Probability Theory

Lecture notes, 2025 Fall

Introduction

In this paper, we discuss the english version of a hungarian lecture notes, which was witten by Szabolcs Mészáros. Additional contributors to the hungary version: Barbara Balázs, László Papp, Dániel Szabó, Balázs Takács, Dávid Tóth, András Tóbiás. Bence Csonka translated and wrote this lecture notes according to the above mentioned work.

Contents

1 Basic Concepts

The practical relevance of probability theory likely needs no justification for any reader: most experimental sciences rely on it in some form. Yet the question remains—why isn't the intuitive "favorable over total" ratio sufficient in most cases?

One reason is that naive approaches sometimes give incorrect or contradictory results. This is illustrated by the many probability paradoxes in the literature. Here's one such example:

If someone hears the name "probability theory", he suddenly thinks of the formula from the secondary school, 'favorable per total'. Why can't we use this formula for everything? The following example shows us that the naive approach could cause some contradictions.

There is a family with two children (they are not twins). What is the probability that the two children are all girls? We will see two different approaches to calculate this probability. Firstly, we have three outcomes: two boys; a boy and a girl; two girls. The number of favorable cases is 1 and the total number of cases is 3. We can easily see that the answer to the question is $\frac{1}{3}$. Now we examine another interpretation. We have the following outcomes: the two children are boys, the first child is a boy and the second child is a girl, the first child is a girl and the second child is a boy, the children are girls. Indeed, we have 4 outcomes and 1 favorable case. The answer to the question $\frac{1}{4}$.

Motivated by this example, we proceed to define the key concepts more rigorously.

1.1 Sample Space

The concept of probability is formalized using the Kolmogorov axioms.

Andrey Kolmogorov, a highly influential 20th-century mathematician, developed a framework to resolve ambiguities like the one above. This framework begins with what are now called the Kolmogorov axioms. The axioms themselves appear as part of the definition of a probability space.

Definition 1.1.1. *Let Ω be an arbitrary set. We define the following terms:*

- **Sample space:** Ω
- **Outcome:** *an element* $\omega \in \Omega$
- **Events:** *selected subsets* $A \subseteq \Omega$
- **Probability:** *a number* $\mathbb{P}(A) \in [0, 1]$ *assigned to an event* A

In the above paradox, there are 6 outcomes, so the sample space has 6 elements. The probability of the event $A = \{\text{draw a gold coin}\}$ is $P(A) = \frac{1}{2}$.

What does it mean that events are "selected" subsets? In short, we consider as events those subsets for which we want to assign a probability. In many elementary cases, all subsets are treated as events, assuming each has a well-defined (even if unknown) probability.

Example 1.1.1. In the case of rolling a die, the sample space is defined as $\Omega = \{1, 2, 3, 4, 5, 6\}$. Its elements correspond to the result of the die roll. Suppose every subset of Ω is considered an event. For instance, $\{2, 4, 6\}$ is an event, often described verbally as "we roll an even number."

Let's examine what operations we can perform on events.

Statement 1.1.1. *Since events are sets, we can define the* ***union*** *$(A \cup B)$,* ***intersection*** *$(A \cap B)$, and* ***complement*** *with respect to Ω $(\overline{A})$.*

The **difference** of two events can be expressed using the above: $A \backslash B = A \cap \overline{B}$. Two events are **disjoint** if $A \cap B = \emptyset$. The term **certain event** is used for the full set Ω, while the empty set (denoted by $\emptyset$) is referred to as the **impossible event**.

Example 1.1.2. Continuing with the dice example: the complement of the event {we roll an even number} is {we roll an odd number}; the intersection of the events {we roll an even number} and {we roll a number greater than 3} is $\{4, 6\}$, while their union is $\{2, 4, 5, 6\}$.

It is easy to see that when events are described by verbal statements (e.g., {we roll an even number}), their union corresponds to a logical "or", their intersection to "and", and the complement to logical negation.

The usual set-theoretic identities remain valid here as well: for example, $A \cup B = B \cup A$, $A \cap \Omega = A$, and so on. The following named identities are particularly worth noting:

Statement 1.1.2. (De Morgan's Laws) *For two sets:*

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \qquad \textit{and} \qquad \overline{A \cap B} = \overline{A} \cup \overline{B},$$

and for countably many sets:

$$\overline{\bigcup_{i=1}^{\infty} A_i} = \bigcap_{i=1}^{\infty} \overline{A_i} \qquad \textit{and} \qquad \overline{\bigcap_{i=1}^{\infty} A_i} = \bigcup_{i=1}^{\infty} \overline{A_i}.$$

The first pair of identities can easily be verified using a Venn diagram.

1.2 Event Algebra

We now consider the collection of all events. This set is called the **event algebra**. Note that the event algebra is not a subset of Ω, but rather a collection of subsets of Ω, that is, a set $\mathcal{F} \subseteq \mathcal{P}(\Omega)$. Here, $\mathcal{P}(\Omega)$ denotes the power set of Ω, i.e., the set of all subsets of Ω.

Example 1.2.1. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, and suppose that $\mathcal{F}$ consists exactly of the sets $\{1, 2, 3\}$, $\{4, 5, 6\}$, the impossible event $\emptyset$, and the certain event Ω. In this case, not every subset of Ω is considered an event. Still, this choice of $\mathcal{F}$ may model certain problems well (cf. Remark 2 in Section 1.1).

One might ask whether the union (or intersection or difference) of two events is also an event. The answer is yes—provided that we assume $\mathcal{F}$, the event algebra, is a so-called σ-algebra (pronounced "sigma algebra"), i.e., it satisfies the following:

Definition 1.2.1. *Let Ω be an arbitrary set, and let $\mathcal{F}$ be a collection of subsets of Ω. We say that $\mathcal{F}$ is a σ-**algebra** on Ω if the following three conditions hold:*

1. $\Omega \in \mathcal{F}$,
2. *if* $A \in \mathcal{F}$*, then* $\overline{A} \in \mathcal{F}$,
3. *if* $A_1, A_2, \dots \in \mathcal{F}$*, then* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

In short, $\mathcal{F}$ is a σ-algebra if it contains the full space, is closed under taking complements, and is closed under countable unions.

From the perspective of probability theory, the definition expresses the following: if we interpret $\mathcal{F}$ as the collection of observable events, then the conditions mean that we must be able to observe:

- events that are certain to occur (1),
- whether a given event does *not* occur (2),
- whether at least one of a countable sequence of observable events occurs (3).

The concept of σ-algebras originates from measure theory—a branch of analysis concerned with generalizations of notions such as area and volume. Kolmogorov employed results from this theory to lay the foundations of probability.

Several useful properties of σ-algebras follow directly from the definition:

Statement 1.2.1. *Let* $\mathcal{F}$ *be a* σ*-algebra over the base set* Ω*. Then the following hold:*

- $\emptyset \in \mathcal{F}$,
- *if* $A, B \in \mathcal{F}$*, then* $A \cup B, A \cap B, A \setminus B \in \mathcal{F}$,
- *if* $A_1, A_2, \dots \in \mathcal{F}$*, then* $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

Proof. From (1), we know that $\Omega \in \mathcal{F}$, and by (2), $\emptyset = \overline{\Omega} \in \mathcal{F}$.

To prove (2), let $A_1 = A$, $A_2 = B$, and $A_i = \emptyset$ for all $i \geq 3$. Then, by (3),

$$A \cup B = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F},$$

so $\mathcal{F}$ is indeed closed under finite unions. Closure under intersections follows from this: if $A, B \in \mathcal{F}$, then by (2), $\overline{A}, \overline{B} \in \mathcal{F}$, and since we have just shown closure under unions, $\overline{A} \cup \overline{B} \in \mathcal{F}$. But by De Morgan's law, $\overline{A} \cup \overline{B} = \overline{A \cap B}$, so $\overline{A \cap B} \in \mathcal{F}$, and using (2) again, $A \cap B \in \mathcal{F}$.

To prove (3), observe that from $A_i \in \mathcal{F}$ it follows that $\overline{A_i} \in \mathcal{F}$ by (2). Applying (3) to the sequence $\overline{A_1}, \overline{A_2}, \dots$, we get

$$\bigcup_{i=1}^{\infty} \overline{A_i} \in \mathcal{F}.$$

Then, by De Morgan's law for countable sets, this is the complement of $\bigcap_{i=1}^{\infty} A_i$, so since $\mathcal{F}$ is closed under complements, we conclude that $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$. □

1.3 Probability Space

We have discussed the sample space and the notion of events, but probabilities themselves have not yet been introduced. As mentioned earlier, we are interested in assigning probabilities $\mathbb{P}(A)$ to those subsets $A \subseteq \Omega$ that are considered events. That is, probability should be a function $\mathbb{P} : \mathcal{F} \to [0, 1]$, where $\mathcal{F}$ is a σ-algebra. But more than that is required.

Definition 1.3.1. *Let $\mathcal{F}$ be a σ-algebra over an arbitrary set Ω. A function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is called a* ***probability measure*** *if $\mathbb{P}(\Omega) = 1$, and the following condition holds:*
If $A_1, A_2, \dots \in \mathcal{F}$ is a sequence of pairwise disjoint events, i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$, then

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

This property says that the probability of a countable union of disjoint events is the (possibly infinite) sum of their individual probabilities. It is called ***σ-additivity***.

Definition 1.3.2. *A triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a* ***probability space*** *(in the sense of Kolmogorov) if $\mathcal{F}$ is a σ-algebra on Ω, and $\mathbb{P}$ is a probability measure.*

Example 1.3.1. In our moderately creative example of rolling a die, let $\Omega = \{1, 2, 3, 4, 5, 6\}$, and let $\mathcal{F}$ be the collection of all subsets. Then for an event A, we define $\mathbb{P}(A) = |A|/|\Omega|$, e.g., $\mathbb{P}(\{1, 2, 5, 6\}) = \frac{4}{6}$. Such a probability space—where $\mathbb{P}(A) = |A|/|\Omega|$—is called a **classical probability space**. A generalization is the **geometric probability space**, where Ω is a subset of the plane, space, or $\mathbb{R}^n$, and $\mathbb{P}(A) = \lambda(A)/\lambda(\Omega)$, where λ denotes area, volume, or n-dimensional volume.

Example 1.3.2. Consider a 5-question test with yes/no answers, where the test taker has a 60% chance of answering each question correctly, independently of the others. Then the sample space $\Omega = \{0, 1\}^5$ models the situation (and $\mathcal{F}$ is the set of all subsets). This is no longer a classical probability space, since the probability of the event consisting only of the outcome $(0, 1, 0, 1, 0)$ is $0.4^3 \cdot 0.6^2 \approx 0.023$, not $1/2^5 \approx 0.031$.

What do we expect from a well-behaved notion of probability? For instance, the following property—though not part of the definition—is easily derived from it:

Statement 1.3.1. *If A and B are disjoint, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.*

Proof. Use the σ-additivity of the probability measure with the choice $A_1 = A$, $A_2 = B$, and $A_i = \emptyset$ for all $i \geq 3$. Then

$$\mathbb{P}(A \cup B) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \mathbb{P}(A) + \mathbb{P}(B) + \sum_{i=3}^{\infty} \mathbb{P}(\emptyset).$$

This is only possible if $\mathbb{P}(\emptyset) = 0$ (otherwise the right-hand side would diverge, while the left-hand side must lie in $[0, 1]$). The statement follows. □

Corollary 1.3.1. *For arbitrary events $A, B \in \mathcal{F}$, the following hold:*

1. $\mathbb{P}(A) + \mathbb{P}(\overline{A}) = 1$,
2. $\mathbb{P}(A \cap B) + \mathbb{P}(A \cap \overline{B}) = \mathbb{P}(A)$,
3. *If $B \subseteq A$, then $\mathbb{P}(B) \leq \mathbb{P}(A)$.*

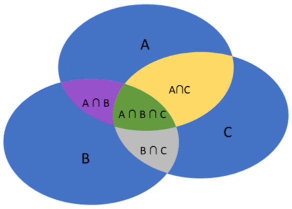
Proof. For the first statement, apply finite additivity to the disjoint events A and $\overline{A}$. For the second, use the disjoint decomposition $A = (A \cap B) \cup (A \cap \overline{B})$. The third follows from the second, since if $B \subseteq A$, then $A \cap B = B$, and $\mathbb{P}(A \cap \overline{B}) \geq 0$. □

1.4 Poincaré Formula

How can we compute the probability of a union of events when the events are not necessarily disjoint? The answer is given by the Poincaré formula (also known as the inclusion-exclusion formula), which only requires the finite additivity property discussed above.

Example 1.4.1. Let A, B, C be three events—for instance, hitting three regions on a target, where we aim to hit at least one of them.

To compute $\mathbb{P}(A \cup B \cup C)$, we might start from $\mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$. But note that in doing so, we've counted the probabilities of pairwise intersections twice, so we need to subtract $\mathbb{P}(A \cap B) + \mathbb{P}(A \cap C) + \mathbb{P}(B \cap C)$ for a better approximation. However, now the probability of $A \cap B \cap C$ has been subtracted three times, after having been added three times—so it has effectively been excluded, although it should be counted once. Thus we arrive at:



$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$$

This is the special case of the Poincaré formula for three events; the general form extends this to any number n of events.

To state the general theorem concisely, we introduce some notation. Let $[n] = \{1, 2, \ldots, n\}$, and let $A_1, A_2, \ldots, A_n \in \mathcal{F}$ be events. Let $I = \{i_1, i_2, \ldots, i_k\}$ be a k-element subset of $[n]$. Then we may consider the event $\bigcap_{i \in I} A_i$, the probability that all the k events indexed by I occur simultaneously. Define

$$S_k \overset{\text{def}}{=} \sum_{I \subseteq [n],\, |I|=k} \mathbb{P}\left(\bigcap_{i \in I} A_i\right),$$

that is, we sum the probabilities of all intersections of k distinct events. For example, when $k = 1$, we simply get $S_1 = \sum_{i=1}^{n} \mathbb{P}(A_i)$.

Theorem 1.4.1 (Poincaré Formula)**.** *Let $A_1, A_2, \ldots, A_n \in \mathcal{F}$ be events. Then*

$$\mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) = \sum_{k=1}^{n} (-1)^{k+1} S_k.$$

A natural question is what happens if we truncate the sum after the first few terms, ignoring the rest (e.g., because in practice they are small and computationally expensive). Even then, we can extract meaningful information about the probability of the union.

Statement 1.4.1 (Bonferroni Inequalities)**.** *Let $A_1, A_2, \ldots, A_n \in \mathcal{F}$ be events, and let $1 \le m_1, m_2 \le n$ be integers such that m_1 is odd and m_2 is even. Then:*

$$\mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) \le \sum_{k=1}^{m_1} (-1)^{k+1} S_k, \quad \text{and} \quad \mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) \ge \sum_{k=1}^{m_2} (-1)^{k+1} S_k.$$

We will not prove these two statements here. See the exercises for examples.

Corollary 1.4.1 (Boole's Inequality)**.** *Let $A_1, A_2, \ldots, A_n \in \mathcal{F}$ be events. Then:*

$$\mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) \le \sum_{j=1}^{n} \mathbb{P}(A_j), \quad \text{and} \quad \mathbb{P}\left(\bigcap_{j=1}^{n} A_j\right) \ge 1 - \sum_{j=1}^{n} \mathbb{P}(\overline{A_j}).$$

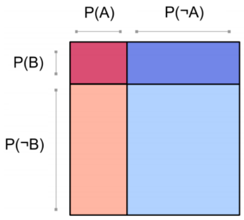
Proof. Apply the first Bonferroni inequality with $m_1 = 1$, which yields the first inequality, since S_1 is the sum $\sum_j \mathbb{P}(A_j)$. The second inequality follows from the first and De Morgan's law, applied to the events $\overline{A_j}$. □

2 Properties of Probability

Throughout this chapter, we assume that a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is given. We will discuss the concepts of independence and conditional probability.

2.1 Independence

Previously, we encountered the case when the probability of the union of two events adds up, i.e., $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. This required the events to be disjoint. However, there are cases where probabilities multiply under certain assumptions.



Definition 2.1.1. *Two events A and B are said to be **independent** if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Independence is quite different from disjointness. It formalizes the idea that the occurrence of one event does not influence the occurrence of the other.

Example 2.1.1. If event A occurs with probability 1/3, and event B occurs with probability 1/4, and we assume no interaction between them, we estimate the probability of both occurring as $\frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}$.
Note that independence is defined purely in terms of probabilities, so two events can be independent even if they seem to influence each other. For instance, when rolling two dice, the events {first die shows 1} and {the two dice show the same number} are independent.

Statement 2.1.1. *If A and B are independent, then so are A and $\overline{B}$.*

Proof. Using the identity $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap \overline{B})$ and the independence of A and B, we have:

$$\mathbb{P}(A \cap \overline{B}) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)(1 - \mathbb{P}(B)) = \mathbb{P}(A)\mathbb{P}(\overline{B}).$$

□

Now let us define independence for multiple events.

Definition 2.1.2. *The events $A_1, \ldots, A_n$ are said to be **mutually independent** if for every subset $I \subseteq [n]$,*

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

In other words, the probability of the intersection of any subcollection equals the product of their individual probabilities.

This may seem like a complicated condition, but it turns out to be the right one. One might wonder why we don't simply require independence for the full intersection $\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n)$. However, that alone is not sufficient.[1]
Also, pairwise independence does not imply mutual independence. The following example shows how subtle this concept is:

Example 2.1.2. Flip two fair coins. Let:

- $A_1 = \{\text{first coin is heads}\}$,
- $A_2 = \{\text{second coin is heads}\}$,
- $A_3 = \{\text{number of heads is even}\}$.

Any pair among A_1, A_2, A_3 is independent, but the triple $\{A_1, A_2, A_3\}$ is not mutually independent, since

$$\mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) = \frac{1}{8}, \quad \text{but} \quad \mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(\text{both coins are heads}) = \frac{1}{4}.$$

This has a linear algebra analogue: the vectors $(1,0), (0,1), (1,1)$ are pairwise linearly independent but not jointly.

2.2 Conditional Probability

How can we "measure" how much one event depends on another?

Definition 2.2.1. *Let $A, B \in \mathcal{F}$ be events with $\mathbb{P}(A) > 0$. The* ***conditional probability*** *of B given A is defined as*

$$\mathbb{P}(B \mid A) \overset{\text{def}}{=} \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}.$$

Read: "the probability of B, given A."

Note that A and B are independent if and only if $\mathbb{P}(B \mid A) = \mathbb{P}(B)$. That is, B is independent of A if its probability is unaffected by whether A occurred. In fact, independence can be defined this way, whenever $\mathbb{P}(A) > 0$.

Example 2.2.1. Consider rolling a die. Let $A = \{\text{even number}\}$. Then:

- $\mathbb{P}(\text{rolling a 6} \mid A) = \frac{1}{3}$,
- $\mathbb{P}(\text{rolling a 1} \mid A) = 0$,
- $\mathbb{P}(\text{rolling more than 3} \mid A) = \frac{2}{3}$,

[1] Mutual independence implies the independence of all subcollections, but to ensure this, we need the full definition above.

- $\mathbb{P}(\text{even number} \mid A) = 1$.

Of course, conditional probability is not just for measuring relationships between events. In many problems, we are given conditional information—for example: "if I arrive at the exam well-prepared, I have at least $1-\varepsilon$ chance of passing."[2]
Let us examine some properties of conditional probability:

Statement 2.2.1. *Let $A \in \mathcal{F}$ be a fixed event with $\mathbb{P}(A) > 0$. Then the map*

$$B \mapsto \mathbb{P}(B \mid A)$$

is also a probability measure on $\mathcal{F}$.

This is useful because we can replace $\mathbb{P}(\cdot)$ with $\mathbb{P}(\cdot \mid A)$ in any earlier theorem about probability, and it remains valid.

Proof. Clearly, $\mathbb{P}(\Omega \mid A) = \frac{\mathbb{P}(\Omega \cap A)}{\mathbb{P}(A)} = 1$. Let $B_1, B_2, \ldots$ be pairwise disjoint events. Then, since $\mathbb{P}$ is a probability measure:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i \;\middle|\; A\right) = \frac{\mathbb{P}\left(\bigcup_{i=1}^{\infty}(B_i \cap A)\right)}{\mathbb{P}(A)} = \sum_{i=1}^{\infty} \frac{\mathbb{P}(B_i \cap A)}{\mathbb{P}(A)} = \sum_{i=1}^{\infty} \mathbb{P}(B_i \mid A).$$

□

Using conditional probability, we can state the probabilistic version of case analysis:

Statement 2.2.2 (*Law of Total Probability*)**.** *Let $A_1, \ldots, A_n \in \mathcal{F}$ be pairwise disjoint events such that $\bigcup_{i=1}^{n} A_i = \Omega$ and $\mathbb{P}(A_i) > 0$ for all i. Then:*

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i).$$

Definition 2.2.2. *A collection of pairwise disjoint events $A_1, \ldots, A_n \in \mathcal{F}$ is called a* ***partition*** *(or* ***complete system of events****) if $\bigcup_{i=1}^{n} A_i = \Omega$.*

Proof of the proposition. Substituting the definition of conditional probability, we get:

$$\sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i) = \sum_{i=1}^{n} \mathbb{P}(B \cap A_i).$$

Since the union $\bigcup_{i=1}^{n}(B \cap A_i) = B \cap \Omega = B$, and the sets are disjoint, the additivity of $\mathbb{P}$ gives the result. □

[2] Conditional probability also connects to Bertrand's box paradox from the first lecture.

Example 2.2.2 (Monty Hall paradox)**.** Suppose there are three doors, behind one of which is a car, and behind the other two are goats. The puzzle proceeds as follows: first, we choose a door. Then the host opens one of the remaining doors, behind which there is a goat. At this point, we are given the opportunity to switch our choice. The question is: should we switch, assuming we prefer winning the car over ending up with a goat?

The surprising answer, contrary to the intuition that "it doesn't matter," is yes. Indeed, if we stick with our initial choice, the probability of winning is clearly $\frac{1}{3}$. But if we switch, then

$$\begin{aligned}\mathbb{P}(\text{finally car}) &= \mathbb{P}(\text{finally car} \mid \text{initially goat}) \cdot \mathbb{P}(\text{initially goat}) \\ &\quad + \mathbb{P}(\text{finally car} \mid \text{initially car}) \cdot \mathbb{P}(\text{initially car}) = 1 \cdot \frac{2}{3} + 0 \cdot \frac{1}{3} = \frac{2}{3},\end{aligned}$$

since if we initially pick a goat, the host has no choice but to reveal the other goat.

There are also situations where we must work with probabilities under several successive, dependent conditions.

Example 2.2.3. We draw three cards, without replacement, from a shuffled 52-card French deck.
What is the probability of drawing a king first, a queen second, and a jack third?
Although the result of each draw influences the next (e.g., drawing a king reduces the number of remaining cards, hence the chance of drawing another king), the correct result is given by the following computation:
Let K_1 denote the event that the first card is a king, Q_2 the event that the second card is a queen, and J_3 the event that the third card is a jack. Then the desired probability is

$$\mathbb{P}(K_1) \cdot \mathbb{P}(Q_2 \mid K_1) \cdot \mathbb{P}(J_3 \mid Q_2 \cap K_1) = \frac{4}{52} \cdot \frac{4}{51} \cdot \frac{4}{50} \approx 0.0005.$$

This approach is generalized by the following proposition.

Statement 2.2.3 (*Multiplication rule*)**.** *Let $A_1, \ldots, A_n \in \mathcal{F}$ be events with $\mathbb{P}(A_i) > 0$ for all i. Then*

$$\mathbb{P}\Big(\bigcap_{i=1}^{n} A_i\Big) = \mathbb{P}(A_1) \cdot \prod_{i=2}^{n} \mathbb{P}\Big(A_i \,\Big|\, \bigcap_{k=1}^{i-1} A_k\Big).$$

The proof follows directly by expanding the definition of conditional probability and simplifying the resulting product.

2.3 Bayes' Theorem

Among phenomena involving conditional probability, Bayes' theorem and the paradox it resolves are of particular importance. (The paradox is also known under other names, e.g., the false positive paradox or base rate fallacy.)

Bayes paradox: During a chest X-ray, the probability that tuberculosis is detected in a person who actually has the disease is 0.95. The probability that a healthy person is mistakenly diagnosed with the disease is 0.001. The proportion of people with tuberculosis in the population is 0.0001. What is the probability that a person is actually healthy, given that the test result is positive?

Statement 2.3.1. (Simple Bayes' theorem) *Let $A, B \in \mathcal{F}$ be events such that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Then*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

The proof follows immediately by substituting the definitions. The theorem is often applied in combination with the law of total probability:

Statement 2.3.2. (Bayes' theorem) *Let $B, A_1, A_2, \ldots, A_n \in \mathcal{F}$ be events such that $\mathbb{P}(B) > 0$, $\mathbb{P}(A_i) > 0$ for all i, and $A_1, \ldots, A_n$ form a complete system of events. Then*

$$\mathbb{P}(A_1 \mid B) = \frac{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1)}{\sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)}.$$

Proof. We apply the simple Bayes' theorem to A_1 and B, and expand the denominator using the law of total probability:

$$\mathbb{P}(A_1 \mid B) = \frac{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1)}{\sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)},$$

which is exactly the desired statement. □

Example 2.3.1. Let us return to the example above. Let $A_1 = \{\text{the person is healthy}\}$, $A_2 = \overline{A_1}$ and $B = \{\text{the test is positive}\}$. Then

$$\mathbb{P}(A_1 \mid B) = \frac{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1)}{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1) + \mathbb{P}(B \mid A_2)\mathbb{P}(A_2)} = \frac{0.001 \cdot 0.9999}{0.001 \cdot 0.9999 + 0.95 \cdot 0.0001} \approx 0.9132$$

which does not paint a very reassuring picture of a test that might otherwise be considered 95% reliable. The result is only an apparent contradiction: it arises from the fact that the vast majority of the population is healthy, so there are many more "opportunities" for false positives than for false negatives.

3 Discrete Random Variables

The definitions introduced in the previous two lectures (event algebra, conditional probability) are indeed fundamental concepts of the topic, but they are not sufficient to naturally express certain problems.

For example, how could we formulate with the tools introduced so far that the outcomes of two dice rolls are independent? Or that the expected value of a single die roll is 3.5, and the expected value of a random number uniformly selected from 0 to 1 is $\frac{1}{2}$? To do this, we need to talk not only about events, but also about random quantities—so-called *random variables.*

3.1 Random Variable

Definition 3.1.1. *Let $X : \Omega \to \mathbb{R}$ be a function. For a given $x \in \mathbb{R}$, let*

$$\{X < x\} \overset{\text{def}}{=} \{\omega \in \Omega \mid X(\omega) < x\},$$

that is, the set of outcomes for which X is less than x. These sets are called the ***level sets*** *of X.*
We call the function X a ***random variable*** *if for every $x \in \mathbb{R}$,*

$$\{X < x\} \in \mathcal{F},$$

that is, the level sets of X are events.

Example 3.1.1. We have already seen random variables in earlier examples, we just didn't name them as such. Some examples of random variables:

1. The result of a single die roll. In this case, the condition from the definition of a random variable $\{X < x\} \in \mathcal{F}$ for all real x"—is equivalent to requiring that for every k, the set of outcomes where we roll a k is an event.

2. The square of the result of a die roll. Its possible values are 1, 4, 9, 16, 25, and 36, each occurring with probability $\frac{1}{6}$. Formally, we may choose the sample space as $\Omega = \{1, 2, 3, 4, 5, 6\}$, with $\mathcal{F}$ and $\mathbb{P}$ as before, and define the random variable as $Y(\omega) = \omega^2$.

3. An urn contains 2 white and 3 red balls. We draw without replacement until we draw a white ball. The number of red balls drawn before the white one is a random variable.

4. We can also define a random variable based on an event. Let

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases}$$

   This is called the **indicator random variable** associated with the event A.

3.2 Expected Value in the Finite Case

When dealing with a random quantity, one of the most natural questions is: "Okay, it's random—but what is its average value?"
The concept of expected value captures this idea of the "average" among the possible outcomes.

Definition 3.2.1. *A random variable X is called* ***simple*** *if it can take only finitely many values. The* ***expected value*** *of a simple random variable is defined as:*

$$\mathbb{E}(X) \overset{\text{def}}{=} \sum_{k\in \mathrm{Ran}(X)} k \cdot \mathbb{P}(X = k),$$

where $\mathrm{Ran}(X)$ *is the finite range of X, and $\mathbb{P}(X = k)$ denotes the probability of the event $\{X = k\}$.*

What does this mean? Why does this strange sum represent an "average"? The formula tells us to take the weighted average of the values of the random variable X, with the weights being the corresponding probabilities.
The name "expected value" is somewhat misleading: the result is not necessarily a value one might "expect" to occur. For instance, if we blindly grab a slipper, then either one foot has the correct slipper or neither—but the expected number of correctly worn slippers is 1, assuming both outcomes are equally likely.
It is important that the formula includes the factor k in the summation. Without it, we would always get $\sum_{k\in \mathrm{Ran}(X)} \mathbb{P}(X = k) = 1$ for any simple random variable X.

Example 3.2.1. Let us compute the expected values of the random variables from the previous examples:

1. In the case of a single die roll, we have $\mathrm{Ran}(X) = \{1, 2, 3, 4, 5, 6\}$ and $\mathbb{P}(X = k) = \frac{1}{6}$ for all $k \in \mathrm{Ran}(X)$, so:

$$\mathbb{E}(X) = \sum_{k=1}^{6} k \cdot \mathbb{P}(X = k) = \sum_{k=1}^{6} k \cdot \frac{1}{6} = \frac{21}{6} = 3.5.$$

2. For the square of the die roll, similarly:

$$\mathbb{E}(Y) = \sum_{k\in\{1,4,9,16,25,36\}} k \cdot \mathbb{P}(Y = k) = (1 + 4 + 9 + 16 + 25 + 36) \cdot \frac{1}{6} = \frac{91}{6} \approx 15.1667.$$

3. Let Z denote the number of red balls drawn until the first white one is drawn:

$$\mathbb{E}(Z) = \sum_{k=0}^{3} k \cdot \mathbb{P}(Z = k) \overset{\text{multiplication rule}}{=}$$

$$= 0 \cdot \frac{2}{5} + 1 \cdot \frac{3}{5} \cdot \frac{2}{4} + 2 \cdot \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} + 3 \cdot \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \cdot \frac{2}{2} = \frac{0 + 3 + 4 + 3}{10} = 1.$$

4. Expected value of an indicator random variable:

$$\mathbb{E}(\mathbf{1}_A) = \sum_{k\in\{0,1\}} k \cdot \mathbb{P}(\mathbf{1}_A = k) = 0 \cdot \mathbb{P}(\mathbf{1}_A = 0) + 1 \cdot \mathbb{P}(\mathbf{1}_A = 1) = \mathbb{P}(A).$$

   In this sense, the expected value extends the notion of probability from indicator variables to (so far, only simple) random variables.

Random variables, like real-valued functions, support the usual operations: if X and Y are random variables, then $X+Y$ is the function defined by $(X+Y)(\omega) = X(\omega)+Y(\omega)$. It can be shown that the sum is also a random variable.[3] Similarly, we can define the difference, product, and—provided the denominator is never zero—the quotient of random variables.
One of the most frequently used properties of expectation is that it is linear.
This means, on the one hand, that for any $c \in \mathbb{R}$, we have $\mathbb{E}(cX) = c \cdot \mathbb{E}(X)$ (which is easy to verify). On the other hand, expectation is also additive:

Statement 3.2.1. *Let X and Y be simple random variables. Then $\mathbb{E}(X+Y) = \mathbb{E}X+\mathbb{E}Y$.*

Proof. Let $M = \mathrm{Ran}(X+Y)$, $K = \mathrm{Ran}(X)$ and $L = \mathrm{Ran}(Y)$. Expanding the definitions:

$$\mathbb{E}(X+Y) = \sum_{m\in M} m \cdot \mathbb{P}(X+Y=m) = \sum_{m\in M} m \cdot \mathbb{P}\Big(\bigcup_{\substack{k\in K,\, l\in L\\ k+l=m}} \{X=k,\, Y=l\}\Big) =$$

$$= \sum_{m\in M} \sum_{\substack{k\in K,\, l\in L\\ k+l=m}} (k+l) \cdot \mathbb{P}(X=k,\, Y=l) = \sum_{k\in K}\sum_{l\in L}(k+l) \cdot \mathbb{P}(X=k,\, Y=l)$$

$$= \sum_{k\in K} k \cdot \mathbb{P}\Big(\bigcup_{l\in L}\{X=k,\, Y=l\}\Big) + \sum_{l\in L} l \cdot \mathbb{P}\Big(\bigcup_{k\in K}\{X=k,\, Y=l\}\Big)$$

$$= \sum_{k\in K} k \cdot \mathbb{P}(X=k) + \sum_{l\in L} l \cdot \mathbb{P}(Y=l) = \mathbb{E}X + \mathbb{E}Y,$$

which is exactly what we wanted to prove. □

Additivity is a useful tool even when the problem does not explicitly involve the sum of random variables.

Example 3.2.2. Let us prove the Poincaré formula for three sets:

$$\mathbb{P}(A_1 \cup A_2 \cup A_3) = \sum_{i=1}^{3} \mathbb{P}(A_i) - \sum_{1\le i<j\le 3} \mathbb{P}(A_i \cap A_j) + \mathbb{P}(A_1 \cap A_2 \cap A_3).$$

[3] For non-simple random variables, the proof is not immediate. It is useful to use the fact that the rational numbers are dense, and that $\{X+Y<x\} = \bigcup_{r\in\mathbb{Q}}(\{X<r\} \cap \{Y<x-r\})$.

$$\mathbb{P}(\cup_i A_i) = 1 - \mathbb{P}\big(\cap_i \overline{A_i}\big) = 1 - \mathbb{E}\Big(\mathbf{1}_{\cap_i \overline{A_i}}\Big) = 1 - \mathbb{E}\Big(\prod_i \mathbf{1}_{\overline{A_i}}\Big) = 1 - \mathbb{E}\Big(\prod_i (1 - \mathbf{1}_{A_i})\Big) =$$

$$= 1 - \mathbb{E}\Bigg(1 - \mathbf{1}_{A_1} - \mathbf{1}_{A_2} - \mathbf{1}_{A_3} + \mathbf{1}_{A_1 \cap A_2} + \mathbf{1}_{A_2 \cap A_3} + \mathbf{1}_{A_1 \cap A_3} - \mathbf{1}_{A_1 \cap A_2 \cap A_3}\Bigg)$$

$$= \sum_{i=1}^{3} \mathbb{P}(A_i) - \sum_{1 \le i < j \le 3} \mathbb{P}(A_i \cap A_j) + \mathbb{P}(A_1 \cap A_2 \cap A_3),$$

which is exactly what we wanted to prove.
Note that in the first line of the computation, we did not use the fact that we had only 3 sets. In fact, the same reasoning works for any finite number of sets, and so the Poincaré formula is proved in general.

We have seen that to compute the expectation of a random variable, it suffices to know the values $\mathbb{P}(X = k)$. The collection of these probabilities is called the **distribution** of the simple random variable.

Let us now look at some notable distributions:

Definition 3.2.2. *A random variable X has a* ***binomial distribution*** *with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ if*

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{for } k \in \{0, 1, 2, \ldots, n\}.$$

Notation: $X \sim B(n; p)$.

Example 3.2.3. We toss a biased coin n times, where each toss shows heads with probability p. Then the number of heads is a binomially distributed random variable.

More generally, if we perform independent trials with the same probability of success, then the number of successes in n trials is binomially distributed with parameters n and p. Formally, let $A_1, \ldots, A_n$ be mutually independent events such that $\mathbb{P}(A_i) = p$ for all i. Then:

$$\mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_n} \sim B(n; p),$$

that is, a random variable with distribution $B(n; p)$ can always be seen as the sum of n indicator random variables of jointly independent events.
This observation is immediately useful: if $X \sim B(n; p)$, then

$$\mathbb{E}(X) = \mathbb{E}(\mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_n}) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) = n \cdot p$$

by the additivity of expectation.

Definition 3.2.3. *A random variable X has a* ***uniform distribution*** *on a finite set $S \subseteq \mathbb{R}$ of size n if*

$$\mathbb{P}(X = k) = \frac{1}{n}$$

for all $k \in S$. If $S = \{1, 2, \ldots, n\}$, then the expected value of X is $\mathbb{E}(X) = \frac{1+2+\cdots+n}{n} = \frac{n+1}{2}$.

4 Continuous Random Variables

So far, we have dealt with random variables that take values in a finite or countably infinite set—typically some subset of the integers.
However, there are random quantities that are better modeled as taking on any real value. Examples include many physical quantities or the time elapsed until a certain event occurs. These will now be our focus.

4.1 Cumulative Distribution Function

In exercises, we have already encountered cases where we "uniformly randomly" selected a number from an interval, or a point from a two-dimensional shape. These are exactly those types of random quantities—random variables, in last week's terminology—that are not discrete, i.e., they do not take values from a countable set.
Why is this a problem? When X is a discrete random variable, it can be described via its **distribution**, i.e., a sequence of numbers between 0 and 1:

$$\mathbb{P}(X = k_1),\ \mathbb{P}(X = k_2),\ \ldots$$

where $k_1, k_2, \ldots$ are the possible values of X. Now let X be a number chosen uniformly at random from the interval $[0, 1]$.
By this, we mean a random variable such that $\mathbb{P}(X \leq t) = t$ for $t \in [0, 1]$, for example $\mathbb{P}(X \leq \frac{1}{2}) = \frac{1}{2}$. In this case, $\mathbb{P}(X = k) = 0$ for any k, so the above (discrete-style) distribution tells us essentially nothing about the random variable.
One might object: "Okay, but we still know that X takes values from $[0, 1]$, so why not just refer to that?" The following example shows why that is insufficient.

Example 4.1.1. Let X be uniformly distributed over $[0, 1]$, and consider the random variable $Y = X^2$.
Then Y also takes values in $[0, 1]$, but it behaves differently than X.
Indeed, $\mathbb{P}(X \leq \frac{1}{2}) = \frac{1}{2}$, while

$$\mathbb{P}\Big(Y \leq \frac{1}{2}\Big) = \mathbb{P}\Big(X^2 \leq \frac{1}{2}\Big) = \mathbb{P}\Big(X \leq \frac{1}{\sqrt{2}}\Big) = \frac{1}{\sqrt{2}}.$$

Thus, Y is more likely to take smaller values—its "distribution is more concentrated near 0" than that of X. We'll clarify what this means shortly.

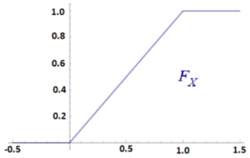
In general, the distribution of X can be described using its so-called cumulative distribution function.

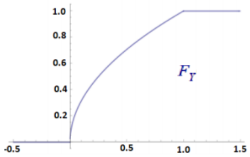
Definition 4.1.1. *Let X be a random variable. Then the function*

$$F_X : \mathbb{R} \to [0, 1] \qquad F_X(x) = \mathbb{P}(X < x)$$

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

Note that $\{X < x\}$ is an element of $\mathcal{F}$, i.e., it is an event (since X is a random variable), so it makes sense to talk about $\mathbb{P}(X < x)$.
The CDFs of X and Y from the example above are:





$$F_X(x) = \mathbb{P}(X < x) = \begin{cases} 0 & \text{if } x \leq 0, \\ x & \text{if } x \in (0,1], \\ 1 & \text{if } x > 1, \end{cases}$$

$$F_Y(x) = \mathbb{P}(Y < x) = \mathbb{P}(X^2 < x) = \mathbb{P}(X < \sqrt{x}) = \begin{cases} 0 & \text{if } x \leq 0, \\ \sqrt{x} & \text{if } x \in (0,1], \\ 1 & \text{if } x > 1. \end{cases}$$

Let us now consider what a CDF can generally tell us. Clearly,

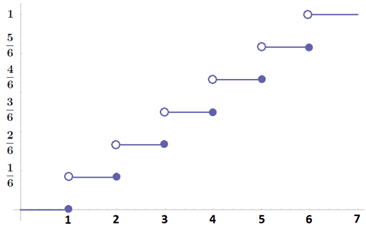
$$F_X(b) - F_X(a) = \mathbb{P}(X < b) - \mathbb{P}(X < a) = \mathbb{P}(a \leq X < b)$$

for any real numbers $a < b$, due to the additivity of $\mathbb{P}$. Moreover, CDFs can be characterized as follows:

Statement 4.1.1. *A function $F : \mathbb{R} \to [0,1]$ is the cumulative distribution function of some random variable if and only if:*

1. *F is (not necessarily strictly) increasing;*
2. *F is left-continuous, i.e., for every x, the left-hand limit of F at x equals $F(x)$;*
3. *$\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.*

Note that left-continuity is far weaker than continuity. For example, the CDF of a discrete random variable is never continuous. Indeed, discrete random variables also have CDFs, which are always left-continuous, but not right-continuous. See the CDF of a die roll to the right.



Proof. Let X be a random variable, and let $x < y$. We verify that F_X satisfies the three conditions above.
Indeed, F_X is increasing since $\{X < x\} \subseteq \{X < y\}$, hence

$$F_X(x) = \mathbb{P}(X < x) \leq \mathbb{P}(X < y) = F_X(y),$$

using the consequence of the properties of the probability space in subsection.
Now for the second property. Left-continuity of F_X is equivalent (by a transfer principle) to the following: for any increasing sequence $(x_n)_{n\in\mathbb{N}}$ with $x_n \neq x$ and $x_n \to x$, we must have $\lim_{n\to\infty} F_X(x_n) = F_X(x)$. We show this holds. First,

$$\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} \mathbb{P}(X < x_n) = \mathbb{P}(X < x_0) + \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{k=1}^{n} \{x_{k-1} \leq X < x_k\}\Big),$$

using the additivity of probability and the identity $\{X < x_n\} = \{X < x_0\} \cup \bigcup_{k=1}^{n} \{x_{k-1} \leq X < x_k\}$. The second term becomes:

$$\lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{k=1}^{n} \{x_{k-1} \leq X < x_k\}\Big) = \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(x_{k-1} \leq X < x_k) = \mathbb{P}(x_0 \leq X < x),$$

using σ-additivity and the fact that $\cup_{k=1}^{\infty} \{x_{k-1} \leq X < x_k\} = \{x_0 \leq X < x\}$. Plugging this back in, we obtain:

$$\lim_{n\to\infty} F_X(x_n) = \mathbb{P}(X < x_0) + \mathbb{P}(x_0 \leq X < x) = \mathbb{P}(X < x) = F_X(x).^{4}$$

A similar argument shows that the third property holds as well.[5]

Conversely, suppose F satisfies the three conditions. We construct a random variable X such that $F = F_X$. Let U be a random variable uniformly distributed on $[0, 1]$.[6] Define X by

$$X = \inf\{y \in \mathbb{R} \mid U < F(y)\}.$$

Then, by the definition of X and the infimum:

$$\mathbb{P}(X < x) = \mathbb{P}\big(\inf\{y \in \mathbb{R} \mid U < F(y)\} < x\big) =$$

$$= \mathbb{P}(\text{there exists } y \in \mathbb{R} \text{ with } y < x \text{ and } U < F(y)).$$

We show that such a y exists if and only if $U < F(x)$. Indeed, if such a y exists, then $U < F(y) \leq F(x)$ since F is increasing.

Conversely, if $U < F(x)$, then by the left-continuity of F, there exists some $y < x$ such that $U < F(y)$. Hence:

$$\mathbb{P}(X < x) = \mathbb{P}(U < F(x)) = F(x),$$

since $0 \leq F(x) \leq 1$, and from the earlier example, we know that $\mathbb{P}(U < z) = z$ for any $0 < z < 1$.

This completes the proof. □

4.2 Probability Density Function

We now examine another important function associated with random variables: the density function. One reason for this is that the graph of the cumulative distribution function (CDF) does not always clearly reveal the properties of the random variable. Recall the case of X and X^2, where X is uniformly distributed on $[0, 1]$. Can we determine from the

[4] This also shows that the right-hand limit of F_X at x equals $\mathbb{P}(X \leq x)$.

[5] This is known as the continuity property of probability.

[6] One should prove that such a uniformly random variable exists. A precise definition would require the notion of Lebesgue measure, which we omit here.

CDF graph where X^2 is most likely to fall within a radius of 0.01? Or how many times more likely it is for X^2 to fall near $\frac{1}{4}$ than near $\frac{3}{4}$?
The first question is easy to answer: at 0 (more precisely, at 0.01), since the CDF increases most steeply there—in other words, those x values contribute the most to the increase of $F_{X^2}(x) = \mathbb{P}(X^2 < x)$.
The second question is trickier: the answer is $\sqrt{3}$, and it involves comparing the slopes of the tangents to the CDF—see below. In both cases, we are effectively analyzing slopes.
If the phrase "slope of the tangent" raises any alarm bells, rest assured—we will take derivatives. As the first example shows, what would help us is the derivative of the CDF, i.e., a function that measures the steepness of the distribution function.
Even though F_{X^2} is not differentiable at 0 and 1, this issue can be sidestepped using a generalization of the derivative. Following a "good-enough" philosophy, we will simply ignore non-differentiable points (as long as the function is at least continuous). This will not affect our computations.

Definition 4.2.1. *A random variable X is said to be **continuous** if there exists a non-negative real-valued function $f_X : \mathbb{R} \to \mathbb{R}$ such that the improper Riemann integral $\int_{-\infty}^{\infty} f_X(z)\,\mathrm{d}z$ is finite,*[7] *and for all $x \in \mathbb{R}$,*

$$\int_{-\infty}^{x} f_X(z)\,\mathrm{d}z = F_X(x),$$

*where the integral is an improper Riemann integral. The function f_X is called the **probability density function** (PDF) of X.*

This definition is not very constructive: it is difficult to compute f_X from the given integrals.
In fact, it doesn't guarantee uniqueness, because if f is a PDF of X, then modifying f at a single point still yields a valid PDF (since integrals are unaffected).
How can we compute something that isn't uniquely defined?

Statement 4.2.1. *If F_X is continuous and differentiable except at finitely many points, then X is a continuous random variable, and the function*

$$f(x) = \begin{cases} F_X'(x) & \textit{if } F_X \textit{ is differentiable at } x, \\ 0 & \textit{otherwise} \end{cases} \qquad (x \in \mathbb{R})$$

is a density function for X.

Example 4.2.1. According to the proposition, for X uniformly distributed on $[0, 1]$, we have:

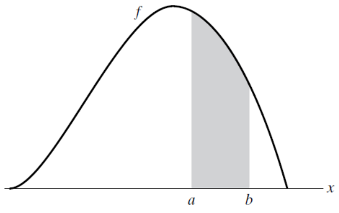
$$f_X(x) = \begin{cases} 1 & \text{if } 0 < x < 1, \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad f_{X^2}(x) = \begin{cases} \frac{1}{2\sqrt{x}} & \text{if } 0 < x < 1, \\ 0 & \text{otherwise} \end{cases} \qquad (x \in \mathbb{R})$$

[7] More generally, we could consider density functions that are Lebesgue integrable, but we won't do that here.

These are the density functions of X and X^2. Intuitively, the value of the density function at x indicates how likely X is to fall near x[8].

$$f_{X^2}\left(\frac{1}{4}\right)/f_{X^2}\left(\frac{3}{4}\right) = \frac{1}{1}\Big/\frac{1}{\sqrt{3}} = \sqrt{3}.$$

The usefulness of the density function goes far beyond giving us information about small neighborhoods. Let us now list some of its properties.



Statement 4.2.2. *Let X be a continuous random variable. Then for all real numbers $a < b$,*

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)\,\mathrm{d}x.$$

Proof. By the additivity of probability:

$$\mathbb{P}(a < X < b) = \mathbb{P}(X < b) - \mathbb{P}(X < a) - \mathbb{P}(X = a) =$$
$$= \int_{-\infty}^b f_X(x)\,\mathrm{d}x - \int_{-\infty}^a f_X(x)\,\mathrm{d}x - 0 = \int_a^b f_X(x)\,\mathrm{d}x,$$

using the fact that the integral is additive over intervals. □

Just like cumulative distribution functions, density functions can also be characterized:

Statement 4.2.3. *A non-negative function $f : \mathbb{R} \to \mathbb{R}$ is the density of some continuous random variable X if and only if f is Riemann integrable and*

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

Clearly, if X is continuous, then its density f_X satisfies this equation. We omit the converse direction.

4.3 Expected Value in the Continuous Case

Previously, we defined the expected value of non-negative random variables as:

$$\mathbb{E}(X) \overset{\text{def}}{=} \sup_{\substack{Z \text{ simple,} \\ Z \le X}} \mathbb{E}(Z), \qquad \text{where} \qquad \mathbb{E}(Z) = \sum_{k \in \mathrm{Ran}(X)} k \cdot \mathbb{P}(Z = k).$$

Note that this definition applies not only to discrete cases—it also yields a value for continuous random variables, even if it is not obvious how.

However, the requirement of non-negativity is somewhat restrictive. To remove it, we use the fact that expectation, as defined above, is additive.

We have already proved this for simple random variables.

[8] Provided that the density function is continuous at x. If f_{X^2} has a removable discontinuity at a point, then the function value at that point carries no meaning.

Statement 4.3.1. *Let X and Y be non-negative random variables. Then $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.*

This identity allows us to define the expected value for arbitrary (not necessarily non-negative or simple) random variables.

Definition 4.3.1. *Let X be any random variable. Define its **positive part** by $X^+ = \max(X, 0)$ and its **negative part** by $X^- = \max(-X, 0)$. Then X^+ and X^- are non-negative random variables, and $X = X^+ - X^-$. If either $\mathbb{E}(X^+) < \infty$ or $\mathbb{E}(X^-) < \infty$, define*

$$\mathbb{E}(X) \overset{\text{def}}{=} \mathbb{E}(X^+) - \mathbb{E}(X^-),$$

which may be a real number, $+\infty$, or $-\infty$. If both $\mathbb{E}(X^+) = \mathbb{E}(X^-) = \infty$, then the expected value is undefined.

While these definitions make sense for continuous random variables, they are not very practical for computation. The next result helps in this regard.

Statement 4.3.2. *Let X be a continuous random variable such that*

$$\int_{-\infty}^{\infty} |t| \cdot f_X(t)\,\mathrm{d}t < \infty.$$

Then

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} t \cdot f_X(t)\,\mathrm{d}t. \tag{1}$$

This condition is necessary to exclude cases where $\mathbb{E}(X)$ is not defined.

Definition 4.3.2. *A random variable X is said to be **uniformly distributed** on the interval (a, b) if its density function is*

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

Notation: $X \sim U(a; b)$.

This is indeed a valid density function, since

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_a^b \frac{1}{b-a}\,\mathrm{d}x = \left[\frac{x}{b-a}\right]_a^b = \frac{b-a}{b-a} = 1.$$

The expected value of a uniformly distributed random variable is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}x = \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2},$$

which also makes intuitive sense: the average value of a uniform distribution on an interval is the midpoint of that interval.
Note how similar the formulas for the discrete case and the continuous case look.
This is no coincidence—they both stem from the same general concept of expectation.
This is further reflected in the next result, where we can treat the two cases in parallel.

Statement 4.3.3 (Expected Value of a Transformed Variable)**.** *Let X be a random variable, and let $g : \mathbb{R} \to \mathbb{R}$ be a function. Suppose $\mathbb{E}(g(X))$ exists.*
If X is discrete, then

$$\mathbb{E}(g(X)) = \sum_{j=1}^{\infty} g(k_j) \cdot \mathbb{P}(X = k_j),$$

where $\mathrm{Ran}(X) = \{k_1, k_2, \dots\}$*. If X is continuous, then*

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x) \, \mathrm{d}x.$$

Example 4.3.1. Let X be a random variable with density $f_X : \mathbb{R} \to \mathbb{R}$ given by $f(x) = 2e^{-2x}$ for $x \in [0, \infty)$ and 0 otherwise. (Verify that this is indeed a valid density function.) Then

$$\mathbb{E}(e^X) = \int_{-\infty}^{\infty} e^x \cdot f_X(x) \, \mathrm{d}x = \int_0^{\infty} e^x \cdot 2e^{-2x} \, \mathrm{d}x = 2 \int_0^{\infty} e^{-x} \, \mathrm{d}x = 2.$$

5 Notable Distributions

In this lecture, we begin by discussing a paradox related to geometric probabilities and continuous random variables. Then we explore various probability distributions, both in discrete and continuous settings, applying concepts from the previous two lectures.

5.1 Bertrand's Paradox

Continuous random variables (and their distributions) can be described in several ways. This can be done using their density function or cumulative distribution function, but sometimes only indirect information is available. Bertrand's paradox[9] highlights how non-trivial it can be to determine a random variable's distribution in such cases.

Bertrand's Paradox: Select a chord of a circle at random. What is the probability that it is longer than the side of an equilateral triangle inscribed in the circle?
The paradox lies in the fact that there are valid solutions yielding different answers: $\frac{1}{3}$, $\frac{1}{2}$, and $\frac{1}{4}$. How is this possible? Let's look at the different approaches:

1. Select two points P and Q independently and uniformly at random on the circumference, and define the chord as the segment connecting them. First draw point P, then inscribe an equilateral triangle in the circle with vertex P. Clearly, the chord will be longer than the triangle's side if and only if Q lies between the other two vertices of the triangle. Thus, the desired probability is $\frac{1}{3}$.

2. Choose a radius of the circle uniformly at random, then select a point P uniformly at random along that radius (independently of the choice of the radius). The chord is defined as the line perpendicular to the radius at point P. First select the radius, then inscribe an equilateral triangle whose one side is perpendicular to the radius, and denote its intersection point with the radius by S. The chord will be longer than the triangle's side if and only if P lies closer to the center than S. Since S bisects the radius, the desired probability is $\frac{1}{2}$.

3. Select a point P uniformly at random in the interior of the circle. The chord is defined as the one for which P is the midpoint. There is always a unique such chord (except when P is the center, but that occurs with probability zero, so we can ignore it). Consider an inscribed equilateral triangle and its incircle. The chord will be longer than the triangle's side if and only if P lies within the smaller circle. From the previous case, we know that if the large circle has radius r, then the small one has radius $\frac{r}{2}$. Hence the desired probability is the ratio of the two areas: $\left(\frac{r}{2}\right)^2\pi \Big/ r^2\pi = \frac{1}{4}$.

[9] This is the same Joseph Bertrand as the one who proposed the box paradox in the first lecture.

The resolution of the paradox lies in the fact that the term “random chord” is not precisely defined. Depending on the method used to generate the chord, we obtain different distributions, and hence different probabilities.

5.2 The Memoryless Property

Among the notable distributions that often arise in examples are the (discrete) geometric distribution and the (continuous) exponential distribution. These two are closely related and will be discussed together.

Definition 5.2.1. *A random variable X is said to have a* ***geometric distribution*** *with parameter p (where $p \in (0,1)$) if*

$$\mathbb{P}(X = k) = (1-p)^{k-1}p$$

for all positive integers k. Notation: $X \sim \mathrm{Geo}(p)$.

Example 5.2.1. Suppose we try to thread a needle. Assume each attempt succeeds with probability 0.1, independently of previous attempts. Then the number of required trials follows a geometric distribution with parameter 0.1.

The expected value of a geometrically distributed random variable is:[10]

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} k \cdot \mathbb{P}(X = k) = \sum_{k=1}^{\infty} k \cdot (1-p)^{k-1}p =$$

$$= p\sum_{k=1}^{\infty}\sum_{i=1}^{k}(1-p)^{k-1} = p\sum_{i=1}^{\infty}\sum_{k=i}^{\infty}(1-p)^{k-1} = p\sum_{i=1}^{\infty}\frac{(1-p)^{i-1}}{1-(1-p)} = \sum_{j=0}^{\infty}(1-p)^j = \frac{1}{p}.$$

Definition 5.2.2. *A random variable Z is said to have an* ***exponential distribution*** *with parameter λ (where $\lambda > 0$) if*

$$f_Z(x) = \begin{cases} \lambda e^{-\lambda x} & \textit{if } x > 0, \\ 0 & \textit{otherwise,} \end{cases} \qquad \textit{that is,} \qquad F_Z(x) = \begin{cases} 1 - e^{-\lambda x} & \textit{if } x > 0, \\ 0 & \textit{otherwise.} \end{cases}$$

Notation: $Z \sim \mathrm{Exp}(\lambda)$.

This is indeed a density function: it is nonnegative and

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{0}^{\infty} \lambda e^{-\lambda x}\,\mathrm{d}x = \left[-e^{-\lambda x}\right]_0^{\infty} = 0 + 1 = 1.$$

[10] The series $\sum_{k=1}^{\infty} k \cdot (1-p)^{k-1}$ can be computed via differentiation of a power series.

The expected value of an exponentially distributed random variable is:

$$\mathbb{E}(Z) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_0^{\infty} \lambda x e^{-\lambda x}\,\mathrm{d}x =$$

$$= \big[x(-e^{-\lambda x})\big]_0^{\infty} - \int_0^{\infty} (-e^{-\lambda x})\,\mathrm{d}x = 0 + \int_0^{\infty} e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda}.$$

This distribution often arises in situations where one is waiting for a success in a stream of frequent, successive trials. The parameter λ (called the rate) indicates how many such "successes" occur on average per unit time.

A distribution is notable not only because it appears in many applications, but also because it has special properties. One such property, shared by the above two distributions, is *memorylessness*.

Definition 5.2.3. *A random variable X is said to be* ***memoryless*** *on a set $G \subseteq \mathbb{R}$ if for all $s, t \in G$,*

$$\mathbb{P}(X > t + s \mid X > s) = \mathbb{P}(X > t), \tag{2}$$

and $\mathbb{P}(X \in G) = 1$, i.e., X takes values in G with probability 1.

Thinking of X as the waiting time until an event occurs, the equation above means that if the event has not yet occurred by time s, the probability we have to wait at least t more time units is the same as if we started waiting from time 0. In other words, time does not affect the likelihood of the event occurring.

Question: Do such random variables even exist? If so, what are their distributions?

Statement 5.2.1. *Let X be a non-constant memoryless random variable on the set G.*

1. *If $G = \{1, 2, 3, \dots\}$, then X has a geometric distribution.*
2. *If $G = [0, \infty)$, then X has an exponential distribution.*

Thus, these two distributions are analogues of each other. They describe how long we have to wait for an event where the passage of time does not change the probability of occurrence. The difference is that in the first case, time is modeled discretely, and in the second case, continuously.

Proof. Expanding the definition of conditional probability, we see that equation is equivalent to:

$$\mathbb{P}(X > t + s) = \mathbb{P}(X > t)\mathbb{P}(X > s) \qquad (\forall s, t \in G).$$

First, let $G = \{1, 2, 3, \dots\}$, and let $p = \mathbb{P}(X = 1)$. For $s = 1$, we get for any positive integer t:

$$\mathbb{P}(X > t + 1) = \mathbb{P}(X > t)\mathbb{P}(X > 1) = \mathbb{P}(X > t) \cdot (1 - p) =$$

$$= \mathbb{P}(X > t-1) \cdot (1-p)^2 = \cdots = (1-p)^{t+1}.$$

From this, the probability mass function of X follows:

$$\mathbb{P}(X = t) = \mathbb{P}(X > t-1) - \mathbb{P}(X > t) = (1-p)^{t-1} - (1-p)^t =$$

$$= \big(1 - (1-p)\big)(1-p)^{t-1} = p(1-p)^{t-1},$$

which is the geometric distribution. There is only a problem if $p \in \{0, 1\}$. If $p = 0$, then $\mathbb{P}(X = t) = 0$ for all $t \in \{1, 2, \dots\}$, so these equations do not define a valid distribution. But by assumption, X must take values in G with probability 1, so this leads to a contradiction. Similarly, if $p = 1$, then X is constantly 1, contradicting the assumption that X is non-constant. Therefore, X must be geometrically distributed.
Now let $G = [0, \infty)$, and define $g(t) := \ln \mathbb{P}(X > t)$[11]. This is well-defined unless X is constantly 0, which we excluded. Taking logarithms in equation (2) yields:

$$g(t+s) = \ln \mathbb{P}(X > t+s) = \ln \mathbb{P}(X > t) + \ln \mathbb{P}(X > s) = g(t) + g(s),$$

so g is an additive function on the nonnegative real numbers. Furthermore, we know that g is decreasing, since the distribution function $t \mapsto F_X(t) = \mathbb{P}(X < t)$ is increasing, and so $t \mapsto \mathbb{P}(X > t)$ is decreasing. Therefore, g must be decreasing as well.

Lemma 5.2.1. *Let $g : [0, \infty) \to \mathbb{R}$ be a monotone decreasing function satisfying $g(t+s) = g(t) + g(s)$ for all $s, t \in [0, \infty)$. Then $g(t) = -\lambda t$, where $\lambda = -g(1) \geq 0$.*[12]

We omit the proof. By the lemma, we have $\ln \mathbb{P}(X > t) = g(t) = -\lambda t$, where $\lambda = -\ln \mathbb{P}(X > 1)$. Thus,

$$\mathbb{P}(X \leq t) = 1 - \mathbb{P}(X > t) = 1 - e^{-\lambda t},$$

where $\lambda \neq 0$, since otherwise $\mathbb{P}(X > t) = 1$ would hold for all t, which is impossible. What we really need is $\mathbb{P}(X < t)$, the distribution function. Note that the function $t \mapsto \mathbb{P}(X \leq t)$ is continuous, so

$$\mathbb{P}(X < t) = \lim_{s \nearrow t} \mathbb{P}(X \leq s) = \mathbb{P}(X \leq t) = 1 - e^{-\lambda t},$$

which is precisely the distribution function of the exponential distribution. □

The connection between the geometric and exponential distributions also manifests in the fact that the geometric distribution can be derived from the exponential. Conversely, the exponential distribution is the so-called limit distribution of the geometric distribution when we compress the possible occurrence times of the geometric variable into multiples of $\frac{1}{n}$, as $n \to \infty$ and $np \to \lambda$.

[11] Here ln denotes the natural logarithm.

[12] Without the monotonicity condition, the statement does not hold; some regularity assumption (e.g. continuity, boundedness, Lebesgue measurability, etc.) is needed.

Statement 5.2.2. *If X is exponentially distributed with parameter λ, then $\lceil X \rceil$ is geometrically distributed with parameter $1 - e^{-\lambda}$.*

Proof. Clearly, $\mathbb{P}(\lceil X \rceil = 0) = \mathbb{P}(-1 < X \leq 0) = 0$. For any positive integer k,

$$\mathbb{P}(\lceil X \rceil = k) = \mathbb{P}(k - 1 < X \leq k) = F_X(k) - F_X(k-1) =$$

$$= (1 - e^{-\lambda k}) - (1 - e^{-\lambda(k-1)}) = e^{-\lambda(k-1)} - e^{-\lambda k} = e^{-\lambda(k-1)}(1 - e^{-\lambda}),$$

which with the notation $p = 1 - e^{-\lambda}$ equals $(1-p)^{k-1}p$, that is, $\lceil X \rceil$ is geometrically distributed with parameter $1 - e^{-\lambda}$. □

5.3 Poisson Distribution

The Poisson distribution is a widely used discrete distribution. Intuitively, it arises when we consider the number of occurrences of many independent events, each having a very small probability of happening.

Definition 5.3.1. *A random variable X is said to have a **Poisson distribution** with parameter $\lambda > 0$ if*

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

for all non-negative integers k. Notation: $X \sim \mathrm{Pois}(\lambda)$.

This is indeed a valid probability distribution since the total probability is

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

The parameter of the Poisson distribution has a clear interpretation, since

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \lambda \cdot \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} e^{-\lambda} = \lambda,$$

meaning that the expected value of the variable is precisely the parameter λ (note that this is not the case for either the geometric or the exponential distribution).

Example 5.3.1. A stuntman gets injured on average 2 times per year. What is the probability that they get injured 4 times this year? Since a stuntman has many opportunities to get injured, and these can be considered independent, the Poisson distribution is a good approximation for the number of injuries. Let Y denote this number. According to the problem, $\lambda = 2$, so

$$\mathbb{P}(Y = 4) = \frac{2^4}{4!} e^{-2} = \frac{2}{3} e^{-2} \approx 0.0902.$$

The fact that the number of occurrences of many low-probability events approximates a Poisson distribution can also be proven. If there are n events, each occurring independently with probability p, then the number of occurrences X follows a binomial distribution $B(n;p)$. The expected number of occurrences is $n \cdot p$, which we denote by λ. The following proposition shows that for large n, the distribution of $B(n;p)$ approximates $\mathrm{Pois}(\lambda)$.

Statement 5.3.1. *Let n be a positive integer, $\lambda \in (0,\infty)$, and define $p_n := \frac{\lambda}{n}$. Then*

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

for any $k \in \{0,1,2,\dots\}$.

Proof. Fix k, then for any n:

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} =$$

$$= \frac{n!}{(n-k)!n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1-\frac{\lambda}{n}\right)^n \cdot \left(1-\frac{\lambda}{n}\right)^{-k}.$$

Here, $\frac{\lambda^k}{k!}$ remains constant as $n \to \infty$, and

$$\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}, \qquad \left(1-\frac{\lambda}{n}\right)^{-k} \to 1.$$

So we just need to handle the first factor:

$$\frac{n!}{(n-k)!n^k} = \frac{n(n-1)\dots(n-k+1)}{n^k},$$

which is a product of k terms, each of the form $\frac{n-i}{n}$, and each tends to 1. Since k is fixed and only n tends to infinity, the product also tends to 1. □

Example 5.3.2. In a Hungarian literature exam, the probability of having exactly 3 typos is twice the probability of having exactly 1 typo (this example is not representative). Assume that typos occur independently and with equal probability. What is the probability that the paper contains no typos?
Let X be the number of typos in a paper. Since only a finite number of typos can occur, the assumptions (independence, identical probability) would suggest that X is binomially distributed. However, both the maximum number of typos and the probability of a single typo are difficult to estimate (especially given the available information).

Instead, we can apply the above result, which states that the distribution of X can be approximated by a Poisson distribution:

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad (k \in \mathbb{N}).$$

According to the given condition,

$$2 = \frac{\mathbb{P}(X = 3)}{\mathbb{P}(X = 1)} = \frac{\lambda^3}{3!} e^{-\lambda} \bigg/ \frac{\lambda}{1!} e^{-\lambda} = \frac{\lambda^2}{6},$$

so $\lambda = 2\sqrt{3}$. This means there are, on average, $2\sqrt{3}$ typos in such a paper. Now the solution is straightforward:

$$\mathbb{P}(X = 0) = \frac{(2\sqrt{3})^0}{0!} e^{-2\sqrt{3}} = e^{-2\sqrt{3}} \approx 0.0313.$$

6 Relations between Random Variables

So far we have only studied random variables individually. In such cases, it was sufficient to consider their distribution, i.e., in the discrete case, the sequence of probabilities of the form $\mathbb{P}(X = k)$, and in the continuous case, the distribution function F_X or the density function f_X. The distribution conveyed all essential information about the random variable.

However, one should not confuse the distribution with the random variable itself: even though both the number of heads and the number of tails in 100 coin tosses follow a $B(100; \frac{1}{2})$ binomial distribution, we clearly cannot say that we always get the same number of heads as tails. This distinction becomes especially important when we consider multiple random variables simultaneously.

In this chapter, we examine the independence of two random variables and the degree of their linear relationship.

6.1 Independence

We have already introduced the concept of independence for events in Lecture 2: events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Using this, we now define the independence of random variables.

Definition 6.1.1. *Let X and Y be random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that X and Y are **independent** if, for all $x, y \in \mathbb{R}$, the events $\{X < x\}$ and $\{Y < y\}$ are independent.*

We need to refer to the probability space because it may happen that $X : \Omega_1 \to \mathbb{R}$ and $Y : \Omega_2 \to \mathbb{R}$, meaning they are defined on different probability spaces. In such a case, we cannot speak of independence between X and Y, since they "live in different worlds".

Example 6.1.1. The result of a dice roll and today's amount of precipitation are intuitively independent, and they are also independent in the sense of the above definition, as we already noted when discussing event independence.

However, independence is not always so obvious. For example, let Z be a uniformly distributed random variable on the set $\{1, 2, 3, \ldots, 11, 12\}$, and let X be the remainder of Z modulo 3, and Y the remainder modulo 4. Then X and Y are independent.

How can we verify the independence of two random variables? In this lecture, we focus on the discrete case. The following statement provides a method for checking independence.

Statement 6.1.1. *Two discrete random variables are independent if and only if for all $x, y \in \mathbb{R}$, the events $\{X = x\}$ and $\{Y = y\}$ are independent, i.e.,*

$$\mathbb{P}(\{X = x\} \cap \{Y = y\}) = \mathbb{P}(X = x)\mathbb{P}(Y = y).$$

Remark 6.1.1. The definition also implies that every event expressible using X and Y is independent, for example, $\{X = x\}$ and $\{1 \leq Y \leq 5\}$ are independent. In general, we can consider the **σ-algebra generated by** X, that is, the smallest σ-algebra—denoted $\sigma(X)$—containing the events $\{X < x\}$ for all $x \in \mathbb{R}$. Then X and Y are independent if and only if every pair of events $A \in \sigma(X)$ and $B \in \sigma(Y)$ are independent.

A key difference from the case of events is that for events, it takes the same effort to verify independence as to refute it—we simply compute $\mathbb{P}(A \cap B)$ and compare it to $\mathbb{P}(A)\mathbb{P}(B)$. However, for random variables, it is usually much easier to disprove independence than to prove it: it suffices to find a pair of events $\{X = x\}$ and $\{Y = y\}$ that are not independent. Why is independence useful? For example, because it helps in calculating expected values.

Statement 6.1.2. *If X and Y are independent random variables and $\mathbb{E}(XY)$, $\mathbb{E}(X)$, and $\mathbb{E}(Y)$ exist, then*

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Proof. We prove the statement for simple random variables X and Y. The general case follows by approximation, which we omit here.
First, assume that $X = \mathbf{1}_A$ and $Y = \mathbf{1}_B$ for some events A and B. Then

$$\mathbb{E}(\mathbf{1}_A \mathbf{1}_B) = \mathbb{E}(\mathbf{1}_{A\cap B}) = \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) = \mathbb{E}(\mathbf{1}_A)\mathbb{E}(\mathbf{1}_B),$$

so the statement holds in this special case.
Now consider general simple random variables. Then we can write X and Y as linear combinations of indicator variables:

$$X = \sum_{k \in \mathrm{Ran}(X)} k \cdot \mathbf{1}_{\{X=k\}}, \qquad Y = \sum_{l \in \mathrm{Ran}(Y)} l \cdot \mathbf{1}_{\{Y=l\}}.$$

Using linearity of expectation and the result above, we obtain:

$$\mathbb{E}(XY) = \mathbb{E}\left(\sum_k k \cdot \mathbf{1}_{\{X=k\}} \sum_l l \cdot \mathbf{1}_{\{Y=l\}}\right) = \sum_k \sum_l k \cdot l \cdot \mathbb{E}(\mathbf{1}_{\{X=k\}}\mathbf{1}_{\{Y=l\}}),$$

$$= \sum_k \sum_l k \cdot l \cdot \mathbb{E}(\mathbf{1}_{\{X=k\}}) \cdot \mathbb{E}(\mathbf{1}_{\{Y=l\}}) = \mathbb{E}(X) \cdot \mathbb{E}(Y),$$

as claimed. □

Remark 6.1.2. One might wonder why we do not use the equation $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ as a definition for independence. The reason is that this property is weaker: it does not imply independence of the random variables. However, the following holds: if for all nonnegative measurable functions f and g, we have $\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y))$, then X and Y are independent.

6.2 Joint Discrete Distributions

In the case of discrete random variables, in order to examine their independence, we need the probabilities $\mathbb{P}(X = k,\ Y = l)$ (i.e., $\mathbb{P}(\{X = k\} \cap \{Y = l\})$), which describe the so-called **joint distribution** of the variables. (We will discuss the continuous case in Lecture 8.)

Example 6.2.1. Let X and Y be discrete random variables with $\mathrm{Ran}(X) = \{2, 3, 5\}$ and $\mathrm{Ran}(Y) = \{0, 1, 2\}$, and suppose the joint probabilities $\mathbb{P}(X = k,\ Y = l)$ are given in the following table. Are X and Y independent? What is the value of $\mathbb{E}(XY)$?

X Y	2	3	5
0	0.05	0.15	0.10
1	0.10	0.20	0.10
2	0.05	0.20	0.05

A joint distribution given in such a table is a valid joint distribution of two random variables precisely when all entries are non-negative and their **sum equals 1**. We can verify that this holds in the example above.
To determine independence, we also need the marginal distributions $\mathbb{P}(X = k)$ and $\mathbb{P}(Y = l)$.

Definition 6.2.1. *Let X and Y be simple (i.e., discrete and finite-range) random variables. If the joint distribution $\mathbb{P}(X = k,\ Y = l)$ is given for all $k \in \mathrm{Ran}(X)$ and $l \in \mathrm{Ran}(Y)$, then the distributions of X and Y are called the* ***marginal distributions*** *of the joint distribution.*

The marginal distributions can be computed using the additivity of probability:

$$\mathbb{P}(X = k) = \sum_{l \in \mathrm{Ran}(Y)} \mathbb{P}(X = k,\ Y = l), \qquad \mathbb{P}(Y = l) = \sum_{k \in \mathrm{Ran}(X)} \mathbb{P}(X = k,\ Y = l),$$

that is, the row and column sums of the table give the distributions of X and Y. In our example:

$$\mathbb{P}(X = 2) = 0.20, \qquad \mathbb{P}(X = 3) = 0.55, \qquad \mathbb{P}(X = 5) = 0.25,$$

$$\mathbb{P}(Y = 0) = 0.30, \qquad \mathbb{P}(Y = 1) = 0.40, \qquad \mathbb{P}(Y = 2) = 0.30.$$

Therefore, by the definition of independence, X and Y are not independent, since for instance

$$\mathbb{P}(X = 5,\ Y = 0) = 0.10 \neq \mathbb{P}(X = 5) \cdot \mathbb{P}(Y = 0) = 0.25 \cdot 0.30 = 0.075.$$

Let us also compute the value of $\mathbb{E}(XY)$ for X and Y in this example. No new definitions are needed, since XY is itself a random variable with value set $\{k \cdot l \mid k \in \operatorname{Ran}(X),\ l \in \operatorname{Ran}(Y)\}$. Thus,

$$\mathbb{E}(XY) = \sum_{m \in \operatorname{Ran}(XY)} m \cdot \mathbb{P}\Big(\bigcup_{\substack{k \in \operatorname{Ran}(X) \\ l \in \operatorname{Ran}(Y) \\ k \cdot l = m}} \{X = k,\ Y = l\}\Big) = \sum_{k \in \operatorname{Ran}(X)} \sum_{l \in \operatorname{Ran}(Y)} k \cdot l \cdot \mathbb{P}(X = k,\ Y = l)$$

$$= 0 \cdot 0.05 + 0 \cdot 0.15 + 0 \cdot 0.10 + 2 \cdot 0.10 + 3 \cdot 0.20 + 5 \cdot 0.10 + 4 \cdot 0.05 + 6 \cdot 0.20 + 10 \cdot 0.05 = 3.2.$$

Note that although the variables are not independent, the expectation $\mathbb{E}(XY)$ can still be computed without issue.

6.3 Covariance

As we saw in the example above, even if two random variables are not independent, the degree of their dependence can still be low (i.e., intuitively, the products $\mathbb{P}(X = k)\mathbb{P}(Y = l)$ are close to the actual joint probabilities $\mathbb{P}(X = k,\ Y = l)$). How can we quantify the strength of dependence between random variables? There are several possibilities for this and we start with the concept of covariance.

Definition 6.3.1. *The **covariance** of random variables X and Y is defined as*

$$\operatorname{cov}(X, Y) \overset{\text{def}}{=} \mathbb{E}\big((X - \mathbb{E}X)(Y - \mathbb{E}Y)\big),$$

provided that the expectations exist and are finite.

Statement 6.3.1. *If $\operatorname{cov}(X, Y)$ is well-defined, then*

$$\operatorname{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Proof. Expanding the definition:

$$\mathbb{E}\big((X - \mathbb{E}X)(Y - \mathbb{E}Y)\big) = \mathbb{E}(XY) - \mathbb{E}\Big(\mathbb{E}(X)Y\Big) - \mathbb{E}\Big(X\mathbb{E}(Y)\Big) + \mathbb{E}\Big(\mathbb{E}(X)\mathbb{E}(Y)\Big) =$$

$$= \mathbb{E}(XY) + (-1 - 1 + 1)\mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y),$$

which proves the claim. □

Corollary 6.3.1. *Let X and Y be random variables for which $\operatorname{cov}(X, Y)$ is defined.*

1. *If Y is constant, then $\operatorname{cov}(X, Y) = 0$.*
2. *If X and Y are independent, then $\operatorname{cov}(X, Y) = 0$.*
3. *$\operatorname{cov}(X, Y) = 0$ does not imply that X and Y are independent.*

Proof. Let $Y = c \in \mathbb{R}$ be constant. Then by the previous result:

$$\operatorname{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(Xc) - \mathbb{E}(X)c = 0.$$

For the second part, we apply Proposition 6.1.2 from the previous subsection, so $\operatorname{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$.
For the third part, let $\operatorname{Ran}(X) = \{-1, 0, 1\}$ with probabilities $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively. Define $Y = |X|$. Then

$$\operatorname{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0 - 0 \cdot \frac{1}{2} = 0,$$

but X and Y are not independent, since

$$\mathbb{P}(X=0)\mathbb{P}(Y=1) = \frac{1}{2} \cdot \frac{1}{2}, \quad \text{while} \quad \mathbb{P}(X=0, Y=1) = 0.$$

□

Example 6.3.1.

1. We have seen that $\mathbb{E}(XY) = 3.2$. We can compute $\mathbb{E}(X) = 3.8$ and $\mathbb{E}(Y) = 1$, so:

$$\operatorname{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 3.2 - 3.8 \cdot 1 = -0.6.$$

2. Let X be uniformly distributed on $\{1, 2, \ldots, 10\}$, and let Y be uniformly distributed on $\{1, -1\}$, independent of X. Then:

$$\operatorname{cov}\big(X,\, 0.9 \cdot X + 0.1 \cdot Y\big) = \mathbb{E}(0.9 \cdot X^2 + 0.1 \cdot XY) - \mathbb{E}(X)\mathbb{E}(0.9 \cdot X + 0.1 \cdot Y) =$$

$$= 0.9 \cdot \mathbb{E}(X^2) + 0.1 \cdot \mathbb{E}(XY) - 0.9 \cdot \mathbb{E}(X)^2 - 0.1 \cdot \mathbb{E}(XY)$$

$$= 0.9 \sum_{k=1}^{10} k^2 \cdot \frac{1}{10} - 0.9 \cdot \left(\frac{11}{2}\right)^2 \approx 7.425.$$

This shows how the additivity of expectation can simplify computing covariance.

Remark 6.3.1. If zero covariance does not characterize independence, why do we consider it at all? The reason is mainly that the covariance is symmetric and bilinear:

$$\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X) \quad \text{and} \quad \operatorname{cov}(X, aY + bZ) = a \cdot \operatorname{cov}(X,Y) + b \cdot \operatorname{cov}(X,Z) \quad (a, b \in \mathbb{R}),$$

whenever the covariances are defined. Thus, covariance is analogous to the scalar product of vectors.

6.4 Variance and Standard Deviation

The special case of covariance when $Y = X$ gives rise to the following:

Definition 6.4.1. *The* ***variance*** *of a random variable X is defined as:*

$$\operatorname{cov}(X, X) = \mathbb{E}\big((X - \mathbb{E}X)^2\big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

Notation: $\mathbb{D}^2(X)$ (or alternatively, $\operatorname{Var}(X)$). The square root of the variance is called the ***standard deviation****, denoted $\mathbb{D}(X)$.*

Remark 6.4.1. In other words, the variance measures the average squared deviation of X from its mean. If we think of covariance as analogous to the scalar product of vectors, then variance corresponds to the squared length, and standard deviation to the length of a vector.

Note that variance does not measure ”self-dependence” but rather the spread of X's values around the mean. Although we could define alternative measures of spread (e.g., $\mathbb{E}(|X - \mathbb{E}X|)$), variance is widely used partly due to the following result:

Statement 6.4.1. *If X and Y are independent, then:*

$$\mathbb{D}^2(X + Y) = \mathbb{D}^2(X) + \mathbb{D}^2(Y).$$

Proof. Expanding the variance:

$$\mathbb{D}^2(X + Y) = \mathbb{E}((X + Y)^2) - (\mathbb{E}(X + Y))^2 =$$

$$= \mathbb{E}(X^2) + \mathbb{E}(Y^2) + 2\mathbb{E}(XY) - \mathbb{E}(X)^2 - \mathbb{E}(Y)^2 - 2\mathbb{E}(X)\mathbb{E}(Y)$$
$$= \mathbb{D}^2(X) + \mathbb{D}^2(Y) + 2\operatorname{cov}(X, Y).$$

Since X and Y are independent, $\operatorname{cov}(X, Y) = 0$, which completes the proof. □

Remark 6.4.2. From the proof, we see that even without independence, we have:

$$\mathbb{D}^2(X + Y) = \mathbb{D}^2(X) + \mathbb{D}^2(Y) + 2\operatorname{cov}(X, Y).$$

Other basic properties of variance:

Statement 6.4.2. *Assume that $\mathbb{D}(X)$ exists and is finite. Then for any $c \in \mathbb{R}$:*

$$\mathbb{D}(X + c) = \mathbb{D}(X), \qquad \mathbb{D}(cX) = |c|\mathbb{D}(X),$$

i.e., standard deviation is translation-invariant and absolutely homogeneous.

Proof. Expanding the definitions:

$$\mathbb{D}^2(X+c) = \mathbb{E}\big((X+c-\mathbb{E}(X+c))^2\big) = \mathbb{E}((X-\mathbb{E}X)^2) = \mathbb{D}^2(X),$$

$$\mathbb{D}^2(cX) = \mathbb{E}\big((cX-\mathbb{E}(cX))^2\big) = \mathbb{E}(c^2(X-\mathbb{E}X)^2) = c^2\mathbb{D}^2(X),$$

taking square roots yields the claim. □

Example 6.4.1.

1. Let K be uniformly distributed on $\{1,2,3,4,5,6\}$. Then:

$$\mathbb{E}(K^2) = \frac{91}{6}, \qquad \mathbb{E}(K) = \frac{7}{2},$$

$$\mathbb{D}^2(K) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12} \approx 2.9167, \quad \mathbb{D}(K) \approx 1.7078.$$

2. Let $X = \mathbf{1}_A$ be an indicator variable with $\mathbb{P}(A) = p$. Then:

$$\mathbb{D}^2(\mathbf{1}_A) = \mathbb{E}(\mathbf{1}_A^2) - \mathbb{E}(\mathbf{1}_A)^2 = p - p^2 = p(1-p).$$

3. Let $X \sim B(n;p)$. Since $X = \mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_n}$, where the A_i are independent with probability p:

$$\mathbb{D}^2(X) = \mathbb{D}^2(\mathbf{1}_{A_1}) + \cdots + \mathbb{D}^2(\mathbf{1}_{A_n}) = np(1-p).$$

4. Let $T \sim \text{Geo}(p)$. Then:

$$\mathbb{D}^2(T) = \mathbb{E}(T^2) - \mathbb{E}(T)^2 = \sum_{k=1}^{\infty} k^2(1-p)^{k-1}p - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}.$$

5. Let $Y \sim \text{Pois}(\lambda)$. Then:

$$\mathbb{D}^2(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \mathbb{E}(Y^2 - Y) + \mathbb{E}(Y) - \mathbb{E}(Y)^2 =$$

$$= \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} e^{-\lambda} + \lambda - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

6. Let $Z \sim \text{Exp}(\lambda)$. Then:

$$\mathbb{D}(Z) = \frac{1}{\lambda}.$$

6.5 Correlation

We previously noted that covariance can indicate dependence between variables. But how should we interpret a nonzero covariance, such as $\mathrm{cov}(X, Y) = -0.6$? How strongly are X and Y related?
This cannot be answered from the covariance alone; we use a derived quantity:

Definition 6.5.1. *Let X and Y be random variables. If $\mathrm{cov}(X, Y)$, $\mathbb{D}(X)$, and $\mathbb{D}(Y)$ are defined, then their **correlation** is:*

$$\mathrm{corr}(X, Y) \overset{\text{def}}{=} \frac{\mathrm{cov}(X, Y)}{\mathbb{D}(X)\mathbb{D}(Y)}.$$

It can be shown that $-1 \leq \mathrm{corr}(X, Y) \leq 1$. In the extreme cases, X and Y are perfectly linearly related:

Statement 6.5.1. *Let X and Y be random variables. If $\mathrm{corr}(X, Y) \in \{1, -1\}$, then there exist real numbers a and b such that $Y = aX + b$ almost surely, and the sign of a agrees with the sign of the correlation.*

Example 6.5.1. For the discrete joint distribution from Example 6.2.1,

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\mathbb{D}(X)\mathbb{D}(Y)} = \frac{-0.6}{\sqrt{9.2} \cdot \sqrt{0.6}} \approx -0.2554.$$

This suggests that X and Y tend to deviate in opposite directions from their respective means, though the linear relationship is relatively weak.

As with covariance, zero correlation does not imply independence. Correlation measures the *linear* relationship between two variables. Thus, even if two variables are related, if their relationship is nonlinear, correlation may not detect it. For instance, there exists a random variable X such that $\mathrm{corr}(X, X^2) = 0$.

7 Continuous Joint Distributions and Convolution

When introducing the concept of a random variable, we first considered the discrete case, and then turned to continuous random variables. For joint distributions, we take a similar path: earlier we examined how to describe the joint behavior and independence of two discrete random variables; now we do the same for continuous random variables.
At the end of this chapter, we add one more method to our toolkit for determining the distribution of the sum of two independent random variables. This distribution is called the **convolution** of the original two distributions. At first glance, the task may not seem complicated—after all, addition should not be too difficult. In the third subsection, we will first see why it actually is, and then why it is not as bad after all.

7.1 Random Vectors

It is common to encounter situations where we must speak about several random variables at once. This may be because we want to examine the relationship between two random quantities (e.g., the running time and memory usage of a random program), or because our variable cannot be described by a single parameter (e.g., the position of a drone in space). In both cases, it is natural to treat our random variables as a vector.

Definition 7.1.1. *Let $X_1, \dots, X_n$ be random variables for some positive integer n. Then the function*

$$\underline{X} \overset{\text{def}}{=} (X_1, \dots, X_n) : \Omega \to \mathbb{R}^n$$

*is called a **random vector**. The **(joint) distribution function** of $\underline{X}$ is the scalar-valued function*

$$F_{\underline{X}} : \mathbb{R}^n \to [0,1], \quad F_{\underline{X}}(x_1, \dots, x_n) = \mathbb{P}(X_1 < x_1, \dots, X_n < x_n).$$

Example 7.1.1. We are chatting separately with two of our friends, Aladár and Béla, via some messaging application. If there is a match on TV, then Aladár replies in twice the usual time (because he is watching it), while Béla replies in half the time (perhaps because he is complaining that he cannot hear his own thoughts over the shouting from the neighbours). Suppose that in the absence of a match, they reply independently according to an exponential distribution with parameter $\lambda = 6$.[13] The probability that there is a match is $\frac{1}{5}$. What is their joint distribution of reply times?
Let X denote Aladár's reply time, Y Béla's, and M the event that there is a match. (Here X is a scalar random variable, not a vector.) Let Z be the random variable

$$Z = \begin{cases} 2 & \text{if there is a match,} \\ 1 & \text{otherwise.} \end{cases}$$

[13] Whether it is realistic to assume an exponential distribution depends on the circumstances.

By assumption, there exist U and V such that

$X = U{\cdot}Z, \quad Y = V/Z, \qquad$ where $U, V \sim \mathrm{Exp}(6)$ are independent of each other and of M.

If $x, y > 0$, then by the law of total probability

$$F_{(X,Y)}(x,y) = \mathbb{P}(X < x, Y < y) = \mathbb{P}(U \cdot Z < x,\, V/Z < y) =$$

$$= \mathbb{P}(U \cdot Z < x, V/Z < y \mid M)\mathbb{P}(M) + \mathbb{P}(U \cdot Z < x, V/Z < y \mid \overline{M})\mathbb{P}(\overline{M}).$$

Here we can use that M and $\overline{M}$ determine the value of Z, and U and V are independent, hence

$$\begin{aligned} &= \mathbb{P}(U \cdot 2 < x, V/2 < y)\frac{1}{5} + \mathbb{P}(U < x, V < y)\frac{4}{5} \\ &= \mathbb{P}\Big(U < \frac{x}{2}\Big)\,\mathbb{P}(V < 2y)\frac{1}{5} + \mathbb{P}(U < x)\mathbb{P}(V < y)\frac{4}{5} \\ &= (1 - e^{-6\frac{x}{2}})(1 - e^{-6\cdot 2y})\frac{1}{5} + (1 - e^{-6x})(1 - e^{-6y})\frac{4}{5}. \end{aligned}$$

If $x \le 0$ or $y \le 0$, then $F_{(X,Y)}(x,y) = 0$.

In the continuous case—just as in the univariate case—we define the joint probability density function of a random vector.

Definition 7.1.2. *Let $\underline{X} = (X_1, \ldots, X_n)$ be a random vector. A function $f_{\underline{X}} : \mathbb{R}^n \to [0, \infty)$ is the **(joint) probability density function** of $\underline{X}$ if $f_{\underline{X}}$ has an improper Riemann integral over $\mathbb{R}^n$ and*

$$\int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f_{\underline{X}}(z_1, \ldots, z_n)\,\mathrm{d}z_1 \ldots \mathrm{d}z_n = F_{\underline{X}}(x_1, \ldots, x_n)$$

*for all $x_1, \ldots, x_n \in \mathbb{R}$. We call $\underline{X}$ **continuous** if it has a joint probability density function.*

Statement 7.1.1. *Let $\underline{X} = (X_1, \ldots, X_n)$ be a random vector. If $\underline{X}$ is continuous, then the following function is the probability density function of $\underline{X}$:*

$$f_{\underline{X}}(x_1, \ldots, x_n) = \begin{cases} \partial_{x_1} \ldots \partial_{x_n} F_{\underline{X}}(x_1, \ldots, x_n) & \textit{if it exists,} \\ 0 & \textit{otherwise.} \end{cases}$$

(The partial derivatives can be taken in any order.)

Remark 7.1.1. Unlike in the univariate case, here it is an assumption (not a consequence) that the probability density function exists (i.e., that $\underline{X}$ is continuous). For example, if $X_1 = X_2$, where X_1 is uniformly distributed on the interval $[0, 1]$, then $F_{(X_1,X_2)}$ is continuous and twice continuously differentiable except along certain lines, yet $\partial_{x_1}\partial_{x_2}F_{(X_1,X_2)}(x_1, x_2) = 0$, which clearly cannot be a density function. Thus, the vector (X_1, X_2) has no joint probability density function in the case $X_1 = X_2$.

As in the univariate case, multivariate probability density functions can also be characterised.

Statement 7.1.2. *Let $f : \mathbb{R}^n \to [0, \infty)$. Then*

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) \, \mathrm{d}x_1 \dots \mathrm{d}x_n = 1$$

if and only if there exists a random vector $\underline{X} = (X_1, \dots, X_n)$ whose probability density function is f.[14]

What did the joint distribution look like in the discrete case? There we did not speak of a joint density function; instead we simply had a table of values $\mathbb{P}(X = x, Y = y)$ representing the joint distribution. The analogue of the previous statement is that the sum of the numbers in the table is 1.
What is the use of the joint density function? In the univariate case, it was used, among other things, to compute probabilities of the form $\mathbb{P}(a < Z < b)$. The same role is preserved in the multivariate case:

Statement 7.1.3. *Let $H \subseteq \mathbb{R}^n$ be a Jordan-measurable set and $\underline{X} = (X_1, \dots, X_n)$ a random vector. Then*

$$\mathbb{P}(\underline{X} \in H) = \int_H f_{\underline{X}}(\underline{x}) \, \mathrm{d}\underline{x}.$$

Example 7.1.2. Returning to the above example, what is the probability that Aladár replies before Béla? By the first proposition, we can obtain the density function by differentiating with respect to x and y, provided it exists:

$$f_{(X,Y)}(x, y) = \begin{cases} 3e^{-3x} \cdot 12e^{-12y}\frac{1}{5} + 6e^{-6x} \cdot 6e^{-6y}\frac{4}{5} & \text{if } x, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

It can be checked by integration that this is indeed the density function of (X, Y), but we omit this. The desired probability is then

$$\mathbb{P}(X < Y) = \int_{\{x<y\}} f_{(X,Y)}(x, y) \, \mathrm{d}x \, \mathrm{d}y = \int_0^\infty \int_0^y \left(3e^{-3x} \cdot 12e^{-12y}\frac{1}{5} + 6e^{-6x} \cdot 6e^{-6y}\frac{4}{5}\right) \mathrm{d}x \, \mathrm{d}y,$$

which, after some calculation, turns out to be 0.44 (exactly, not just rounded).

One final missing concept is the marginal distribution.

Definition 7.1.3. *If $\underline{X} = (X_1, \dots, X_n)$ is a random vector, then the distribution of the random variable X_i is called the* ***i-th marginal distribution*** *(or* marginal *for short) of $\underline{X}$.*

[14] Joint distribution functions can also be characterised, but their description is more complicated than in the univariate case.

Statement 7.1.4. *If $\underline{X} = (X_1, \dots, X_n)$ is a continuous random vector, then each X_i is also continuous, and its density function is*

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x_1, \dots, x_n) \,\mathrm{d}x_1 \dots \mathrm{d}x_{i-1} \,\mathrm{d}x_{i+1} \dots \mathrm{d}x_n \qquad (\forall x_i \in \mathbb{R}),$$

that is, the integral of the joint density over all variables except x_i.

Example 7.1.3. In the above example, a marginal of (X, Y) is, for instance, the distribution of Y, i.e.,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, y) \,\mathrm{d}x = \int_0^{\infty} \Big(3e^{-3x} \cdot 12e^{-12y}\frac{1}{5} + 6e^{-6x} \cdot 6e^{-6y}\frac{4}{5}\Big) \,\mathrm{d}x$$

$$= \frac{12}{5}e^{-12y}\big[-e^{-3x}\big]_0^{\infty} + \frac{24}{5}e^{-6y}\big[-e^{-6x}\big]_0^{\infty} = \frac{12}{5}e^{-12y} + \frac{24}{5}e^{-6y},$$

for $y > 0$, and 0 otherwise. The distribution of Y is a so-called mixture of exponential distributions.

Remark 7.1.2. It makes sense to talk about random vectors and their joint distribution functions even if $X_1, \dots, X_n$ are discrete. This is not the case for density functions. Moreover, even if X_1 and X_2 are each continuous, they may still fail to have a joint density function. See the example in the previous remark.

7.2 Independence of Random Vectors

A special case in the relationship of random variables is when they are independent. We already defined the independence of two random variables in Chapter 6. Although the discussion there mainly concerned discrete distributions, the definition makes sense for non-discrete variables as well. This is generalised by the following definition.

Definition 7.2.1. *The random variables $X_1, \dots, X_n$ are* ***(jointly) independent*** *if the events*

$$\{X_1 < x_1\}, \dots, \{X_n < x_n\}$$

are independent for all $x_1, \dots, x_n \in \mathbb{R}$.

From now on we will omit the word "jointly". If we speak of some other type of independence (such as pairwise independence), we will state it explicitly.

How can we check the joint independence of n random variables? In the discrete case, we examined the probabilities of events $\{X_i = x_i\}$. When the product $\mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n)$ matched the probability of the intersection for all parameter choices, this proved independence. In the continuous case, this approach does not work. Instead, we can say:

Statement 7.2.1. *The random variables $X_1, \ldots, X_n$ are independent if and only if*

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

for all $x_1, \ldots, x_n \in \mathbb{R}$.

The above statement, while useful for continuous random variables, applies more generally: just as the distribution function is defined for any random variable, the theorem holds for arbitrary $X_1, \ldots, X_n$.
Returning to the continuous case: sometimes we do not have the distribution function at hand, but we do know the joint density function. Then we can use the following.

Statement 7.2.2. *Let $X_1, \ldots, X_n$ be continuous random variables. They are independent if and only if $(X_1, \ldots, X_n)$ is a continuous random vector and*

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

for all $x_1, \ldots, x_n \in \mathbb{R}$.

In short, the joint density function of independent random variables factorises. In particular, in such a case the joint density function exists.

Example 7.2.1. In the setting of the previous example, is Aladár's reply time independent of Béla's? We can check this using the previous proposition. For $x, y > 0$, we have

$$f_{(X,Y)}(x, y) = 3e^{-3x} \cdot 12e^{-12y}\frac{1}{5} + 6e^{-6x} \cdot 6e^{-6y}\frac{4}{5},$$

and, similarly to the computation of f_Y,

$$f_X(x) = \frac{3}{5}e^{-3x} + \frac{24}{5}e^{-6x}.$$

Thus X and Y are not independent, because

$$f_X(x) \cdot f_Y(y) = \Big(\frac{3}{5}e^{-3x} + \frac{24}{5}e^{-6x}\Big)\Big(\frac{12}{5}e^{-12y} + \frac{24}{5}e^{-6y}\Big)$$

is not the same as

$$f_{(X,Y)}(x, y) = 3e^{-3x} \cdot 12e^{-12y}\frac{1}{5} + 6e^{-6x} \cdot 6e^{-6y}\frac{4}{5},$$

except for special x, y. Independence would require $f_X(x) \cdot f_Y(y) = f_{(X,Y)}(x, y)$ for all x, y.

7.3 Convolution

Since we introduced the concept of expectation, its most frequently used property has been additivity, i.e., $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$. This holds whether X and Y are independent or not. We might thus think that determining the distribution of $X+Y$ should be straightforward, at least when X and Y are independent. Let us look at some examples to see why this is not the case.

Example 7.3.1.

1. Let $X \sim B(n;p)$ be binomially distributed. We know that X is the sum of independent indicator random variables, each taking only values 1 or 0. Even though the component variables are as simple as possible and independent, the distribution of X is significantly more complicated.[15]

2. Let X and Y be independent $\mathrm{Geo}(p)$ random variables, and let $Z = X+Y$ (for example, Z could represent the number of seconds until the second phone call arrives at a call centre). Then $X+Y$ is not geometric, but rather has the so-called negative binomial distribution with order parameter $r=2$.

3. Let U and V be independent, uniformly distributed random variables on $[0,1]$. What is the distribution of $U+V$? At first glance, one might guess it is uniform on $[0,2]$. This is false, for the same reason that the sum of two dice rolls is much more likely to be 7 than 12. The result is called the Irwin–Hall distribution with parameter $n=2$.

Definition 7.3.1. *Let X and Y be independent random variables. Then the distribution of $X+Y$ is called the* ***convolution*** *of the distributions of X and Y.*

Statement 7.3.1. *Let X and Y be independent discrete random variables taking non-negative integer values. Then*

$$\mathbb{P}(X+Y=k) = \sum_{i=0}^{k} \mathbb{P}(X=i)\mathbb{P}(Y=k-i)$$

for all $k \in \mathbb{N}$.

Proof. We only need to use the additivity of probability:

$$\mathbb{P}(X+Y=k) = \mathbb{P}\Big(\bigcup_{i+j=k} \{X=i\} \cap \{Y=j\}\Big)$$

$$= \sum_{i=0}^{k} \mathbb{P}\big(\{X=i\} \cap \{Y=k-i\}\big) = \sum_{i=0}^{k} \mathbb{P}(X=i)\mathbb{P}(Y=k-i).$$

□

[15] If you think the binomial distribution is not complicated, you might try to convince yourself using the de Moivre–Laplace theorem from the previous chapter.

Example 7.3.2. Let $X \sim \mathrm{Pois}(\lambda)$ and $Y \sim \mathrm{Pois}(\mu)$ be independent. What is the distribution of $X+Y$? By the above proposition,

$$\mathbb{P}(X+Y=k) = \sum_{i=0}^{k} \mathbb{P}(X=i)\mathbb{P}(Y=k-i)$$

$$= \sum_{i=0}^{k} \frac{\lambda^i}{i!}e^{-\lambda}\frac{\mu^{k-i}}{(k-i)!}e^{-\mu} = e^{-(\lambda+\mu)}\sum_{i=0}^{k}\frac{k!}{k!\cdot i!\cdot (k-i)!}\lambda^i\mu^{k-i}$$

$$= e^{-(\lambda+\mu)}\frac{1}{k!}(\lambda+\mu)^k,$$

using the binomial theorem. We see that the result is $\mathrm{Pois}(\lambda+\mu)$ distributed.

Computing the convolution in the continuous case is somewhat more involved. Here, the pairs $(i, k-i)$ form a continuum, so the sum becomes an integral, and the distribution is replaced by the density function.

Statement 7.3.2. *Let X and Y be independent continuous random variables. Then the function*

$$z \mapsto \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x$$

is the density function of $X+Y$.

Example 7.3.3. Let X and Y be independent $\mathrm{Exp}(\lambda)$ random variables. By the previous proposition,

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x = \int_0^z \lambda e^{-\lambda x}\lambda e^{-\lambda(z-x)}\,\mathrm{d}x$$

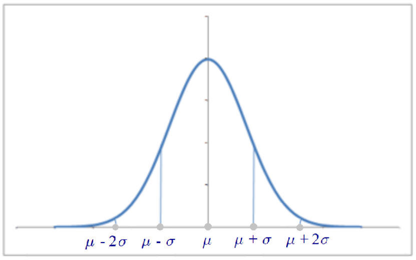
$$= \lambda^2 e^{-\lambda z}\int_0^z 1\,\mathrm{d}x = \lambda^2 e^{-\lambda z} z$$

for all $z > 0$. The resulting distribution is called the gamma distribution.

8 Normal Distribution

We have already discussed notable distributions, but not the most notable one yet. This is the Gaussian normal distribution, or Gaussian distribution, which is of central importance both theoretically and practically.

In applications, it typically appears when dealing with the distribution of a random variable arising as the result of a large number of independent but individually negligible effects, such as the result of a physical measurement. The theoretical background for this, the central limit theorem, will be discussed in Lecture 9.



8.1 Definition of the Distribution

Let us begin by stating what we are talking about, and only afterwards explain why this distribution is so notable.

Definition 8.1.1. *A random variable Y is* ***normally distributed*** *with real parameters μ and σ^2 ($\sigma^2 > 0$) if Y is a continuous random variable with density function*

$$f_Y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (x \in \mathbb{R}).$$

Notation: $Y \sim N(\mu; \sigma^2)$.
If $\mu = 0$ and $\sigma^2 = 1$, the distribution is called the ***standard normal distribution****. Its density function is denoted by φ, that is,*

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

for all $x \in \mathbb{R}$.

We should immediately check that the above f_Y is indeed a probability density function. It is enough to show that it is nonnegative and that its integral over the entire real line is 1. Nonnegativity is obvious (since e^x is always positive), but we must compute the integral. Let us begin with the standard case.

Statement 8.1.1.

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

Proof. Instead of the original integral, let us first examine its square. It can be rewritten as:

$$\left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \, \mathrm{d}x\right)\left(\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} \, \mathrm{d}y\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} \, \mathrm{d}y \, \mathrm{d}x.$$

This is an improper double integral. Since $e^{-\frac{x^2+y^2}{2}}$ is nonnegative everywhere, its double integral equals[16]

$$\lim_{R\to\infty} \iint_{B_R} e^{-\frac{x^2+y^2}{2}} \,\mathrm{d}x\,\mathrm{d}y,$$

where B_R denotes the closed disc of radius R centred at the origin. We can compute this by switching to polar coordinates: $x = r\cos(\alpha)$ and $y = r\sin(\alpha)$. Then

$$\iint_{B_R} e^{-\frac{x^2+y^2}{2}} \,\mathrm{d}x\,\mathrm{d}y = \int_0^{2\pi} \int_0^R e^{-\frac{r^2}{2}} r \,\mathrm{d}r\,\mathrm{d}\alpha = \int_0^{2\pi} \Big[-e^{-\frac{r^2}{2}} \Big]_0^R \,\mathrm{d}\alpha$$

$$= \int_0^{2\pi} \big(-e^{-\frac{R^2}{2}} + 1 \big) \,\mathrm{d}\alpha = 2\pi \big(1 - e^{-\frac{R^2}{2}} \big).$$

As $R \to \infty$, the term $e^{-\frac{R^2}{2}}$ tends to 0, so the square of our desired quantity is 2π. Since e^{-x^2} is positive, its integral is also positive, hence taking the square root gives the claim.[17] □

8.2 Standardisation

The proposition shows that the standard normal distribution is indeed a probability distribution. What about the other normal distributions? We could integrate their density functions directly, but instead we take a more conceptual approach.

Lemma 8.2.1. *Let $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$, and let X be a continuous random variable with density function f_X. Then the density function of $Y = \sigma X + \mu$ is*

$$f_Y(x) = \frac{1}{\sigma} f_X\Big(\frac{x-\mu}{\sigma}\Big) \tag{3}$$

for all $x \in \mathbb{R}$.

Proof. By the definition of the density function, we must show that $\frac{1}{\sigma} f_X\big(\frac{x-\mu}{\sigma}\big)$ is nonnegative and that

$$\int_{-\infty}^a \frac{1}{\sigma} f_X\Big(\frac{x-\mu}{\sigma}\Big) \,\mathrm{d}x = F_Y(a)$$

for all $a \in \mathbb{R}$, where F_Y is the distribution function of Y. Nonnegativity is clear, since f_X is nonnegative and $\sigma > 0$. We can compute the integral as follows:

$$\int_{-\infty}^a \frac{1}{\sigma} f_X\Big(\frac{x-\mu}{\sigma}\Big) \,\mathrm{d}x \overset{z=\frac{x-\mu}{\sigma}}{=} \int_{-\infty}^{\frac{a-\mu}{\sigma}} \frac{1}{\sigma} f_X(z)\sigma \,\mathrm{d}z = F_X\Big(\frac{a-\mu}{\sigma}\Big)$$

[16] This follows from the basic properties of double integrals, provided the function is integrable on every bounded rectangle.

[17] We also need to ensure the integral makes sense. This follows from the fact that $e^{-\frac{x^2}{2}}$ is positive and continuous, so the integral exists (possibly infinite). The above calculation determines its value.

$$= \mathbb{P}\Big(X < \frac{a-\mu}{\sigma}\Big) = \mathbb{P}(\sigma X + \mu < a) = \mathbb{P}(Y < a) = F_Y(a),$$

using that $\int_{-\infty}^{b} f_X(z)\,\mathrm{d}z = F_X(b) = \mathbb{P}(X < b)$ for any real b. □

Corollary 8.2.1. *Let $\mu, \sigma \in \mathbb{R}$ with $\sigma > 0$. A random variable Y has distribution $N(\mu; \sigma^2)$ if and only if there exists $X \sim N(0; 1)$ such that $Y = \sigma X + \mu$.*

In other words, every normal distribution is just a linear transformation of the standard normal distribution. Therefore, in most cases it is sufficient to understand the standard case.

Proof. First, suppose $X \sim N(0; 1)$ and $Y = \sigma X + \mu$. By the lemma, f_Y can be computed from f_X, where $f_X = \varphi$. Hence

$$f_Y(x) = \frac{1}{\sigma} f_X\Big(\frac{x-\mu}{\sigma}\Big) = \frac{1}{\sigma}\frac{1}{\sqrt{2\pi}} e^{-\frac{(\frac{x-\mu}{\sigma})^2}{2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

so Y is indeed $N(\mu; \sigma^2)$ distributed.
Conversely, if $Y \sim N(\mu; \sigma^2)$, let $X = \frac{1}{\sigma}(Y - \mu)$. Rearranging gives $Y = \sigma X + \mu$. By the lemma, (3) holds for all $x \in \mathbb{R}$. Applying the substitution $x = \sigma z + \mu$, we get

$$f_X(z) = \sigma f_Y(\sigma z + \mu) = \sigma \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\sigma z + \mu - \mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

for all $z \in \mathbb{R}$. Thus X is standard normal. □

Remark 8.2.1. Most of the distributions we have studied so far do not have the property that their versions with different parameters are simple transformations of each other. For example, if X has binomial distribution $B(n; p)$ and thus takes values in $\{0, 1, 2, \ldots, n\}$ (all with positive probability), then $3X$ cannot take the value 2, so it cannot be binomially distributed. Similarly, $\sigma X + \mu$ can only be binomial if $\sigma = 1$ and $\mu = 0$.
The only exception so far is the exponential distribution: if $X \sim \mathrm{Exp}(\lambda)$, then $\sigma X \sim \mathrm{Exp}(\sigma^{-1}\lambda)$ for all positive real λ and σ.

So, we have seen the density function of the normal distribution, but what about its distribution function?

Notation The distribution function of the standard normal distribution is denoted by Φ, that is,

$$\Phi(x) = \int_{-\infty}^{x} \varphi(z)\,\mathrm{d}z = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,\mathrm{d}z.$$

The reason for "hiding" the distribution function in integral form is that Φ cannot be expressed in closed form. This does not mean the integral does not exist or is infinite: it exists and is finite for all real x, and it can be well-approximated numerically. It simply cannot be expressed in elementary form without limits. An important identity for Φ is:

$$\Phi(-x) = 1 - \Phi(x) \qquad (\forall x \in \mathbb{R}),$$

which follows from the fact that φ is symmetric about 0, i.e., $\varphi(x) = \varphi(-x)$.
What do μ and σ in the definition of the normal distribution mean?

Statement 8.2.1. *If $Y \sim N(\mu; \sigma^2)$, then $\mathbb{E}(Y) = \mu$ and $\mathbb{D}^2(Y) = \sigma^2$.*

In other words, the normal distributions are parameterised by their mean and variance. From this relationship follows the method of **standardisation**: if Y has a normal distribution, then

$$\frac{Y - \mathbb{E}Y}{\mathbb{D}(Y)} \sim N(0; 1).$$

This is very useful in practice: it is often easier to transform the random variable and compute with the density or distribution function of the standard normal, see also the example below.

Proof. By the Corollary 8.2.1corollary, there exists X such that $Y = \sigma X + \mu$. Hence,

$$\mathbb{E}(Y) = \mathbb{E}(\sigma X + \mu) = \sigma\mathbb{E}(X) + \mu, \quad \text{and} \quad \mathbb{D}^2(Y) = \mathbb{D}^2(\sigma X + \mu) = \mathbb{D}^2(\sigma X) = \sigma^2\mathbb{D}^2(X).$$

Therefore, it suffices to compute for the standard normal X that $\mathbb{E}(X) = 0$ and $\mathbb{D}^2(X) = 1$. Indeed, by the result on expectation of transformed variables

$$\mathbb{E}(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{x^2}{2}}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \big[- e^{-\frac{x^2}{2}} \big]_{-\infty}^{\infty} = \frac{1}{\sqrt{2\pi}}(0-0) = 0,$$

$$\mathbb{E}(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \Big[x\big(-e^{-\frac{x^2}{2}}\big)\Big]_{-\infty}^{\infty} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \big(-e^{-\frac{x^2}{2}}\big)\,\mathrm{d}x = 0+1 = 1.$$

Hence, $\mathbb{D}^2(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 1 - 0 = 1.$ □

Example 8.2.1. We take a sample from a tank after production. Suppose the temperature of the sample, measured in Celsius degrees, is distributed as $N(-2; 1.69)$. What is the probability that the sample's temperature is above $0°C$?
Let the temperature be denoted by Y. We want $\mathbb{P}(Y > 0)$. Consider the standardised variable:

$$X \overset{\text{def}}{=} \frac{Y+2}{\sqrt{1.69}} \sim N(0; 1), \quad \text{i.e.} \quad \mathbb{P}\Big(\frac{Y+2}{1.3} < x\Big) = \mathbb{P}(X < x) = \Phi(x) \quad (x \in \mathbb{R}).$$

Thus,

$$\mathbb{P}(Y > 0) = 1 - \mathbb{P}(Y \leq 0) = 1 - \mathbb{P}(Y < 0) = 1 - \mathbb{P}\Big(\frac{Y+2}{1.3} < \frac{2}{1.3}\Big) = 1 - \Phi\Big(\frac{2}{1.3}\Big),$$

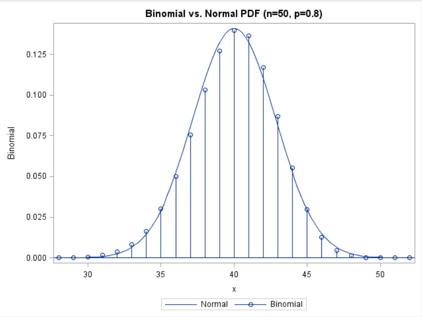
where the value of Φ can be computed by software or looked up in standard tables. We get $\mathbb{P}(Y > 0) \approx 0.0620 \approx 6\%$.

Remark 8.2.2. The standard deviation can also be visually identified: the inflection points of the density function, i.e., where the graph changes convexity, are exactly at $\mu - \sigma$ and $\mu + \sigma$. Specifically, the standard normal density φ is concave on $[-1, 1]$ and convex outside this interval. This follows from $\varphi''(x) = (x^2 - 1)\varphi(x)$, so $\varphi''(x) < 0$ exactly when $x \in (-1, 1)$.

8.3 De Moivre–Laplace Theorem

Why would someone want to use such a distribution? Empirically, the normal distribution often provides a good approximation for distributions of measured results. Examples include the height or weight of a country's population, or the daily average temperature in a given month, etc. Common to these is that the outcome is influenced by many small, roughly independent factors.

The following theorem states that for large n, the binomial distribution can be approximated by a normal distribution.



Theorem 8.3.1. *Let $p \in (0,1)$ and $S_n \sim B(n;p)$. Then for all real $a < b$,*

$$\lim_{n\to\infty} \mathbb{P}\left(a < \frac{S_n - \mathbb{E}(S_n)}{\mathbb{D}(S_n)} < b\right) = \int_a^b \varphi(x)\,\mathrm{d}x = \Phi(b) - \Phi(a),$$

where for the binomial distribution $\mathbb{E}(S_n) = np$ and $\mathbb{D}(S_n) = \sqrt{np(1-p)}$.

One interesting fact is that although $p \neq \frac{1}{2}$ and the binomial distribution is asymmetric, the properly normalised variable's limiting distribution is symmetric about zero.

Example 8.3.1. 31.4% of mathematicians wear sandals. Selecting 100 mathematicians at random, what is the approximate probability that fewer than 25 pairs of sandals will be found on them? (Idealising, assume a mathematician either wears a full pair of sandals or none.)

Let S_n be the number of sandals, with $n = 100$ and $p = 0.314$. We compute $\mathbb{E}S_n = 31.4$ and $\mathbb{D}(S_n) = \sqrt{100 \cdot 0.314 \cdot (1 - 0.314)} \approx 4.6412$. By the above theorem,

$$X = \frac{S_n - 31.4}{4.6412}$$

is approximately standard normal. Hence,

$$\mathbb{P}(S_n < 25) = \mathbb{P}\left(\frac{S_n - 31.4}{4.6412} < \frac{25 - 31.4}{4.6412}\right) = \mathbb{P}(X < -1.3790) \approx \Phi(-1.3790) = 1 - \Phi(1.3790) \approx 0.0839 \approx$$

Using the theorem saved us from tedious binomial probability calculations. The quality of the approximation is supported by further results.

Remark 8.3.1. There might be some déjà vu: for the Poisson distribution, we also said that the binomial distribution converges to Poisson as $n \to \infty$. Now we say it converges to normal. Is there a contradiction? No, because these arise from different parameter regimes.

Here p is fixed, $n \to \infty$, and the variable is standardised, whereas in the Poisson case $p \to 0$, $n \to \infty$ such that $np \to \lambda$.

8.4 Outlook: Heuristic for the De Moivre–Laplace Theorem

We will not prove the theorem, but explore the expression on the left and why the normal distribution appears on the right. (The above theorem and example are the main points for applications, the following is an informal sketch.)
Denote

$$h := \frac{1}{\sqrt{np(1-p)}}.$$

Consider the distribution of the variable

$$\frac{S_n - \mathbb{E}(S_n)}{\mathbb{D}(S_n)} = h(S_n - np),$$

where

$$\operatorname{Ran}\big(h(S_n - np)\big) = \{h(k - np) \mid k = 0, 1, \dots, n\},$$

since S_n is binomial. Note that as n grows, h decreases, so the values get more densely packed. The probability of exactly $h(k-np)$ is $\binom{n}{k}p^k(1-p)^{n-k}$.
From this we create a piecewise linear function $f_n : \mathbb{R} \to \mathbb{R}$ which resembles a density function. For each $k \in \{0, \dots, n\}$,

$$f_n\big(h(k-np)\big) := \frac{1}{h}\binom{n}{k}p^k(1-p)^{n-k},$$

and f_n is linear between these points, zero outside the domain.
How does f_n relate to the theorem? The left side can be approximated by the integral $\int_a^b f_n(x)\,\mathrm{d}x$, more precisely,

$$\lim_{n\to\infty}\left|\int_a^b f_n(x)\,\mathrm{d}x - \mathbb{P}\big(a < h(S_n - np) < b\big)\right| = 0.$$

Assuming this, it suffices to show that $\int_a^b f_n \to \int_a^b \varphi$ for all $a < b$. It can be shown elementarily that for x not of the form $h(k-np)$,

$$\lim_{n\to\infty}\frac{f_n'(x)}{f_n(x)} = -x.$$

This implies that f_n asymptotically satisfies the differential equation

$$f'(x) = -xf(x),$$

except at some special points x. Although many technical questions arise, such as existence of $f(x) = \lim_{n\to\infty} f_n(x)$, continuity, differentiability, whether f is a density function, and whether the convergence of derivatives holds, assuming these, we get

$$\frac{f'(x)}{f(x)} = (\ln f(x))' = -x,$$

almost everywhere, integrating which yields

$$f(x) = e^{-\frac{x^2}{2}} e^c, \quad c \in \mathbb{R}.$$

Since f is a density, $e^c = \frac{1}{\sqrt{2\pi}}$, so f_n indeed converges to the standard normal density.

9 Limit Theorems

In earlier chapters, we have already encountered the concept of so-called limit distributions: we saw that the probabilities of the Poisson distribution can be approximated by those of the binomial distribution under appropriate parameterization, and for the normal distribution we mentioned the de Moivre–Laplace theorem. But what exactly does it mean for distributions to "converge" to another distribution? And can we prove a limit distribution only for special types of distributions? In this lecture, we will address these questions.

We begin by looking at inequalities that estimate the probabilities of the "tails" of a distribution of a random variable. These inequalities are useful tools in the proofs of limit theorems. We will then discuss two of the most fundamental theorems in probability theory: the law of large numbers and the central limit theorem.

9.1 Chebyshev's Inequality

In many problems, the exact distribution of a random variable is unknown, which is especially true in practical situations. However, we would still like to estimate (for example, from above) the probability that X takes an extreme value.

Statement 9.1.1 (Markov's Inequality)**.** *Let X be a nonnegative random variable. Then, for every $a > 0$,*

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

Proof. Define a new random variable: $Y = a$ if $X \geq a$, and 0 otherwise; in other words, $Y = a\mathbf{1}_{\{X \geq a\}}$. Since X is nonnegative, we have $Y \leq X$ on the whole sample space. By the monotonicity of expectation, $\mathbb{E}(Y) \leq \mathbb{E}(X)$. Thus,

$$\mathbb{E}(X) \geq \mathbb{E}(Y) = 0 \cdot \mathbb{P}(Y = 0) + a \cdot \mathbb{P}(Y = a) = a \cdot \mathbb{P}(X \geq a),$$

and rearranging yields the desired inequality. □

Markov's inequality is not necessarily a strong estimate by itself. For example, if $a < \mathbb{E}(X)$, it only asserts that a probability is less than a number greater than 1, which is not very informative. Its typical application is when a is much larger than $\mathbb{E}(X)$, in which case it states intuitively that the probability of X taking a value more extreme than a decreases at least as fast as $x \mapsto \frac{c}{x}$, where $c = \mathbb{E}(X)$.

The following consequence is more commonly used:

Corollary 9.1.1 (Chebyshev's Inequality)**.** *Let X be a random variable with finite variance $\mathbb{D}^2(X)$. Then, for every $a > 0$,*

$$\mathbb{P}\big(|X - \mathbb{E}(X)| \geq a\big) \leq \frac{\mathbb{D}^2(X)}{a^2}.$$

Proof. After rewriting the left-hand side, apply Markov's inequality to the nonnegative variable $\big(X - \mathbb{E}(X)\big)^2$ (and to a^2 instead of a):

$$\mathbb{P}\big(|X - \mathbb{E}(X)| \geq a\big) = \mathbb{P}\Big(\big(X - \mathbb{E}(X)\big)^2 \geq a^2\Big) \leq \frac{\mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big)}{a^2} = \frac{\mathbb{D}^2(X)}{a^2},$$

using the fact that $\big(X - \mathbb{E}(X)\big)^2$ is always nonnegative.
This is exactly the claimed statement. □

Note that while Markov's inequality applies only to nonnegative random variables, Chebyshev's inequality holds for any real-valued random variable.

Example 9.1.1. A database receives on average 50 queries per unit time, with a standard deviation of 5.[18] Give a lower bound on the probability that the number of queries in a unit time is between 40 and 60.
Let X denote the number of queries in a unit time. By Chebyshev's inequality,

$$\mathbb{P}(40 < X < 60) = \mathbb{P}(|X - 50| < 10) = 1 - \mathbb{P}(|X - 50| \geq 10) \geq 1 - \frac{\mathbb{D}^2(X)}{a^2} = 1 - \frac{5^2}{10^2} = \frac{3}{4}.$$

Observe that this bound requires no assumption on the distribution other than knowing its mean and variance.

Markov's inequality can serve as the basis for stronger bounds by applying other functions instead of squaring, or by imposing stronger conditions on X.

Corollary 9.1.2 (Parameterized Chernoff Bound)**.** *Let X be a random variable. Then, for all $a, t > 0$*[19]

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(e^{tX})}{e^{ta}}.$$

Proof. For any $t > 0$, the function $x \mapsto e^{tx}$ is monotonically increasing.
Thus, by Markov's inequality,

$$\mathbb{P}\big(X \geq a\big) = \mathbb{P}\big(e^{tX} \geq e^{ta}\big) \leq \frac{\mathbb{E}(e^{tX})}{e^{ta}},$$

using the fact that both e^{tX} and e^{ta} are positive. □

[18] Estimating the variance from empirical data (e.g., from repeated experiments) belongs to the domain of statistics and will not be discussed here.

[19] It may happen that $\mathbb{E}(e^{tX}) = \infty$, in which case the inequality holds trivially by the convention "∞/positive constant $= \infty$", but is useless.

Example 9.1.2. Let $X \sim \text{Pois}(5)$. Give an upper bound on $\mathbb{P}(X \geq 10)$.
By the Chernoff bound,

$$\mathbb{P}(X \geq 10) \leq \frac{\mathbb{E}(e^{tX})}{e^{10t}} = e^{-10t} \sum_{k=0}^{\infty} e^{tk} \frac{5^k}{k!} e^{-5} = e^{-10t} \sum_{k=0}^{\infty} \frac{(5e^t)^k}{k!} e^{-5} = e^{-10t+5e^t-5}.$$

This bound holds for any $t > 0$, so in particular we can choose t to minimize the exponent. Differentiation shows the minimum occurs at $t = \ln(2)$, with value $e^5/1024 \approx 0.1449$. In contrast, Markov's inequality would give only $\mathbb{E}(X)/10 = \frac{1}{2}$ as an upper bound.

In the remainder of the chapter, we apply these ideas to prove the two most important theorems of the topic.

9.2 Law of Large Numbers

In everyday language, the law of large numbers is often taken to mean that if we try something many times, eventually we will succeed. This claim (or its precise formulation) is a special case of what in probability theory is called the law of large numbers. In fact, the "law" is more general: it speaks not only about probabilities, but also about expectations.
Let us first formalize the above informal statement. Let $A_1, \ldots, A_n$ be independent events, each occurring with probability p, where $0 < p < 1$. These events correspond to repeating the same experiment n times. The claim is that for sufficiently large n, at least one will occur, even if p is small. Formally,

$$\lim_{n \to \infty} \mathbb{P}(A_1 \cup \cdots \cup A_n) \to 1.$$

Note that when we introduced the concept of probability, we implicitly assumed something stronger: namely, that in independent trials, the proportion of successes converges to the success probability itself (and thus will not be zero if $p > 0$). This is also a special case of the law of large numbers.
To express the latter in formulas, let $X_i = 1$ if A_i occurs, and 0 otherwise (i.e., $X_i = \mathbf{1}_{A_i}$). Let $\overline{X_n} = \frac{X_1 + \ldots + X_n}{n}$ denote the average. The "intuitive truth" we previously accepted states that as $n \to \infty$,

$$\overline{X_n} \to p.$$

The problem is that this statement is not precise: in what sense does a sequence of random variables converge? As the following theorem shows, there are several possible senses.

Theorem 9.2.1. *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with $\mathbb{E}X_n = \mu$ and $\mathbb{D}(X_n) = \sigma$ for all n, where $\mu, \sigma \in \mathbb{R}$ are fixed. Let $\overline{X_n} = \frac{X_1 + \ldots + X_n}{n}$.*

- ***Weak Law of Large Numbers:*** *For any $\varepsilon > 0$,*

$$\lim_{n \to \infty} \mathbb{P}\left(\left|\overline{X_n} - \mu\right| \geq \varepsilon\right) = 0.$$

- ***Strong Law of Large Numbers:***

$$\mathbb{P}\Big(\lim_{n\to\infty} \overline{X_n} = \mu\Big) = 1,$$

where the limit is understood pointwise in ω.

As we see, the "law of large numbers" is not a single theorem, but a family of theorems under various conditions.
The theorem generalizes the earlier discussion in that X_i need not be $\{0,1\}$-valued, but can have any distribution. If X_i is $\{0,1\}$-valued, then the weak law is called the **Bernoulli weak law of large numbers**.

Proof of the Weak Law of Large Numbers. By the linearity of expectation, $\mathbb{E}(\overline{X_n}) = \mu$. Thus, by Chebyshev's inequality,

$$\mathbb{P}\big(\big|\overline{X_n} - \mu\big| \geq \varepsilon\big) = \mathbb{P}\Big(\big|\overline{X_n} - \mathbb{E}(\overline{X_n})\big| \geq \varepsilon\Big) \leq \frac{\mathbb{D}^2(\overline{X_n})}{\varepsilon^2}.$$

Since $X_1, \dots, X_n$ are independent,

$$\mathbb{D}^2(\overline{X_n}) = \mathbb{D}^2\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\big(\mathbb{D}^2(X_1) + \cdots + \mathbb{D}^2(X_n)\big) = \frac{\sigma^2}{n},$$

which tends to 0 as $n \to \infty$. The claim follows. □

Why is the first called "weak" and the second "strong"? First, the strong law implies the weak one (proof omitted). Second, they differ in the mode of convergence: in the weak law we have *convergence in probability*, while in the strong law we have *almost sure convergence.*
In probability theory, convergence in probability means that the fraction of outcomes ω for which the deviation exceeds ε becomes negligible. However, it might still happen that for any fixed ω, there are infinitely many n with large deviation. The strong law rules this out, stating that almost surely, after some point, the deviation will always remain within ε.
Why then discuss the "weaker" weak law at all? Because under weaker assumptions (e.g., dropping independence or identical distribution), the weak law may still hold while the strong law fails. For example, daily average temperatures measured over a long time are neither independent nor identically distributed—summer follows winter, hopefully.

9.3 Central Limit Theorem

The law of large numbers tells us *where* we converge (to the expectation), but not *how fast.* In other words, it says nothing about the rate at which the deviation from the mean decreases as n increases.
To answer this, we need the concept of convergence in distribution, since the deviation from the mean is not a fixed number but a distribution (increasingly concentrated around the mean) as n grows.

Definition 9.3.1. *Let $X_1, X_2, \dots, X_n, \dots$ be a sequence of random variables, and let F_{X_n} denote the distribution function of X_n, while F_Z denotes the distribution function of Z. The sequence $(X_n)_{n\in\mathbb{N}}$ is said to* ***converge in distribution*** *to a random variable Z if*

$$F_{X_n}(x) \to F_Z(x) \qquad (n \to \infty)$$

for every $x \in \mathbb{R}$ at which F_Z is continuous.[20] *Notation: $X_n \xrightarrow{d} Z$.*

Theorem 9.3.1 (Central Limit Theorem)**.** *Let $X_1, X_2, \dots$ be independent, identically distributed random variables. Assume that $0 < \mathbb{D}^2 X_1 < \infty$. Let $Z \sim N(0,1)$. Then*

$$\frac{X_1 + \cdots + X_n - n\mathbb{E}(X_1)}{\sqrt{n}\mathbb{D}(X_1)} \xrightarrow{d} Z \qquad (n \to \infty).$$

Expanding the definition of convergence in distribution, this means that the distribution function of the left-hand side converges to that of Z, denoted by Φ, at all continuity points of Φ. Since Φ is continuous everywhere, the theorem can be restated as follows:

Corollary 9.3.1 (Central Limit Theorem, probabilistic form)**.** *Let $X_1, X_2, \dots$ be independent, identically distributed random variables. Assume that $0 < \mathbb{D}^2 X_1 < \infty$. Then*

$$\mathbb{P}\left(\frac{X_1 + \cdots + X_n - n\mathbb{E}(X_1)}{\sqrt{n}\mathbb{D}(X_1)} < a\right) \to \Phi(a)$$

for every $a \in \mathbb{R}$ as $n \to \infty$.

From this formulation it is clear that the central limit theorem generalizes the de Moivre–Laplace theorem: here we consider any independent, identically distributed random variables with finite variance, while the de Moivre–Laplace theorem deals only with sums of independent indicator variables (i.e., binomial random variables).
In both cases, the statement is that the distribution function can be approximated by Φ.

Remark 9.3.1. The sum in the theorem can also be expressed as

$$\frac{X_1 + \cdots + X_n - n\mathbb{E}(X_1)}{\sqrt{n}\mathbb{D}(X_1)} = \frac{\overline{X_n} - \mathbb{E}(\overline{X_n})}{\mathbb{D}(\overline{X_n})},$$

which follows from rearrangement and the properties of the standard deviation. In other words, the theorem concerns the standardized form of $\overline{X_n}$.

Example 9.3.1. A winery sells on average 100 liters of wine per working day, with a standard deviation of 20. Assume the daily amounts are independent and identically distributed. In the remaining 50 working days of the year, they would need to sell 4750 liters to surpass last year's sales. What is the probability of achieving this?

[20] It can be shown that the above definition of convergence in distribution is equivalent to the following condition: $\lim_{n\to\infty} \mathbb{E}f(X_n) = \mathbb{E}f(X)$ for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$.

Let $X_1, X_2, \dots, X_{50}$ denote the daily amounts sold. We have $\mathbb{E}(X_1) = 100$ and $\mathbb{D}(X_1) = 20$. By the central limit theorem, the appropriately standardized sum is approximately normal, so

$$\mathbb{P}\Big(\sum_{i=1}^{50} X_i \geq 4750\Big) = 1 - \mathbb{P}\Bigg(\frac{\sum_{i=1}^{50} X_i - 50 \cdot 100}{\sqrt{50} \cdot 20} < \frac{4750 - 5000}{\sqrt{50} \cdot 20}\Bigg)$$

$$\approx 1 - \Phi\Bigg(\frac{-250}{\sqrt{50} \cdot 20}\Bigg) = 1 - \Phi(-1.7678) = \Phi(1.7678) \approx 0.9615.$$

Thus, the probability of success is about 96%.

To prove the theorem, we use an idea from the first subsection, specifically the quantity $\mathbb{E}(e^{tX})$ appearing in the Chernoff bound.

Definition 9.3.2. *The **moment generating function** of a random variable X is defined by*

$$M_X(t) \overset{\text{def}}{=} \mathbb{E}(e^{tX}),$$

for those t where $\mathbb{E}(e^{tX})$ is finite.

The moment generating function is suitable for use in the proof of the central limit theorem because of the following properties:

Statement 9.3.1. *Let Y, Z and $X_1, X_2, \dots$ be random variables. Assume $M_Y(t)$ and $M_Z(t)$ are defined for all $t \in \mathbb{R}$.*

1. *If $M_Y(t) = M_Z(t)$ for all $t \in \mathbb{R}$, then Y and Z have the same distribution.*

2. *If $\lim_{n\to\infty} M_{X_n}(t) = M_Z(t)$ for all $t \in \mathbb{R}$, then $X_n \xrightarrow{d} Z$.*

We assume in the proof that the moment generating functions are defined everywhere. The result still holds without this assumption, but we omit the general proof. We also state without proof the lemma that appears in the argument.

Sketch of the proof of the Central Limit Theorem. For simplicity, we first consider the case where $\mathbb{E}(X_1) = 0$ and $\mathbb{D}(X_1) = 1$. This is sufficient, since if the theorem holds for the standardized variables $\frac{X_i - \mathbb{E}(X_1)}{\mathbb{D}(X_1)}$, then it follows for the original variables by the rearrangement

$$\frac{X_1 + \cdots + X_n - n\mathbb{E}(X_1)}{\sqrt{n}\mathbb{D}(X_1)} = \sqrt{n} \cdot \frac{\sum_{i=1}^{n} \frac{X_i - \mathbb{E}(X_1)}{\mathbb{D}(X_1)}}{n}.$$

Thus, it suffices to show $\sqrt{n}\,\overline{X_n} \xrightarrow{d} Z$. By Proposition 9.3.1, it is enough to prove that

$$\lim_{n\to\infty} M_{\frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}}(t) = M_Z(t) \tag{4}$$

for all $t \in \mathbb{R}$.
First, compute the right-hand side of (4):

$$M_Z(t) = \mathbb{E}(e^{tZ}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz} e^{-z^2/2} \, dz = e^{t^2/2}.$$

Thus, the moment generating function of the standard normal distribution is $t \mapsto e^{t^2/2}$.
Next, compute the moment generating function of X_1:

$$M_{X_1}(t) = 1 + \frac{t^2}{2}\big(1 + r(t)\big), \tag{5}$$

where $r(t)$ denotes the expectation of the remaining terms in the power series expansion of e^{tX_1}. It can be shown that:

Lemma 9.3.1. $\lim_{t \to 0} r(t) = 0$.

From this lemma, the left-hand side of (4) can be written as

$$M_{\frac{\sum_{i=1}^n X_i}{\sqrt{n}}}(t) = \mathbb{E}\left(e^{t\frac{\sum_{i=1}^n X_i}{\sqrt{n}}}\right) = \prod_{i=1}^n M_{X_i}\Big(\frac{t}{\sqrt{n}}\Big) = M_{X_1}\Big(\frac{t}{\sqrt{n}}\Big)^n. \tag{6}$$

Taking logarithms and substituting (5),

$$\ln M_{\frac{\sum_{i=1}^n X_i}{\sqrt{n}}}(t) = n \ln\left(1 + \frac{t^2}{2n}(1 + r_n)\right),$$

where $r_n = r\big(\frac{t}{\sqrt{n}}\big) \to 0$ as $n \to \infty$. Since $\frac{\ln(1+y)}{y} \to 1$ as $y \to 0$, it follows that

$$\lim_{n \to \infty} \ln M_{\frac{\sum_{i=1}^n X_i}{\sqrt{n}}}(t) = \frac{t^2}{2}.$$

Exponentiating, we obtain $\lim_{n\to\infty} M_{\frac{\sum_{i=1}^n X_i}{\sqrt{n}}}(t) = e^{t^2/2} = M_Z(t)$ for all t, as required. □

10 Linear Regression

So far, we have considered the covariance of random variables only in the discrete case, even though the same definition works perfectly well for continuous random variables as well. The reason we postponed this topic was the absence of the concept of a joint density function, which enables us to compute covariance in the continuous case.

As an application of the concepts of variance and covariance, we will also discuss linear regression here. By *linear regression* we primarily mean a statistical model that describes relationships between variables based on an assumed linear dependence. Such models are used for both predictive and explanatory purposes — the former belonging to estimation theory, the latter to hypothesis testing, both subfields of statistics.

However, linear regression also has a purely probabilistic aspect, which does not require the basic concepts of statistics such as samples or estimators. This concerns the question of how to choose numbers α and β for given random variables X and Y so that $\beta X + \alpha$ is as close as possible to Y.

10.1 Variance and Covariance in the Continuous Case

Let X be a continuous random variable with density function f_X. How can we compute the variance of X?

Previously, we have already considered the expectation of X, and even the expectation of a transformation $g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ is any continuous function. Therefore, we can compute the variance of X as follows (as we already did for the normal distribution):

$$\mathbb{D}^2(X) \overset{\text{def}}{=} \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x - \Big(\int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x\Big)^2.$$

The standard deviation is then $\mathbb{D}(X) = \sqrt{\mathbb{D}^2(X)}$.

The interpretation of variance (and standard deviation) is the same here: it measures the average squared deviation from the mean (and its square root). Intuitively, it describes how much the density function "spreads out" around the mean.[21]

Example 10.1.1. Let $Z \sim \mathrm{Exp}(\lambda)$ for some $\lambda > 0$. By two integrations by parts,

$$\mathbb{E}(Z^2) = \int_0^{\infty} z^2 \lambda e^{-\lambda z}\,\mathrm{d}z = \frac{2}{\lambda^2}.$$

Hence

$$\mathbb{D}^2(Z) = \frac{2}{\lambda^2} - \Big(\frac{1}{\lambda}\Big)^2 = \frac{1}{\lambda^2}, \qquad \mathbb{D}(Z) = \frac{1}{\lambda}.$$

[21] The shape of the density function can be described by many other derived quantities, such as the mean absolute deviation, kurtosis (peakedness), or skewness.

Continuing along these lines, we see that the original formula, provided that the expectations exist. Moreover, the following identity still holds:

$$\operatorname{cov}(X,Y) \overset{\text{def}}{=} \mathbb{E}\big((X-\mathbb{E}X)(Y-\mathbb{E}Y)\big) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

In practice, the term $\mathbb{E}(XY)$ is often the most challenging to compute, since the distribution of XY depends on the *joint* distribution of (X,Y), not only on the marginal distributions. The following statement allows us to determine $\mathbb{E}(XY)$ without explicitly computing the distribution of XY.

Statement 10.1.1. *Let $\underline{X} = (X_1,\dots,X_n)$ be a continuous random vector, and let $g : \mathbb{R}^n \to \mathbb{R}$ be such that $\mathbb{E}(g(X_1,\dots,X_n))$ exists. Then*

$$\mathbb{E}(g(X_1,\dots,X_n)) = \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} g(x_1,\dots,x_n)\, f_{\underline{X}}(x_1,\dots,x_n)\,\mathrm{d}x_1\dots\mathrm{d}x_n.$$

If g is continuous and nonnegative, then $\mathbb{E}(g(X_1,\dots,X_n))$ exists (possibly taking the value $+\infty$).

As a special case, if (X,Y) is a continuous random vector, then

$$\mathbb{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y,$$

provided the expectation exists (here $g(x,y) = x \cdot y$ is not nonnegative).

Example 10.1.2. Let X denote the annual rainfall (in 1000 mm), and Y the number of umbrellas sold (in 1000 units). Suppose their joint density is

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{5}(4-2x^2+xy-y^2) & \text{if } 0<x<1,\ 0<y<2, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\mathbb{E}(XY) = \frac{17}{45}, \quad \mathbb{E}(X) = \frac{7}{15}, \quad \mathbb{E}(Y) = \frac{4}{5}.$$

Hence

$$\operatorname{cov}(X,Y) = \frac{17}{45} - \frac{7}{15}\cdot\frac{4}{5} = \frac{1}{225} \approx 0.0044.$$

The properties of variance and covariance discussed earlier hold in the continuous case as well.

Lemma 10.1.1. *Let (X,Y,Z) be a random vector. Then, assuming the expressions are well-defined:*

1. *For $c \in \mathbb{R}$, $\mathbb{D}(X+c) = \mathbb{D}(X)$ and $\mathbb{D}(cX) = |c|\mathbb{D}(X)$.*

2. $\mathbb{D}^2(X+Y) = \mathbb{D}^2(X) + \mathbb{D}^2(Y) + 2\operatorname{cov}(X,Y)$.

3. $\mathbb{D}^2(X) = 0$ *iff* $\mathbb{P}(X = c) = 1$ *for some* $c \in \mathbb{R}$.

4. *If X and Y are independent, then* $\operatorname{cov}(X,Y) = 0$*; in particular* $\mathbb{D}^2(X+Y) = \mathbb{D}^2(X) + \mathbb{D}^2(Y)$.

5. *(Bilinearity) If* $b, c \in \mathbb{R}$*, then* $\operatorname{cov}(X, bY + cZ) = b\operatorname{cov}(X,Y) + c\operatorname{cov}(X,Z)$.

Remark 10.1.1. Point 4 of the lemma can be generalized by the following result.

Lemma 10.1.2. *If X and Y are independent random variables, and $g, h : \mathbb{R} \to \mathbb{R}$ are continuous functions, then $g(X)$ and $h(Y)$ are also independent.*

For a random vector, it is common to arrange variances and covariances into a matrix. This is not merely for compactness: in computations with random vectors, the covariance matrix naturally appears in products with vectors, and its determinant or trace may be of interest (see, for example, the multivariate normal distribution in Lecture 12).

Definition 10.1.1. *The* ***covariance matrix*** *of the random vector $\underline{X} = (X_1, \dots, X_n)$ is the $n \times n$ real matrix*

$$\operatorname{cov}(\underline{X}) = \begin{pmatrix} \operatorname{cov}(X_1,X_1) & \operatorname{cov}(X_1,X_2) & \dots & \operatorname{cov}(X_1,X_n) \\ \operatorname{cov}(X_2,X_1) & \operatorname{cov}(X_2,X_2) & & \vdots \\ \vdots & & \ddots & \\ \operatorname{cov}(X_n,X_1) & \dots & & \operatorname{cov}(X_n,X_n) \end{pmatrix},$$

i.e., $\operatorname{cov}(\underline{X})_{i,j} = \operatorname{cov}(X_i, X_j)$ *for all* $1 \le i,j \le n$.

Since $\mathbb{D}^2(X_i) = \operatorname{cov}(X_i, X_i)$, the diagonal entries are the variances of the components.

Statement 10.1.2. *Let $\underline{X} = (X_1, \dots, X_n)$ be a random vector.*

1. $\operatorname{cov}(\underline{X})$ *is symmetric:* $\operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_j, X_i)$ *for all* i, j.

2. $\operatorname{cov}(\underline{X})$ *is positive semidefinite: for all* $(a_1, \dots, a_n) \in \mathbb{R}^n$,

$$\sum_{i=1}^n \sum_{j=1}^n a_i \operatorname{cov}(X_i, X_j)\, a_j \ge 0,$$

with equality iff $\sum_{i=1}^n a_i X_i$ *is constant with probability* 1.

Proof. Symmetry follows directly from the definition. For positive semidefiniteness, note that by Lemma point 3 and bilinearity,

$$\mathbb{D}^2\Big(\sum_{i=1}^n a_i X_i\Big) = \sum_{i=1}^n \sum_{j=1}^n a_i \operatorname{cov}(X_i, X_j)\, a_j.$$

Since variance is always nonnegative, the right-hand side is nonnegative, and it is zero exactly when $\sum_{i=1}^n a_i X_i$ is almost surely constant. □

Example 10.1.3. For the (X, Y) in Example 10.1.2, we also have

$$\mathbb{D}^2(X) = \frac{7}{90}, \quad \mathbb{D}^2(Y) = \frac{58}{225}.$$

Thus, for $\underline{Z} = (X, Y)$,

$$\operatorname{cov}(\underline{Z}) = \begin{pmatrix} \frac{7}{90} & \frac{1}{225} \\ \frac{1}{225} & \frac{58}{225} \end{pmatrix}.$$

10.2 Linear Regression

Suppose we own a shop selling umbrellas and receive a long-term forecast for next year's rainfall. In the absence of better information, we try to use this forecast to estimate how much stock to order, i.e., approximately how many umbrellas we will sell. How should we make this estimate if past years suggest some relationship between rainfall and the number of umbrellas sold? One possible method is *linear regression*.

Let X denote the annual rainfall and Y the number of umbrellas sold, as in Example 10.1.2. Assume that (X, Y) has the joint density $f_{X,Y}$ from that example. The basic idea of linear regression is to approximate Y by a linear function of X, i.e., in the form $\beta X + \alpha$, as closely as possible.

Note that "best approximation" is not a well-defined concept until we specify a criterion for what makes an approximation good or bad. Several approaches are possible,[22] but the most basic is the **method of least squares**.

Definition 10.2.1. *Let X and Y be random variables. The* ***linear regression*** *of Y on X is the random variable $\beta X + \alpha$, where $\alpha, \beta \in \mathbb{R}$ minimize*

$$\mathbb{E}\Big(\big(Y - (\beta X + \alpha)\big)^2\Big). \tag{7}$$

This optimization problem essentially always has a unique solution:

Statement 10.2.1. *Let X and Y be random variables with finite $\mathbb{D}^2(X)$, $\mathbb{D}^2(Y)$, and $\operatorname{cov}(X, Y)$, and suppose $\mathbb{D}^2(X) \neq 0$. Then the expectation in (7) is minimized exactly when*

$$\beta = \frac{\operatorname{cov}(X, Y)}{\mathbb{D}^2(X)} \qquad \text{and} \qquad \alpha = \mathbb{E}(Y) - \frac{\operatorname{cov}(X, Y)}{\mathbb{D}^2(X)}\mathbb{E}(X).$$

Definition 10.2.2. *The* ***regression line*** *of Y on X is the set*

$$\{(x, y) \in \mathbb{R}^2 \mid y = \beta x + \alpha\}$$

where β and α are given by the above proposition.

[22] Alternative versions of linear regression that define "best approximation" differently include weighted linear regression, ridge regression, and ℓ_1 regression.

Visually, in the (X, Y)-plane, the regression line is the straight line that best approximates the distribution of (X, Y). The model is most useful when the joint distribution of (X, Y) is concentrated near this line.

Remark 10.2.1. The formulas for β and α are not necessarily easy to remember or justify. One heuristic (not a proof) is to choose α and β so that $\beta X + \alpha$ has the same expectation as Y and the same covariance with X as Y does:

$$\mathbb{E}(Y) = \beta\mathbb{E}(X) + \alpha, \quad \operatorname{cov}(X, Y) = \beta\, \mathbb{D}^2(X),$$

which leads directly to the given formulas.
A more compact approach uses the correlation:

$$\operatorname{corr}(X, Y) \overset{\text{def}}{=} \frac{\operatorname{cov}(X, Y)}{\mathbb{D}(X)\mathbb{D}(Y)},$$

a number between -1 and 1 measuring the linear dependence of X and Y. If $\beta X + \alpha$ is the regression of Y on X, then

$$\frac{(\beta X + \alpha) - \mathbb{E}(Y)}{\mathbb{D}(Y)} = \frac{X - \mathbb{E}(X)}{\mathbb{D}(X)} \cdot \operatorname{corr}(X, Y).$$

That is, replacing Y by its regression in the standardized form of Y yields the standardized X scaled by the correlation.

Proof. We must minimize

$$h(\alpha, \beta) = \mathbb{E}\big(Y - (\beta X + \alpha)\big)^2 = \mathbb{E}(Y^2) + \beta^2\mathbb{E}(X^2) + \alpha^2 - 2\beta\mathbb{E}(XY) - 2\alpha\mathbb{E}(Y) + 2\alpha\beta\mathbb{E}(X).$$

This is nonnegative (as an expected square) and is a quadratic polynomial in α and β. A quadratic polynomial attains its global minimum where all partial derivatives vanish:

$$\frac{\partial h}{\partial \beta}: \quad 2\beta\mathbb{E}(X^2) - 2\mathbb{E}(XY) + 2\alpha\mathbb{E}(X) = 0,$$

$$\frac{\partial h}{\partial \alpha}: \quad 2\alpha - 2\mathbb{E}(Y) + 2\beta\mathbb{E}(X) = 0.$$

This linear system has the solution

$$\beta = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\mathbb{E}(X^2) - \mathbb{E}(X)^2} = \frac{\operatorname{cov}(X, Y)}{\mathbb{D}^2(X)},$$

$$\alpha = \mathbb{E}(Y) - \beta\mathbb{E}(X),$$

which are exactly the stated formulas. □

Example 10.2.1. In the introductory example, where X is rainfall and Y is umbrellas sold, the covariance matrix entries give

$$\beta = \frac{1/225}{7/90} = \frac{2}{35}, \qquad \alpha = \frac{4}{5} - \frac{2}{35} \cdot \frac{7}{15} = \frac{58}{75}.$$

Thus, given a forecast for X, we can approximate Y using these coefficients. Interpretation: increased rainfall slightly increases the already high baseline stock requirement.

Since linear regression is only an approximation, it is useful to know the typical error in predicting Y (here "error" means the variance of the difference).

Statement 10.2.2. *If $\beta X + \alpha$ is the regression of Y on X, then*

$$\mathbb{D}^2\big(Y - (\beta X + \alpha)\big) = \mathbb{D}^2(Y) - \frac{\operatorname{cov}(X,Y)^2}{\mathbb{D}^2(X)}.$$

Proof. Using the variance properties and $\beta = \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2(X)}$:

$$\mathbb{D}^2(Y-(\beta X+\alpha)) = \mathbb{D}^2(Y-\beta X) = \mathbb{D}^2(Y)+\beta^2\mathbb{D}^2(X)-2\beta\operatorname{cov}(Y,X) = \mathbb{D}^2(Y)-\frac{\operatorname{cov}(X,Y)^2}{\mathbb{D}^2(X)}.$$

□

Remark 10.2.2. Equivalently,

$$\mathbb{D}^2(Y - (\beta X + \alpha)) = \mathbb{D}^2(Y) \cdot \big(1 - \operatorname{corr}(X,Y)^2\big).$$

Thus, the stronger the correlation between X and Y, the smaller the fraction of $\mathbb{D}^2(Y)$ contributing to the prediction error.

Example 10.2.2. In the previous example,

$$\mathbb{D}^2\big(Y - (\beta X + \alpha)\big) = \frac{58}{225} - \frac{(1/225)^2}{(7/90)^2} \approx 0.2545.$$

Thus, sales can deviate considerably from the value predicted by the regression.

In statistics, "linear regression" typically means something slightly different: it is assumed that the distributions of the random variables (and thus the variances and covariances) are unknown, and the coefficients α and β themselves are estimated from a finite sample. This leads to different formulas and interpretations, but the core idea remains: to find an approximately linear relationship between the variables of interest.

11 Conditional Probability and Multidimensional Distributions

In this chapter, we state the version of the law of total probability where the condition involves a random variable instead of a complete system of events. As an independent topic, we examine in more detail some well–known distributions of random vectors (the so–called multidimensional distributions), with special attention to the multivariate normal distribution.

11.1 Law of Total Probability, Continuous Case

After the previous chapter, the reader may be left with a slight sense of incompleteness: while we have stated the law of total expectation in several forms (using a complete system of events and using the level sets $\{X = x\}$ of a random variable, both in the discrete and continuous case), the law of total probability was formulated only for a complete system of events. What happens to the latter if the condition is a random variable (or a level set of one)?

If X is a discrete random variable, then the law of total probability brings nothing new:

$$\mathbb{P}(A) = \sum_{k \in S_X} \mathbb{P}(A \mid X = k) \cdot \mathbb{P}(X = k),$$

where S_X is the set of k values such that $\mathbb{P}(X = k) > 0$, so that $\mathbb{P}(A \mid X = k)$ makes sense. This is a special case of the original law of total probability. However, if X is continuous, then $\mathbb{P}(A \mid X = k)$ is meaningless. The resolution of this problem is the same as in the case of the regression function.

Definition 11.1.1. *Let X be a random variable and A an event. The **conditional probability** of A given X is the regression function*

$$x \mapsto \mathbb{E}(\mathbf{1}_A \mid X = x),$$

denoted $\mathbb{P}(A \mid X = x)$.

Intuitively, $\mathbb{E}(\mathbf{1}_A \mid X = x)$ is the probability of the event A (or, more precisely, its best average approximation), knowing the value of X.

From here, the law of total probability can be guessed:

Theorem 11.1.1 (Law of Total Probability)**.** *Let X be a continuous random variable and A an event. Then*

$$\mathbb{P}(A) = \int_{-\infty}^{\infty} \mathbb{P}(A \mid X = x) f_X(x)\,\mathrm{d}x,$$

where f_X is the density of X.

Example 11.1.1. Let X denote the preparation time (in days) of a student for the probability theory exam. Suppose X is uniformly distributed on $[\varepsilon, 20]$ (where ε is smaller than 20, and hopefully positive). Given that the student spends x days preparing, the probability of obtaining the top grade is $\left(\frac{x}{21}\right)^2$. What is the probability of obtaining the top grade?
In the notation of the previous theorem: $f_X(x) = \frac{1}{20-\varepsilon}$ for $\varepsilon \leq x \leq 20$ and 0 otherwise. Moreover, $\mathbb{P}(A \mid X = x) = \left(\frac{x}{21}\right)^2$. Thus

$$\mathbb{P}(A) = \int_{\varepsilon}^{20} \left(\frac{x}{21}\right)^2 \frac{1}{20-\varepsilon} \, \mathrm{d}x = \left[\frac{x^3}{3 \cdot 21^2 (20-\varepsilon)}\right]_{\varepsilon}^{20} = \frac{\varepsilon^2 + 20\varepsilon + 20^2}{3 \cdot 21^2}.$$

If $\varepsilon = 1$, this is approximately 0.3182.

Remark 11.1.1. A special case of conditional probability is the conditional distribution function:

$$F_{Y|X}(y \mid x) = \mathbb{P}(Y < y \mid X = x).$$

11.2 Multidimensional Distributions

Let $\underline{X} = (X_1, \ldots, X_m)$ be a random vector. As in the one–dimensional case, we can speak of the distribution of $\underline{X}$ (for example, described by the joint distribution function), as we already did in the context of joint distributions. Let us now examine some commonly occurring multidimensional distributions.
A notable discrete distribution is the binomial. How can it be generalized to more variables? One natural way is: let $X_1, \ldots, X_m$ be jointly independent with $X_i \sim B(n; p_i)$ for some $n \in \mathbb{N}$ and $0 < p_i < 1$ $(i = 1, \ldots, m)$. This yields a meaningful multidimensional distribution, but it is not the only way to generalize the binomial.

Example 11.2.1. We relabel a fair die: one face has a 1, two faces have a 2, and three faces have a 3. We roll the die 13 times. Let X_i denote the number of i's rolled. What is $\mathbb{P}(X_1 = 3, X_2 = 4, X_3 = 6)$?
The probability can be computed combinatorially:

$$\mathbb{P}(X_1 = 3, X_2 = 4, X_3 = 6) = \frac{13!}{3!4!6!} \left(\frac{1}{6}\right)^3 \left(\frac{1}{3}\right)^4 \left(\frac{1}{2}\right)^6 \approx 0.05364,$$

since the number of ways to arrange three 1's, four 2's, and six 3's is $\frac{13!}{3!4!6!}$ (a multinomial coefficient), and the probability of each such outcome is $p_1^3 p_2^4 p_3^6$, where p_i is the probability of rolling i.

Definition 11.2.1. *The random vector $\underline{X} = (X_1, \ldots, X_m)$ has a **multinomial distribution** with parameters $n \in \mathbb{N}$ and $(p_1, p_2, \ldots, p_m) \in [0,1]^m$ if $p_1 + \cdots + p_m = 1$ and*

$$\mathbb{P}(X_1 = k_1, \ldots, X_m = k_m) = \frac{n!}{k_1! k_2! \ldots k_m!} p_1^{k_1} \ldots p_m^{k_m}$$

for all $0 \leq k_i \leq n$ $(i = 1, \ldots, m)$ with $\sum_{i=1}^{m} k_i = n$.

If $m = 2$ and $(p_1, p_2) = (p, 1-p)$ for some $p \in [0,1]$, then $X_1 \sim B(n;p)$ (and X_2 carries no extra information, since $X_2 = n - X_1$).
It is clear that the X_i are not independent (since $X_1, \dots, X_{m-1}$ determine X_m), yet the marginals of $\underline{X}$ are binomial $B(n;p_i)$. This example shows that marginals do not determine the joint distribution, and that the natural multivariate generalization of a distribution does not necessarily have independent coordinates.

Definition 11.2.2. *A random vector $\underline{X} = (X_1, \dots, X_n)$ has an n–**dimensional standard normal distribution** if it is continuous and its joint density function is*

$$f_{\underline{X}}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2}, \qquad (x_1, \dots, x_n \in \mathbb{R}).$$

How do we obtain not necessarily standard multivariate normal distributions?

Definition 11.2.3. *A random vector $\underline{Y} = (Y_1, \dots, Y_n)$ has a **multivariate normal distribution** if there exists $\underline{\underline{A}} \in \mathbb{R}^{n\times n}$, $\underline{\mu} \in \mathbb{R}^n$, and an n–dimensional standard normal random vector $\underline{X}$ such that*

$$\underline{Y} = \underline{\underline{A}} \cdot \underline{X} + \underline{\mu},$$

*treating $\underline{X}$ as a column vector. The distribution of $\underline{Y}$ is called **nondegenerate** if $\underline{\underline{A}}$ can be chosen to be nonsingular (i.e., $\det(\underline{\underline{A}}) \neq 0$).*

This definition differs from the parameterization in the one–dimensional case, where we specified a (not necessarily standard) normal distribution by its expectation and variance. Let us examine analogous parameters for the multivariate normal distribution.

Definition 11.2.4. *The **mean vector** of a random vector $\underline{Y} = (Y_1, \dots, Y_n)$ is the vector $(\mathbb{E}Y_1, \dots, \mathbb{E}Y_n) \in \mathbb{R}^n$. It is denoted by $\mathbb{E}\underline{Y}$.*

The covariance matrix can also be expressed in terms of the mean vector. Treating $\underline{Y}$ and $\mathbb{E}\underline{Y}$ as column vectors,

$$\operatorname{cov}(\underline{Y}) = \mathbb{E}\big((\underline{Y} - \mathbb{E}\underline{Y}) \cdot (\underline{Y} - \mathbb{E}\underline{Y})^T\big) \in \mathbb{R}^{n\times n},$$

where the product is the matrix product of an $n\times 1$ and a $1\times n$ matrix, and the expectation is taken coordinatewise.

Statement 11.2.1. *Let $\underline{X} = (X_1, \dots, X_n)$ be a standard normal random vector, and let $\underline{Y} = \underline{\underline{A}} \cdot \underline{X} + \underline{\mu}$. Then $\mathbb{E}\underline{Y} = \underline{\mu}$ and $\operatorname{cov}(\underline{Y}) = \underline{\underline{A}} \cdot \underline{\underline{A}}^T$.*

With these parameters, we can also write the density function of the multivariate normal distribution.

Statement 11.2.2. *Let $\underline{Y}$ be a nondegenerate n–dimensional normal random vector. Let $\underline{\mu}$ denote its mean vector and $\underline{\underline{\Sigma}}$ its covariance matrix. Then the density function of $\underline{Y}$ is*

$$f_{\underline{Y}}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \det(\underline{\underline{\Sigma}})^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\underline{\Sigma}}^{-1} (\underline{x}-\underline{\mu})},$$

where $\det(\underline{\underline{\Sigma}})$ is the determinant of $\underline{\underline{\Sigma}}$, and $\underline{\underline{\Sigma}}^{-1}$ is its inverse matrix.

The exponent is a triple matrix product (vector, matrix, vector) that results in a real number. In the two–dimensional case:

$$\underline{\underline{\Sigma}} = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \quad \Rightarrow \quad \underline{\underline{\Sigma}}^{-1} = \frac{1}{\det(\underline{\underline{\Sigma}})} \begin{pmatrix} c & -b \\ -b & a \end{pmatrix} \quad \det(\underline{\underline{\Sigma}}) = ac - b^2,$$

where $a = \mathbb{D}^2(Y_1)$, $b = \operatorname{cov}(Y_1, Y_2)$, and $c = \mathbb{D}^2(Y_2)$.
An important consequence is that a nondegenerate normal distribution is determined by its mean vector $\underline{\mu}$ and covariance matrix $\underline{\underline{\Sigma}}$. (Note that a given $\underline{\underline{\Sigma}}$ may arise from several different matrices $\underline{\underline{A}}$, so this is not obvious.) In fact, the same is true for the degenerate case, but then no density function exists; we do not discuss this case further here.

Notation. An n–dimensional normal distribution is denoted by $N(\underline{\mu}, \underline{\underline{\Sigma}})$, where $\underline{Y} = \underline{\underline{A}} \cdot \underline{X} + \underline{\mu}$, $\underline{X}$ is n–dimensional standard normal, and $\underline{\underline{\Sigma}} = \underline{\underline{A}} \cdot \underline{\underline{A}}^T$. In particular, the standard normal distribution is denoted by $N(\underline{0}, \underline{\underline{I}})$, where $\underline{0}$ is the n–dimensional zero vector, and $\underline{\underline{I}}$ is the n–dimensional identity matrix.

Note that neither in the standard nor in the general case have we yet spoken about the distributions of the coordinates Y_i, nor even mentioned the one–dimensional normal distribution. So what are the marginals of a normal distribution? The answer is moderately surprising:

Statement 11.2.3. *Let $\underline{Y} \sim N(\underline{\mu}, \underline{\underline{\Sigma}})$, where $\underline{\mu} \in \mathbb{R}^n$ and $\underline{\underline{\Sigma}} \in \mathbb{R}^{n \times n}$. Then $Y_i \sim N(\mu_i, \Sigma_{i,i})$.*

In the standard case, we know more: since the density factorizes (because $\frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2}$), the coordinates X_i are jointly independent one–dimensional standard normals. Thus the normal distribution enjoys a pleasant property that the polynomial or Marshall–Olkin distributions do not: the natural multivariate generalization consists of jointly independent copies of the one–dimensional distribution arranged into a vector.
The normal distribution, due to several other properties, is the prime candidate for the "too good to be true" award; we summarize these properties in the following statement:

Corollary 11.2.1. *Let $(Y_1, Y_2) \sim N(\underline{\mu}, \underline{\underline{\Sigma}})$ be a bivariate normal random vector. Then:*

1. *For any $c_1, c_2 \in \mathbb{R}$, the random variable $c_1Y_1 + c_2Y_2$ is one–dimensional normal, or constant.*

2. *If $\operatorname{corr}(Y_1, Y_2) = 0$, then Y_1 and Y_2 are independent.*

3. *The regression $\mathbb{E}(Y_2 \mid Y_1)$ coincides with the linear regression of Y_2 on Y_1, i.e.,*

$$\mathbb{E}(Y_2 \mid Y_1) = \frac{b}{a} Y_1 + \left(\mu_2 - \frac{b}{a}\mu_1\right), \qquad \textit{where} \quad \underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \underline{\underline{\Sigma}} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

Regarding visualization: what does the density of a normal distribution look like, for example in the bivariate case?

In the standard case, it is a “hill” around the origin (as in the one–dimensional case), rotationally symmetric, i.e., its contour lines are circles. In the nonstandard case, the contour lines become ellipses. Thus a nonstandard normal distribution need not be rotationally symmetric, but remains symmetric with respect to the principal axes of the ellipse(s). Let us examine one such ellipse.

For simplicity, assume $\underline{\mu} = \underline{0}$, so that the center of the ellipse is at the origin. The principal axes are perpendicular, so there exists an orthogonal transformation $\underline{\underline{U}} \in \mathbb{R}^{2\times 2}$ that maps the principal axes to the coordinate axes. It can be computed that then $\underline{\underline{U}} \cdot \underline{Y} \sim N(\underline{0}, \underline{\underline{D}})$, where $\underline{\underline{D}}$ is a diagonal matrix. By point 2 of the corollary, the coordinates of $\underline{\underline{U}} \cdot \underline{Y}$ are independent. In summary, in an appropriate coordinate system, every normal distribution consists of independent one–dimensional normal random variables.

The standard deviations of the independent random variables obtained by diagonalization have already appeared implicitly in the formula of the normal distribution: if $\underline{\underline{D}} = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$, then $\det(\underline{\underline{\Sigma}})^{1/2}$ in the density is exactly $\sigma_1 \cdot \sigma_2$, the product of the standard deviations. The determinant of the covariance matrix is invariant under orthogonal transformations, so it does not matter whether we speak of the covariance matrix of $\underline{Y}$ or of the transformed $\underline{\underline{U}} \cdot \underline{Y}$. More visually, this measures the ratio of the area of the ellipse to the area of the unit circle.

In multivariate distributions, besides the determinant of the covariance matrix, another quantity used to measure (total) variance is the trace $\mathrm{Tr}(\underline{\underline{\Sigma}})$. In terms of the standard deviations of the diagonalized variable, $\mathrm{Tr}(\underline{\underline{\Sigma}}) = \sigma_1^2 + \sigma_2^2$. Intuitively, this measures the average squared distance of $\underline{Y}$ from $\underline{\mu}$.