) = \frac{3}{10}.$$

Remark. Markov's theorem provide an easily accessible upper bound by $\frac{\mathbb{E}[X]}{a}$ of $P(X \geq a)$ (i.e., the probability that X takes larger values than a given number). However, this bound can be imprecise. Moreover, it is only meaningful when $a > \mathbb{E}[X]$, and when $a \leq \mathbb{E}[X]$, $\frac{\mathbb{E}[X]}{a} \geq 1$ and the estimate is trivially true.

Example. Suppose that the grades of 40 students have a mean of 60. Then by Markov's inequality, the number of grades which are ≥ 95 is bounded above by

$$\frac{60}{95} \cdot 40 \approx 25.3.$$

Therefore, the number of grades which are ≥ 95 is at most 25.

Example. Let X be the uniform random variable on $(0, 30)$, that is, the probability density function is

$$f(x) = \begin{cases} \frac{1}{30} & \text{if } 0 < x < 30, \\ 0 & \text{otherwise.} \end{cases}$$

Then the expectation $\mathbb{E}[X] = 15$.

(a). Set $a = 10$. Then

$$\frac{\mathbb{E}[X]}{a} = 1.5, \quad \text{while} \quad P(X \geq 10) = \frac{2}{3}.$$

(b). Set $a = 15$. Then

$$\frac{\mathbb{E}[X]}{a} = 1, \quad \text{while} \quad P(X \geq 15) = \frac{1}{2}.$$

(c). Set $a = 20$. Then

$$\frac{\mathbb{E}[X]}{a} = 0.75, \quad \text{while} \quad P(X \geq 20) = \frac{1}{3}.$$

THEOREM 7.2 (Chebyshev[i]). *Let X be a random variable with expectation $\mathbb{E}[X] = \mu$ and variance $\sigma^2 = \mathrm{Var}[X] > 0$. Then for all $d > 0$,*

$$P(|X - \mu| \geq d) \leq \frac{\sigma^2}{d^2}.$$

As a consequence, for all $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

PROOF. Let $Y = (X - \mu)^2$. Then Y is a non-negative random variable whose expectation

$$\mathbb{E}[Y] = \mathbb{E}\left[(X - \mu)^2\right] = \mathrm{Var}[X] = \sigma^2.$$

Take $a = d^2$ in Markov's inequality. Then

$$P(|X - \mu| \geq d) = P\left((X - \mu)^2 \geq d^2\right) = P\left(Y \geq d^2\right) \leq \frac{\mathbb{E}[Y]}{d^2} = \frac{\sigma^2}{d^2}.$$

□

[i]Pafnuty Chebyshev, *Des valeurs moyennes.* [The expectation values]. (1867).

Example. Suppose that the class of *Introduction to Probability* is consisted of 10 students. Equip this sample space S with a probability P such that each sample point has the same probability (of $\frac{1}{10}$). Assume that the scores of Test 1 are $33, 17, 32, 38, 32, 23, 32, 38, 40, 23$. Let X denote the score a student obtained in Test 1. Then X is a random variable whose expectation $\mu = \mathbb{E}[X] = 30.8$ and variance $\mathrm{Var}[X] = 50.96$ (so the standard deviation $\sigma \approx 7.139$).

(a). Set $k = 1$. Then $\frac{1}{k^2} = 1$, while

$$P(|X - \mu| \geq \sigma) \approx P(|X - 30.8| \geq 7.139) = P(X \geq 38) + P(X \leq 23) = \frac{6}{10}.$$

(b). Set $a = 1.5$. Then $\frac{1}{k^2} \approx 0.816$, while

$$P(|X - \mu| \geq 1.5\sigma) \approx P(|X - 30.8| \geq 10.709) = P(X \geq 42) + P(X \leq 20) = \frac{1}{10}.$$

(c). Set $a = 2$. Then $\frac{1}{k^2} = 0.25$, while

$$P(|X - \mu| \geq 2\sigma) \approx P(|X - 30.8| \geq 14.278) = P(X \geq 46) + P(X \leq 16) = 0.$$

Remark. Chebyshev's inequality provide an easily accessible upper bound by $\frac{1}{k^2}$ of $P(|X-\mu| \geq k\sigma)$ (i.e., the probability that X is deviated from the expectation by k standard deviations). However, this bound can be imprecise. Moreover, it is only meaningful when $k > 1$, and when $k \leq 1$, $\frac{1}{k^2} \geq 1$ and the estimate is trivially true.

Example. Let X be the uniform random variable on $(0, 30)$. Then the expectation $\mu = \mathbb{E}[X] = 15$ and variance $\mathrm{Var}[X] = 75$ (so the standard deviation $\sigma \approx 8.66$).

(a). Set $k = 1$. Then $\frac{1}{k^2} = 1$, while

$$P(|X - \mu| \geq \sigma) \approx P(|X - 15| \geq 8.66) = P(X \geq 23.66) + P(X \leq 6.34) = 0.423.$$

(b). Set $k = 1.5$. Then $\frac{1}{k^2} \approx 0.816$, while

$$P(|X - \mu| \geq 1.5\sigma) \approx P(|X - 15| \geq 12.99) = P(X \geq 27.99) + P(X \leq 2.01) = 0.134.$$

(c). Set $k = 2$. Then $\frac{1}{k^2} = 0.25$, while

$$P(|X - \mu| \geq 2\sigma) \approx P(|X - 15| \geq 17.32) = P(X \geq 32.32) + P(X \leq -2.32) = 0.$$

Remark. An equivalent formulation of Chebyshev's inequality is that

$$P(|X - \mu| < d) \geq 1 - \frac{\sigma^2}{d^2} \quad \text{and} \quad P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2},$$

that is, lower bounds of the probability that the random variable X takes value close to its expectation.

Example. Suppose that the grades of 40 students have a mean of 80 and a variance of 36 (so the standard deviation is 6). Then by Chebyshev's inequality, the number of grades which are in the range $(68, 92)$ (i.e., within two standard deviations from the mean) is bounded below by

$$\left(1 - \frac{1}{2^2}\right) \cdot 40 = 30.$$

Therefore, the number of grades in the range $(68, 92)$ is at least 30.

Example. Let X be the standard random variable. Then the expectation $\mu = \mathbb{E}[X] = 0$ and variance $\mathrm{Var}[X] = 1$ (so the standard deviation $\sigma = 1$).

(a). Set $k = 1$. Then $1 - \frac{1}{k^2} = 0$, while

$$P(|X - \mu| < \sigma) = P(|X| < 1) \approx 0.68.$$

(b). Set $k = 2$. Then $1 - \frac{1}{k^2} = 0.75$, while

$$P(|X - \mu| < 2\sigma) = P(|X| < 2) \approx 0.95.$$

(c). Set $k = 3$. Then $1 - \frac{1}{k^2} \approx 0.889$, while

$$P(|X - \mu| < 3\sigma) = P(|X| < 3) \approx 0.997.$$

Now the weak law of large numbers follows directly from Chebyshev's inequality:

THEOREM 7.3 (The weak law of large numbers). *Let X_i with $i \in \mathbb{N}$ be iid random variables with expectation μ and variance $\sigma^2 < \infty$. Then for any $\varepsilon > 0$,*

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \to 0 \quad \text{as } n \to \infty.$$

PROOF. Let

$$Y = \frac{X_1 + \cdots + X_n}{n}.$$

Then Y is a random variable whose expectation is given by

$$\mathbb{E}[Y] = \mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n}\left(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]\right) = \mu.$$

Since $X_1, ...X_n$ are independent, the variance of Y is given by

$$\mathrm{Var}[Y] = \mathrm{Var}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n^2}\mathrm{Var}\left[X_1 + \cdots + X_n\right] = \frac{1}{n^2}\left(\mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n]\right) = \frac{\sigma^2}{n}.$$

Then by Chebyshev's inequality,

$$P(|Y - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}[Y]}{\varepsilon^2} \leq \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad \text{as } n \to \infty.$$

□

Example. Consider n independent tosses of a fair coin. Let X_i denote the number of heads in the i-th toss, $i = 1, ..., n$, whose expectation $\mu = \frac{1}{2}$ and variance $\sigma^2 = \frac{1}{4}$. Then $X = X_1 + \cdots + X_n$ is the number of heads, whereas $Y = \frac{X}{n}$ is the percentage of tosses which land on heads.

According to the weak law of large numbers, for any $\varepsilon > 0$,

$$P\left(\left|Y - \frac{1}{2}\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}[Y]}{\varepsilon^2} \leq \frac{\sigma^2}{n\varepsilon^2} = \frac{1}{4n\varepsilon^2}.$$

For example, set $n = 100$ and $\varepsilon = \frac{1}{10}$. Then

$$P\left(\left|Y - \frac{1}{2}\right| \geq \frac{1}{10}\right) \leq \frac{1}{4n\varepsilon^2} = 0.25.$$

which implies that

$$P\left(\left|Y - \frac{1}{2}\right| \leq \frac{1}{10}\right) \geq 0.75.$$

That is, with at least probability 0.75, the percentage of tosses which land on heads is in the range $\frac{1}{2} \pm \frac{1}{10}$, i.e., the number of heads is in the range 50 ± 10.

7.3. The strong law of large numbers

In this section, we prove the strong law of large numbers.

THEOREM 7.4 (The strong law of large numbers). *Let X_i with $i \in \mathbb{N}$ be iid random variables with expectation μ and variance σ^2. Suppose that the fourth moment $\mathbb{E}[X_i^4] = K < \infty$.*

$$P\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

Remark (Weak and strong laws of large numbers). Let

$$Y = \frac{X_1 + \cdots + X_n}{n}.$$

Recall that the weak law of large numbers states that for any $\varepsilon > 0$,

$$P(|Y - \mu| \geq \varepsilon) \to 0 \quad \text{as } n \to \infty.$$

that is, the probability that Y deviates from the expectation μ by at least ε converges to 0.

Now the strong law of large numbers states that with probability 1, Y converges to μ, that is, Y does *not* deviate from its expectation asymptotically. Therefore, the strong law of large numbers implies the weak version.

Indeed, if not, then $P(|Y - \mu| \geq \varepsilon)$ does *not* converge to 0, i.e., there is a positive probability that Y deviates from μ by at least $\varepsilon > 0$. With at least the same probability, Y does *not* converge to μ, contradicting with the strong law.

The proof below of the strong law of large numbers uses the fourth moment $\mathbb{E}[X^4]$. We prove a relation with the second moment:

LEMMA 7.5. *Let X be a random variable with fourth moment $\mathbb{E}[X^4] = K$. Then*

$$\left(\mathbb{E}\left[X^2\right]\right)^2 \leq K.$$

PROOF. Notice that $\mathbb{E}[X^4] = \mathbb{E}[(X^2)^2]$ is the second moment of the random variable X^2. Hence,

$$K = \mathbb{E}\left[X^4\right] = \mathbb{E}\left[\left(X^2\right)^2\right] = \operatorname{Var}\left[X^2\right] + \left(\mathbb{E}\left[X^2\right]\right)^2,$$

which implies that

$$\left(\mathbb{E}\left[X^2\right]\right)^2 = K - \operatorname{Var}\left[X^2\right] \leq K,$$

since $\operatorname{Var}[X^2] \geq 0$. □

The proof below also uses some basics of series, for which we briefly recall here:

Remark (Series).

- If
$$\sum_{n=1}^{\infty} a_n$$
converges, then
$$\lim_{n\to\infty} a_n = 0.$$
- If $0 \leq a_n \leq b_n$ for all $n \in \mathbb{N}$ and
$$\sum_{n=1}^{\infty} b_n$$
converges, then
$$\sum_{n=1}^{\infty} a_n$$
also converges.
-
$$\sum_{n=1}^{\infty} \frac{1}{n^r} \begin{cases} \text{converges} & \text{if } r > 1, \\ \text{diverges} & \text{if } 0 < r \leq 1. \end{cases}$$

Now we provide the proof of the strong law of large numbers:

PROOF. It suffices to prove the theorem in the case when $\mu = \mathbb{E}[X_i] = 0$. Indeed, the case for general iid random variables Y_i with $i \in \mathbb{N}$ follows by setting $X_i = Y_i - \mathbb{E}[Y_i]$.

Write

$$X = \sum_{i=1}^{n} X_i \quad \text{and} \quad Y = \frac{X}{n}.$$

Then

$$X^4 = \left(\sum_{i=1}^{n} X_i\right)^4 = \sum_{i,j,k,l=1}^{n} X_i X_j X_k X_l,$$

whose terms can be classified into five groups:

(1). All four of i, j, k, l are identical. There are n such terms, e.g., $X_1X_1X_1X_1$. Each term has expectation equal to

$$\mathbb{E}\left[X_1X_1X_1X_1\right] = \mathbb{E}\left[X_1^4\right] = K.$$

(2). Three of i, j, k, l are identical, which differ with the remaining one, e.g., $X_1^3X_2$. Each term has expectation equal to

$$\mathbb{E}\left[X_1^3X_2\right] = \mathbb{E}\left[X_1^3\right]\mathbb{E}\left[X_2\right] = 0,$$

because $\mathbb{E}[X_1] = \mu = 0$.

(3). Two of i, j, k, l are identical, differing with the remaining two, which are also distinct, e.g., $X_1^2X_2X_3$. Each term has expectation equal to

$$\mathbb{E}\left[X_1^2X_2X_3\right] = \mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_2\right]\mathbb{E}\left[X_3\right] = 0.$$

(4). Two of i, j, k, l are identical, differing with the remaining two, which are identical, e.g., $X_1^2X_2^2$. There are

$$\binom{n}{2}\binom{4}{2} = 3n(n-1)$$

such terms. Each term has expectation equal to

$$\mathbb{E}\left[X_1^2X_2^2\right] = \mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_2^2\right] = \left(\mathbb{E}\left[X_1^2\right]\right)^2 \leq K,$$

using Lemma (7.5).

(5). All four of i, j, k, l are distinct, e.g., $X_1X_2X_3X_4$. Each term has expectation equal to

$$\mathbb{E}\left[X_1X_2X_3X_4\right] = \mathbb{E}\left[X_1\right]\mathbb{E}\left[X_2\right]\mathbb{E}\left[X_3\right]\mathbb{E}\left[X_4\right] = 0.$$

Hence,

$$\mathbb{E}\left[X^4\right] \leq nK + 3n(n-1)K,$$

which implies that

$$\mathbb{E}\left[\frac{X^4}{n^4}\right] \leq \frac{K}{n^3} + \frac{3(n-1)K}{n^3} \leq K\left(\frac{1}{n^3} + \frac{3}{n^2}\right).$$

Notice that

$$\sum_{n=1}^{\infty}\left(\frac{1}{n^3} + \frac{3}{n^2}\right) < \infty.$$

It then follows that

$$\mathbb{E}\left[\sum_{n=1}^{\infty}\frac{X^4}{n^4}\right] = \sum_{n=1}^{\infty}\mathbb{E}\left[\frac{X^4}{n^4}\right] < \infty, \tag{7.1}$$

which implies that with probability 1,

$$\sum_{n=1}^{\infty}\frac{X^4}{n^4} < \infty. \tag{7.2}$$

If not, then there is a positive probability such that

$$\sum_{n=1}^{\infty}\frac{X^4}{n^4} = \infty,$$

which would lead to conclusion that the expectation of this series is infinity, contradicting with (7.1).

Now (7.2) implies that with probability 1,

$$\lim_{n\to\infty} \frac{X^4}{n^4} = \lim_{n\to\infty} \left(\frac{X}{n}\right)^4 = \lim_{n\to\infty} Y^4 = 0,$$

that is, $Y \to 0$ as $n \to \infty$. □

7.4. The central limit theorem

In this section, we prove the central limit theorem.

THEOREM 7.6 (The central limit theorem). *Let X_i with $i \in \mathbb{N}$ be iid random variables with expectation μ and variance σ^2. Let $X = X_1 + \cdots + X_n$. Then for each interval $I \subset \mathbb{R}$,*

$$\lim_{n\to\infty} P\left(\frac{X - n\mu}{\sigma\sqrt{n}} \in I\right) = \int_I \phi(x)\,dx.$$

Thus, the central limit theorem provides a complete characterization of the limit of the random variable $\frac{X-n\mu}{\sigma\sqrt{n}}$. We prove the theorem using the characteristic functions of random variables.

Definition (Characteristic functions). Let X be a random variable. The characteristic function of X is defined as

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right].$$

- Suppose that X is a discrete random variable with probability mass function p. Then
$$\mathbb{E}\left[e^{itX}\right] = \sum_{x\in X(S)} e^{itx} \cdot p(x).$$
- Suppose that X is a continuous random variable with probability density function f. Then
$$\mathbb{E}\left[e^{itX}\right] = \int_{-\infty}^{\infty} e^{itx} \cdot f(x)\,dx.$$

Remark. In the discrete case, $\varphi_X(t)$ is the discrete Fourier transform of the probability mass functions; in the continuous case, the continuous Fourier transform of the probability density function.

In Fourier analysis, one can show that the characteristic function of a random variable X determines completely the probability mass or density function via an inverse Fourier transform. In particular, all information of X can be derived from its characteristic function, which include the expectation, the variance, and all the moments. See Lemma 7.7.

Moreover, to show that two random variables are identical, it suffices to show that their characteristic functions equal. This is our plan of proving the central limit theorem later. That is, to show that a random variable Z converges to the standard normal random variable, we prove that the characteristic function $\varphi_Z(t)$ converges to $e^{-\frac{1}{2}t^2}$, one of the standard normal random variable, see Proposition 7.8.

LEMMA 7.7. *Let X be a random variable with characteristic function $\varphi_X(t)$. Then for all $k \in \mathbb{N}$,*

$$\varphi_X^{(k)}(0) = i^k \mathbb{E}\left[X^k\right].$$

In particular,

$$\varphi_X(0) = 1, \quad \varphi_X'(0) = i\mathbb{E}[X], \quad \varphi_X''(0) = -\mathrm{Var}\left[X^k\right].$$

PROOF. Compute that

$$\varphi_X^{(k)}(t) = \mathbb{E}\left[(iX)^k e^{itX}\right] = i^k \mathbb{E}\left[X^k e^{itX}\right].$$

Hence,

$$\varphi_X^{(k)}(0) = i^k \mathbb{E}\left[X^k\right].$$

□

PROPOSITION 7.8. *Let X be the standard normal distribution. Then the characteristic function is*

$$\mathbb{E}\left[e^{itX}\right] = e^{-\frac{1}{2}t^2}.$$

PROOF. Compute that

$$\begin{aligned}
\mathbb{E}\left[e^{itX}\right] &= \int_{-\infty}^{\infty} e^{itx} \cdot \phi(x)\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} \cdot e^{-\frac{1}{2}x^2}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(x^2 - 2itx\right)}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(x^2 - 2itx + (it)^2 - (it)^2\right)}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-it)^2} e^{-\frac{1}{2}t^2}\, dx \\
&= e^{-\frac{1}{2}t^2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-it)^2}\, dx \\
&= e^{-\frac{1}{2}t^2}.
\end{aligned}$$

□

The proof below of the central limit theorem also uses some basics of Taylor series, for which we briefly recall here:

Remark (Taylor series). Suppose that $u(x) \in C^\infty(\mathbb{R})$. Then

$$u(x) = \sum_{k=0}^{\infty} \frac{u^{(k)}(0)}{k!} x^k.$$

which is the Taylor series at 0. In particular, if $f \in C^\infty(\mathbb{R})$, then as $x \to 0$,

$$u(x) = u(0) + u'(0)x + \frac{1}{2}u''(0)x^2 + o\left(x^2\right).$$

Now we provide the proof of the central limit theorem:

PROOF. It suffices to prove the theorem in the case when $\mu = \mathbb{E}[X_i] = 0$ and $\sigma^2 = \mathrm{Var}[X_i] = 1$. Indeed, the case for general iid random variables Y_i with $i \in \mathbb{N}$ follows by setting $X_i = \frac{Y_i - \mathbb{E}[Y_i]}{\sqrt{\mathrm{Var}[Y_i]}}$.

We then need to show that as $n \to \infty$,

$$Z = \frac{X}{\sqrt{n}} = \frac{X_1 + \cdots + X_n}{\sqrt{n}}$$

converges to the standard normal random variable. To this end, we show that the characteristic function

$$\varphi_Z(t) = \mathbb{E}\left[e^{itZ}\right] \to e^{-\frac{1}{2}t^2} \quad \text{as } n \to \infty,$$

in which $e^{-\frac{1}{2}t^2}$ is the characteristic function of the standard normal random variable by Proposition 7.8. Since $X_1, ..., X_n$ are independent, we compute that

$$\begin{aligned}
\mathbb{E}\left[e^{itZ}\right] &= \mathbb{E}\left[e^{\frac{it(X_1 + \cdots + X_n)}{\sqrt{n}}}\right] \\
&= \mathbb{E}\left[e^{\frac{itX_1}{\sqrt{n}}} \cdots e^{\frac{itX_n}{\sqrt{n}}}\right] \\
&= \mathbb{E}\left[e^{\frac{itX_1}{\sqrt{n}}}\right] \cdots \mathbb{E}\left[e^{\frac{itX_n}{\sqrt{n}}}\right]
\end{aligned}$$

$$= \left(\mathbb{E}\left[e^{\frac{itX_1}{\sqrt{n}}}\right]\right)^n$$

Consider the random variable $\frac{X_1}{\sqrt{n}}$:

$$\mathbb{E}\left[\frac{X_1}{\sqrt{n}}\right] = 0 \quad \text{and} \quad \text{Var}\left[\frac{X_1}{\sqrt{n}}\right] = \frac{1}{n}\text{Var}\left[X_1\right] = \frac{1}{n}.$$

Therefore, its characteristic function satisfies that

$$\varphi'_{X_1}(0) = \mathbb{E}\left[\frac{X_1}{\sqrt{n}}\right] = 0 \quad \text{and} \quad \varphi''_{X_1}(0) = -\text{Var}\left[\frac{X_1}{\sqrt{n}}\right] = -\frac{1}{n}.$$

Hence, the Taylor series of φ_{X_1} at 0 states that

$$\varphi_{X_1}(t) = 1 - \frac{t^2}{2n} + o\left(t^2\right).$$

Finally, as $n \to \infty$,

$$\mathbb{E}\left[e^{itZ}\right] = \left(\mathbb{E}\left[e^{\frac{itX_1}{\sqrt{n}}}\right]\right)^n = \left(1 - \frac{t^2}{2n} + o\left(t^2\right)\right)^n \to e^{-\frac{1}{2}t^2}.$$

Here, we used the fact that

$$\left(1 + \frac{1}{m}\right)^m \to e \quad \text{as } m \to \infty.$$

□

Homework Assignment.

Question 7.1. A manager distributes a total bonus of \$10,000 among 100 employees. Using Markov's inequality, find an upper bound of the number of employees who receive bonuses greater than or equal to \$500.

Answer. The mean of the bonuses is given by $\frac{10000}{100} = 100$. By Markov's inequality, the number of employees who receive bonuses greater than or equal to \$500 is bounded above by

$$\frac{100}{500} \cdot 100 = 20.$$

Therefore, the number of employees who receive bonuses greater than or equal to \$500 is at most 20.

Question 7.2. A manager distributes a total bonus of \$10,000 among 100 employees. Suppose that the variance of the bonuses is 100 (so the standard deviation is \$10). Using Chebyshev's inequality, find the a lower bound of the number of employees which receive bonuses in the range $(70, 130)$.

Answer. The mean of the bonuses is given by $\frac{10000}{100} = 100$. By Chebyshev's inequality, the number of employees who receive bonuses in the range $(70, 130)$ (i.e., within three standard deviations from the mean) is bounded below by

$$\left(1 - \frac{1}{3^2}\right) \cdot 100 \approx 88.9.$$

Therefore, the number of employees who receive bonuses in the range $(70, 130)$ is at least 89.

Question 7.3. Let $X_1, ..., X_9$ be independent Poisson random variables with expectation 1. Let $X = X_1 + \cdots + X_9$.

(a). Using Markov's inequality, find an upper bound of $P(X > 12)$.

(b). Using the central limit theorem, find an approximation of $P(X > 12)$.

Answer. A Poisson random variable with expectation $\mu = 1$ has variance $\sigma^2 = 1$. Hence, $X = X_1 + \cdots + X_9$ has expectation $\mathbb{E}[X] = 9$ and variance $\text{Var}[X] = 9$ (so the standard deviation is 3).

(a). Using Markov's inequality,

$$P(X > 12) \leq \frac{\mathbb{E}[X]}{12} = \frac{3}{4} = 0.75.$$

(b). Using the central limit theorem, the random variable

$$Y = \frac{X - 9}{3}$$

can be approximated by the standard normal random variable, that is, $X = 3Y + 9$. Hence,

$$P(X > 12) = P(3Y + 9 > 12) = P(Y > 1) = 1 - P(Y < 1) \approx 1 - 0.8413 = 0.1587,$$

by Table 5.2.

Question 7.4. A person has 100 light bulbs whose lifetimes are independent exponential random variables with mean 5 hours. If the bulbs are used one at a time, with a failed bulb being replaced immediately by a new one, approximate the probability that there is still a working bulb after 540 hours.

Answer. Let $X_1, ..., X_{100}$ be iid exponential random variables with expectation $\mu = 5$. Then its parameter is $\lambda = \frac{1}{5}$ and variance is $\sigma^2 = 25$. Then $X = X_1 + \cdots + X_n$ is a random variable with expectation 500 and variance 2500 (so the standard deviation is 50). Using the central limit theorem, the random variable

$$Y = \frac{X - 500}{50}$$

can be approximated by the standard normal random variable, that is, $X = 50Y + 500$. Hence,

$$P(X > 540) = P(50Y + 500 > 540) = P(Y > 0.8) = 1 - P(Y < 0.8) \approx 1 - 0.7881 = 0.2119,$$

by Table 5.2.