Lecture notes for probability part of MASD

Torben Krüger

1 Basic Concepts

Here we discuss the content of **Section 1** from "Introduction to probability, statistics, and random processes".

We introduce the basic mathematical model for probability theory. The setup is formulated in sufficient generality to ensure that essentially any question you may have about a probabilistic experiment can be formulated within it.

1.1 Basic set theory

We first recall some basic notions of set theory. Sets are the most fundamental objects in mathematics. They are simply a collection of objects (called elements).

DEFINITION 1.1 (Set). A set A is a collection of elements x. We write $x \in A$ if x is an element of A and write $x \notin A$ if x is not an element of A. The empty set $\emptyset$ does not contain any elements by definition ($x \notin \emptyset$ holds for all x). We write $A \subset B$ and say that A is a subset of B if all elements of A are also elements of B ($x \in A$ implies $x \in B$).

We have several ways of denoting specific sets that we want to talk about. Here we will use the following notations for sets:

1. List of elements in case the set is finite, e.g. $\heartsuit \in \{\spadesuit, \heartsuit, \clubsuit\}$ and $7 \notin \{1, 2, 3, 4, 5\}$.
2. Standard sets $\mathbb{R} = \{\text{real numbers}\}$, $\mathbb{N} = \{\text{positive integers}\}$, $\mathbb{Z} = \{\text{all integers}\}$.
3. Sets of elements with specified properties:

$$\{\text{all elements } x \in A \text{ with property } \mathcal{P}\} := \{x \in A : x \text{ satisfies property } \mathcal{P}\}$$

   For example we write $\{1, 2, 3, 4\} = \{x : x \in \mathbb{N} \text{ and } x \leq 4\} = \{x \in \mathbb{N} : x \leq 4\}$.
4. List when the set is countably infinite (can be enumerated by natural numbers), e.g.

$$\{\text{even positive integers}\} = \{2, 4, 6, 8, \dots\} = \{2n : n \in \mathbb{N}\} \text{ and } \{x_1, x_2, \dots\} = \{x_n : n \in \mathbb{N}\}$$

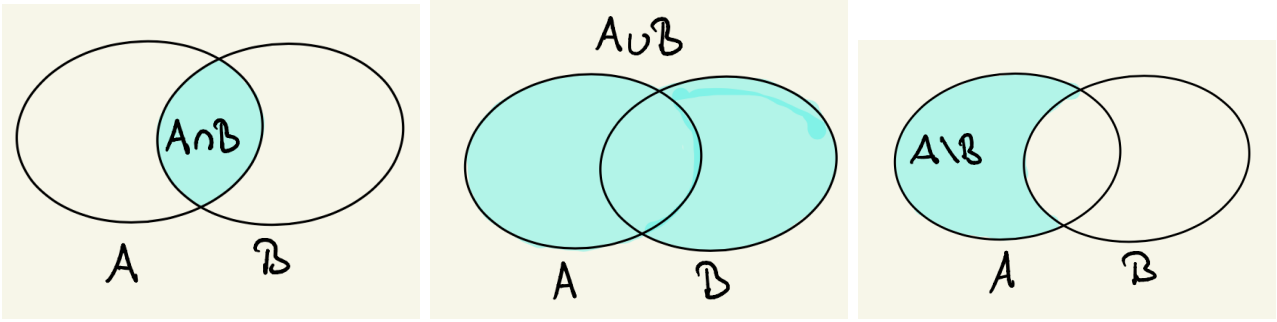
In order to work with sets we have to perform certain operations on them. The following elementary set operations can be performed:

1. Intersection: $A \cap B = \{x : x \in A \text{ and } x \in B\}$ and $\cap_{i=1}^n A_i = \{x : x \in A_i \text{ for all } i \leq n\}$
2. Union: $A \cup B = \{x : x \in A \text{ or } x \in B\}$ and $\cup_{i=1}^n A_i = \{x : x \in A_i \text{ for at least one } i \leq n\}$
3. Complement: $A \setminus B = A - B = \{x : x \in A \text{ and } x \notin B\}$

4. Cartesian product: $A \times B = \{(a, b) : a \in A, b \in B\}$ and $A_1 \times \cdots \times A_n = \{(x_1, \ldots, x_n) : x_i \in A_i\}$. If all $A_i = A$ are the same, then we also write $A^n = \{(x_1, \ldots, x_n) : x_i \in A\}$.

EXAMPLE 1.2. Here are a few examples of set operations:

1. $\{\spadesuit, \heartsuit, \clubsuit\} \cup \{\spadesuit, \diamondsuit\} = \{\spadesuit, \heartsuit, \clubsuit, \diamondsuit\}$, $\cup_{i=1}^{\infty}\{i\} = \mathbb{N}$ and $\cup_{i=1}^{n}\{1, \ldots, i\} = \{1, \ldots n\}$
2. $\{\spadesuit, \heartsuit, \clubsuit\} \cap \{\spadesuit, \diamondsuit\} = \{\spadesuit\}$, $\{\heartsuit, \clubsuit\} \cap \{\spadesuit, \diamondsuit\} = \emptyset$, $\mathbb{N} \cap [2, 3] = \{2, 3\}$ and $\cap_{i=1}^{\infty}\{1, \ldots, i\} = \{1\}$
3. $\{\spadesuit, \heartsuit, \clubsuit\} \setminus \{\spadesuit, \diamondsuit\} = \{\heartsuit, \clubsuit\}$ and $\mathbb{Z} \setminus \mathbb{N} = \{0, -1, -2, \ldots\}$
4. $\{\spadesuit, \heartsuit, \clubsuit\} \cup \{\spadesuit, \diamondsuit\} = \{(\spadesuit, \spadesuit), (\spadesuit, \diamondsuit), (\heartsuit, \spadesuit), (\heartsuit, \diamondsuit), (\clubsuit, \spadesuit), (\clubsuit, \diamondsuit)\}$



We denote by $|A|$ the size (or cardinality) of a set A. If A has finitely many elements, then $|A|$ is simply the number of elements. If A has infinitely many elements we write $|A| = \infty$. In this case there are two distinct situations. Either A is countably infinite, i.e. its elements can be enumerated by the natural numbers $\mathbb{N}$, or A is uncountable. In the latter case there are too many elements in A to enumerate them by $\mathbb{N}$.

EXAMPLE 1.3. Here are some examples of cardinalities for sets we will encounter.

1. The sets $A_1 = \{x \in \mathbb{N} : x \leq 10\}$, $A_2 = \{\spadesuit, \heartsuit, \clubsuit\}$, $A_3 = \emptyset$ are finite with cardinalities $|A_1| = 10$, $|A_2| = 3$ and $|A_3| = 0$.
2. The sets $\mathbb{N}$, $\mathbb{Q}$, $\mathbb{N} \times \mathbb{N}$ and $\{2x : x \in \mathbb{N}\}$ are all infinite and countable.
3. The sets $\mathbb{R}$, $[1, 2]$, $\{x : x \subset \mathbb{N}\} = \{\text{all subsets of } \mathbb{N}\}$ are all infinite and uncountable.

Finally we will need the concept of a function f that assigns for every element x in a set A exactly one element $f(x)$ in a set B. We write $f : A \to B$. Here, A is called the domain of f and B its co-domain. The set $\text{Range}(f) = R_f = \{f(x) : x \in A\} \subset B$ is called the image (or range) of f.

EXAMPLE 1.4. Here are some examples of functions:

1. The functions $f, g, h : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$, $g(x) = 4$ and $h(x) = 2^x$ all assign the value 4 to $2 \in \mathbb{R}$, i.e. $f(2) = g(2) = h(2) = 4$.
2. The addition function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ defined by $f(x, y) = x + y$ has $R_f = \{x \in \mathbb{N} : x \geq 2\}$ as its range.

1.2 Probability spaces

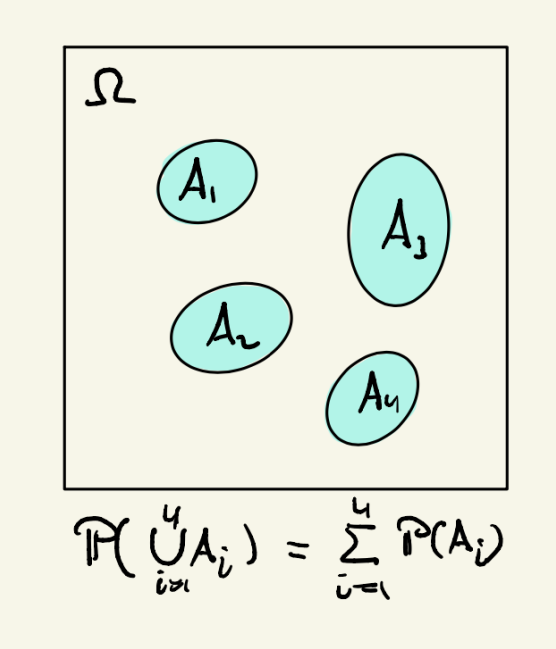
To mathematically model probabilistic experiments we need a set that encodes all conceivable outcomes of our experiment. We call this set S the *sample space* or *universe*. *Events* A are subsets of S, i.e. $A \subset S$. We will see that there is some flexibility in the choice of S, provided it contains enough complexity to encode all experimental outcomes. It is more important what probabilities we assign to particular events, rather than how the universe is chosen exactly. Think of how there is some flexibility to encode certain data in a program.

EXAMPLE 1.5 (Rolling a die). A natural choice for the sample space associated to rolling a die is $S = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. These are all possible outcomes. The event "The die roll shows more than three eyes" is the subset $\{\boxdot, \boxdot, \boxdot\}$ of S. The event "We rolled a 6" is $\{\boxdot\}$. More mathematically convenient we can encode the same outcomes as $S = \{1, 2, 3, 4, 5, 6\}$. If we roll the die three times the sample space may be chosen as $S = \{1, 2, 3, 4, 5, 6\}^3$, encoding the outcome of each roll. The event "The roll sum is 5" is $\{(1, 1, 3), (1, 3, 1), (3, 1, 1), (1, 2, 2), (2, 1, 2), (2, 2, 1)\} \subset S$.

DEFINITION 1.6 (Probability space). A *probability space* $(S, \mathbb{P})$ consists of a sample space (universe) S and an assignment $\mathbb{P}$ that assigns to each event A a number $\mathbb{P}(A)$ that we call the probability of A. The assignment $\mathbb{P}$ is called a (probability) distribution and has to satisfy the following rules:

1. Non-negativity of probabilities: $\mathbb{P}(A) \geq 0$ for all events $A \subset S$.
2. Total probability of the universe: $\mathbb{P}(\text{all possible outcomes}) = \mathbb{P}(S) = 1$
3. Additivity of probabilities: If $A_1, A_2, \ldots$ are disjoint (mutually exclusive, i.e. $A_i \cap A_j = \emptyset$ if $i \neq j$) events, then

$$\mathbb{P}\big(\cup_{i=1}^{\infty} A_i\big) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots = \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{1.1}$$



By choosing $A_i = \emptyset$ for all i in the second property, we conclude $\mathbb{P}(\emptyset) = 0$. In particular, the additivity in (1.1) implies also that

$$\mathbb{P}\big(\cup_{i=1}^{n} A_i\big) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i) \tag{1.2}$$

for finitely many events $A_1, \ldots, A_n$ because we can choose $A_i = \emptyset$ for all $i > n$. Furthermore, if the universe is countable (finite or infinite), then knowing $\mathbb{P}$ is equivalent to knowing $\mathbb{P}(\{x\})$ for all $x \in S$ because

$$\mathbb{P}(A) = \sum_{x \in A} \mathbb{P}(\{x\})\,.$$

REMARK 1.7 (Probability as a volume). Note that the properties of $\mathbb{P}$ are very similar to what you would expect for assigning a volume, provided the volume of the entire space S is normalised to 1. Indeed, volume and integration have the same axiomatic formulation as probabilities.

EXAMPLE 1.8 (Rolling a die continued). The probabilities assigned to the events in the sample space $S = \{1, 2, 3, 4, 5, 6\}$ of our (fair) die roll are determined by $\mathbb{P}(\{x\}) = 1/6$ for all $x \in S$. This leads e.g. to

$$\mathbb{P}(\text{The die roll shows more than three eyes}) = \mathbb{P}(\{4, 5, 6\}) = \mathbb{P}(\{4\}) + \mathbb{P}(\{5\}) + \mathbb{P}(\{6\}) = \frac{3}{6} = \frac{1}{2}\,.$$

If we roll the die three times with sample space $S = \{1, 2, 3, 4, 5, 6\}^3$ we have $\mathbb{P}(\{x\}) = \left(\frac{1}{6}\right)^3 = \frac{1}{216}$ for all $x \in S$. The reason for this is that the probabilities of independent events are multiplied. In this case we have e.g.

$$\mathbb{P}(\text{The roll sum is 5}) = \mathbb{P}(\{(1, 1, 3), (1, 3, 1), (3, 1, 1), (1, 2, 2), (2, 1, 2), (2, 2, 1)\}) = \frac{6}{216} = \frac{1}{36}\,.$$

More generally, to model what happens when we perform several independent experiments we need the following notion.

DEFINITION 1.9 (Independent events). Let S be a probability space with distribution $\mathbb{P}$. Two events $A, B \subset S$ are called independent if

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

Sometimes it is easier to compute the probability that an event does not occur rather than the probability that it does. In the following we will denote by $A^c = \overline{A} = S \setminus A$ the event that A *does not occur* or the *complement of* A. For computing complements we have De Morgan's laws, $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$. **Please convince yourself that these rules are true!**

EXAMPLE 1.10 (Maximum of dice rolls). What is the probability that the maximum of two dice rolls is larger than 4? It is easer to compute the probability of the complementary event, namely that the maximum of two dice rolls is smaller than 5, because this means that both rolls have to be smaller than 5. This has a probability of $(4/6)^2 = \frac{4}{9}$. Thus, $\mathbb{P}(\text{Maximum roll larger than 4}) = 1 - \frac{1}{9} = \frac{5}{9}$.

Of course we could have arrived at this result also by counting the outcomes in the event "Maximum roll larger than 4" in the sample space $\{1, 2, 3, 4, 5, 6\}^2$, which are

$$(1, 5), (2, 5), (3, 5), (4, 5), (5, 5), (6, 5), (1, 6), (2, 6), (3, 6), (4, 6), (5, 6), (6, 6),$$
$$(5, 1), (5, 2), (5, 3), (5, 4), (6, 1), (6, 2), (6, 3), (6, 4).$$

Then we arrive at $\mathbb{P}(\text{Maximum roll larger than 4}) = \frac{20}{36} = \frac{5}{9}$. But this is much more tedious.

In the following theorem we prove properties that follow from the axioms of probability distributions and show how probabilities behave under basic set operations.

THEOREM 1.11 (Properties of probability). Let S be a probability space with probability distribution $\mathbb{P}$. Then the following holds true:

1. Complement probability: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ for any event A.
2. Inclusion-exclusion principle: For any events $A_1, \ldots, A_n$ we have

$$\mathbb{P}(\cup_{i=1}^n A_i) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{I:|I|=k} \mathbb{P}(\cap_{i \in I} A_i)\,.$$

Here the sum $\sum_{I:|I|=k}$ is over all subsets $I \subset \{1, \ldots, n\}$ with k elements.

Proof. To prove Property 1 we use the addition rule (1.2) to get the first equality in

$$\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(A \cup A^c) = \mathbb{P}(S) = 1\,.$$

For the last equality we used Property 1 from Definition 1.6.

Property 2 is proved by induction over n. **Recall how induction proofs work!** For the case $n = 1$ there is nothing to show since both sides of the claimed identity are equal. We consider the case $n = 2$ since we will use the result in the following. We have for two events A, B the identity

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\,. \tag{1.3}$$

Let us prove (1.3). Since $A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)$ is a union of three disjoint sets, we conclude

$$\begin{aligned}\mathbb{P}(A \cup B) &\overset{(1)}{=} \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B)\\ &\overset{(2)}{=} \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B)\\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\,.\end{aligned}$$

Here in (1) we used the addition rule for probabilities. In (2) we used $\mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)$ (as a consequence of the addition rule for probabilities) in the form $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$ for any events A and B.

Now let us assume that Property 2 is proven for $n - 1$ sets. Then we compute

$$\begin{aligned}\mathbb{P}(\cup_{i=1}^n A_i) &\overset{(1)}{=} \mathbb{P}(\cup_{i=1}^{n-1} A_i) + \mathbb{P}(A_n) - \mathbb{P}(A_n \cap \cup_{i=1}^{n-1} A_i)\\ &\overset{(2)}{=} \sum_{k=1}^{n-1}(-1)^{k+1} \sum_{I:|I|=k} \mathbb{P}(\cap_{i\in I} A_i) + \mathbb{P}(A_n) - \mathbb{P}(\cup_{i=1}^{n-1}(A_i \cap A_n))\\ &\overset{(3)}{=} \sum_{k=1}^{n-1}(-1)^{k+1} \sum_{I:|I|=k} \mathbb{P}(\cap_{i\in I} A_i) + \mathbb{P}(A_n) - \sum_{k=1}^{n-1}(-1)^{k+1} \sum_{J:|J|=k} \mathbb{P}(\cap_{j\in J}(A_j \cap A_n))\,,\end{aligned} \tag{1.4}$$

where in (1) we used the case for two sets $A = \cup_{i=1}^{n-1} A_i$) and $B = A_n$ from (1.3) and in (2) as well as (3) we used the case for $n - 1$ sets. In the last expression I and J are both summed over all subsets of $\{1, \dots, n-1\}$ of size k. We rewrite the last summand as

$$-\sum_{k=1}^{n-1}(-1)^{k+1} \sum_{J:|J|=k} \mathbb{P}(\cap_{j\in J}(A_j \cap A_n)) = \sum_{k=2}^{n}(-1)^{k+1} \sum_{I:|I|=k, n\in I} \mathbb{P}(\cap_{j\in J} A_j)\,, \tag{1.5}$$

where $I \subset \{1, \dots, n\}$ of size k contains n in the last sum. Putting (1.4) and (1.5) together finishes the proof. □

1.3 The uniform distribution

We will now present some examples that illustrate how to construct the mathematical setup for a probabilistic model, i.e. how to choose a suitable probability space.

EXAMPLE 1.12 (Uniform distribution). On a finite universe S we consider the *uniform distribution* defined through

$$\mathbb{P}(A) := \frac{|A|}{|S|} = \frac{\text{number of outcomes in } A}{\text{number of outcomes in universe}}\,.$$

EXAMPLE 1.13 (Coin toss). The experiment of tossing a coin n times is modelled by $S = \{0, 1\}^n$ (1 for heads and 0 for tails) with the uniform distribution. The uniform distribution is chosen because no outcome is special and, thus, they are all equally likely. What is the probability that heads is observed at most twice during our experiment? The event $A =$"heads is observed at most twice" is a disjoint union of three events, namely $A = A_0 \cup A_1 \cup A_2$, where $A_0 =$"heads is never observed", $A_1 =$"heads is observed exactly once", $A_2 =$"heads is observed exactly twice". We count the number of elements. First, $A_0 = \{(0, \dots, 0)\}$ contains only one element. Then $|A_1| = n$ because exactly one entry of $(x_1, \dots, x_n) \in S$ is equal to 1. Finally, $|A_2| = n(n-1)$ because after choosing the first nonzero

entry (out to n choices) one can choose the second nonzero entry out of the remaining $n-1$ choices. Altogether we find

$$\mathbb{P}(A) = \mathbb{P}(A_0) + \mathbb{P}(A_1) + \mathbb{P}(A_2) = \frac{|A_0| + |A_1| + |A_2|}{|S|} = \frac{1 + n + n(n-1)}{2^n} = \frac{n^2+1}{2^n}\,.$$

Up to now all our examples were discrete models, i.e. the underlying sample spaces were countable (even finite). However, sometimes random experiments have continuous outcomes. For example, let $(x, y) \in [0, 1]^2$ denote the (x, y)-coordinates on a quadratic map and $S \subset [0, 1]^2$ the area on the map covered by water. During a thunderstorm (covering the whole area) we model the position in S where a lightning will hit next. All points are equally likely. On the other hand, the probability that a lightning will hit exactly a specific point $(x, y) \in S$ vanishes, i.e. $\mathbb{P}(\{(x, y)\}) = 0$, in contrast to discrete models.

EXAMPLE 1.14 (Continuous uniform distribution). On a bounded subset S of $\mathbb{R}^d$ with positive volume (length, area or volume for $d = 1, 2, 3$, respectively) we define the uniform distribution

$$\mathbb{P}(A) = \frac{\mathrm{Vol}(A)}{\mathrm{Vol}(S)} = \frac{\text{volume of } A}{\text{volume of } S}\,.$$

REMARK 1.15 (Warning). There are some pathological examples of events (with abstract constructions) that cannot be assigned a probability in the case of continuous uniform distributions. These are non-measurable sets. But we will always ignore this issue in the following since it is irrelevant for applications.

EXAMPLE 1.16 (Lightning hits). Suppose 1/5 of the area on the map encoded by coordinates in $[0, 1]^2$ is covered by water and your rowing boat (which is on the water) covers an area of 1/1000 on the same map. How likely is it that: a) the next lightning that hits the water will also hit your boat? b) this lightning will hit the point at the very front of the boat? Assuming a continuous uniform distribution, the answers are a) 1/200 and b) 0.

1.4 Conditional probability

DEFINITION 1.17 (Conditional probability). Fix an event B with nonvanishing probability $\mathbb{P}(B) > 0$ on a probability space with probability distribution $\mathbb{P}$. Then we define

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

for any event A. We call $\mathbb{P}(A|B)$ the 'conditional probability of A, given B'.

We will see in the exercises that $A \mapsto \mathbb{P}(A|B)$ is a probability distribution.

THEOREM 1.18 (Law of total probability). Let $B_1, \ldots, B_n$ be a partition of the sample space (i.e. B_i are disjoint and $\cup_{i=1}^n B_i = S$) with $\mathbb{P}(B_i) > 0$. Then

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

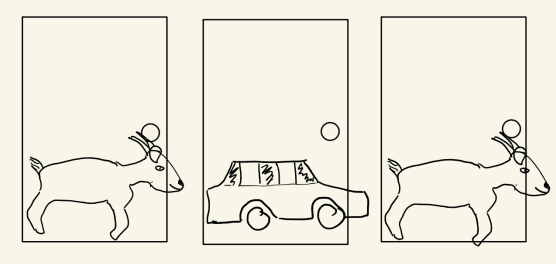
holds for any event A.

Proof. We compute

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

where in the first equality we used the definition of conditional probability and in the last equality we used that $B_1, \ldots, B_n$ is a partition of the sample space. □

EXAMPLE 1.19 (Let's Make a Deal). The rules of a popular TV game show from the 60's are the following: You are presented with three closed doors. Behind two of them there is a goat and behind the other is the price, a brand new car. You pick a door. Then the host (who knows what is behind each door) opens one of the other two doors, behind which there is a goat. Now you can play. You can either stick to your original decision or switch to the other remaining closed door. Should you switch?



The answer is found using Theorem 1.18. Let $X_1, X_2 \in \{g, c\}$ (for goat and car) be your picks in the first and second round. We consider the two possible strategies of play. The first is to stick with your original choice. Then we find

$$\begin{aligned}\mathbb{P}(X_2 = c) &= \mathbb{P}(X_2 = c|X_1 = c)\mathbb{P}(X_1 = c) + \mathbb{P}(X_2 = c|X_1 = g)\mathbb{P}(X_1 = g) \\ &= 1 \cdot \frac{1}{3} + 0 \cdot \frac{2}{3} = \frac{1}{3}\,.\end{aligned}$$

The second strategy is to switch. In this case

$$\begin{aligned}\mathbb{P}(X_2 = c) &= \mathbb{P}(X_2 = c|X_1 = c)\mathbb{P}(X_1 = c) + \mathbb{P}(X_2 = c|X_1 = g)\mathbb{P}(X_1 = g) \\ &= 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}\,.\end{aligned}$$

Thus, although somewhat counterintuitive, switching is better.

THEOREM 1.20 (Bayes' rule). Let $B_1, \ldots, B_n$ be a partition of the sample space with $\mathbb{P}(B_i) > 0$. Then

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{n} \mathbb{P}(A|B_j)\mathbb{P}(B_j)}\,,$$

hold for any event A with $\mathbb{P}(A) > 0$.

Proof. Both sides are identical to $\mathbb{P}(A \cap B_i)$ when we multiply them by $\mathbb{P}(A)$ and use Theorem 1.18. □

EXAMPLE 1.21 (Medical testing). Bayes' rule has important applications in medical testing. We consider the following setup:

- Let S be all people of the population under consideration.
- Let $I \subset S$ be the set of infected and $H = I^c$ all healthy people.
- Let $P \subset S$ be the people for which the test is positive and $N = P^c$ be the people for which the test is negative.

We assume a uniform distribution on the population and are given the following information (from a recent event).

- The ratio of the infected to the total population, namely $\mathbb{P}(I) = 0.003$.
- The sensitivity of the test $\mathbb{P}(P|I) = 0.95$ and the specificity of the test $\mathbb{P}(N|H) = 0.99$.

After being tested positive we want to know how likely it is that we are actually infected. First we compute $\mathbb{P}(P|H) = 1-\mathbb{P}(N|H) = 0.01$ and $\mathbb{P}(H) = 1-\mathbb{P}(I) = 0.997$. Then we use Bayes' rule, namely

$$\mathbb{P}(I|P) = \frac{\mathbb{P}(P|I)\mathbb{P}(I)}{\mathbb{P}(P|I)\mathbb{P}(I)+\mathbb{P}(P|H)\mathbb{P}(H)} = \frac{0.95\times 0.003}{0.95\times 0.003+0.01\times 0.997} \approx 0.22\,.$$

Thus, a positive test is pretty useless to determine our own status. It is just an indication that we have to be monitored more closely and get tested again. The test should however ensure that somebody infected will be monitored, i.e. that an infected person is not released because of a negative test. This is the case here because

$$\mathbb{P}(I|N) = \frac{\mathbb{P}(N|I)\mathbb{P}(I)}{\mathbb{P}(N|I)\mathbb{P}(I)+\mathbb{P}(N|H)\mathbb{P}(H)} = \frac{(1-\mathbb{P}(P|I))\mathbb{P}(I)}{(1-\mathbb{P}(P|I))\mathbb{P}(I)+\mathbb{P}(N|H)(1-\mathbb{P}(I))} \approx 0.00015\,.$$

2 Combinatorics

Here we discuss the content of **Section 2** from "Introduction to probability, statistics, and random processes".

We will now discuss how to compute probabilities in several discrete experiments. Many questions about probabilities can be reduced to the examples from below by analogy. The first basic principle of combinatorics is the following: When an experiment can be performed as two successive steps with a and b possible outcomes, respectively, then the experiment has ab possible outcomes (multiplication principle).

Example 2.1 (Multiplication principle). You have to choose a username consisting of two lowercase letters (a to z), followed by three digits $(0, 1, \ldots, 9)$ and one of the letters A, B or C. An example is $ky649B$. How many possible username choices are there? By the multiplication principle there are $26\times 26\times 10\times 10\times 10\times 3 = 2.028000\times 10^6$ choices.

2.1 Sampling types

In many random experiments we consecutively sample (or draw) elements k times from a given set with n elements and want to know the number of all possible outcomes of the sampling. We distinguish the following types of sampling experiments.

- *With or without replacement:* In the case of sampling with replacement we imagine drawing from a set and putting back (replace) the element we picked before drawing the next one. Here repeated drawing of the same element is allowed because we put back the drawn element into the pool of possible choices. If we sample without replacement, repetitions are not allowed, i.e. after drawing an element we draw the next one from the remaining subset.

- *Ordered or unordered:* In ordered sampling we record the order in which the elements have been drawn, while in unordered sampling the ordered is irrelevant.

For the set $\{a, b, c\}$ and a sample size of 2 we get the following possible outcomes:

- Ordered sampling with replacement: $a,a \mid a,b \mid a,c \mid b,a \mid b,b \mid b,c \mid c,a \mid c,b \mid c,c$

- Ordered sampling without replacement: $a,b \mid a,c \mid b,a \mid b,c \mid c,a \mid c,b$

- Unordered sampling without replacement: $a,b \mid a,c \mid b,c$

- Unordered sampling with replacement: $a,a \mid a,b \mid a,c \mid b,b \mid b,c \mid c,c$

2.2 Ordered sampling with replacement

EXAMPLE 2.2 (Ordered sampling with replacement). Consider an urn with n enumerated balls in it. At random we pick a ball, write down its number and return the ball to the urn. Then we pick another ball. How many possible outcomes of the experiments are there if we repeat this step k times? The answer is n^k by multiplying the possibilities of each step. Thus, the probability of drawing a specific sequence of numbers is n^{-k}.

EXAMPLE 2.3 (Cardinality of power sets). Let S be a finite set with $|S| = n$. How many subsets does S have? We can think of this as sampling n times from a set of two values ("element is in the subset" or "element is not in the subset") and get 2^n as the number of all subsets of S.

2.3 Ordered sampling without replacement

DEFINITION 2.4 (Factorials). We define $n! := n \cdot (n-1) \dots 2 \cdot 1$ (speak "n factorial"). In particular, $0! := 1$, $1! := 1$, $2! := 2$, $3! := 6$, etc.

EXAMPLE 2.5 (Ordered sampling without replacement). Consider an urn with n enumerated balls. Now we choose the balls without returning them and repeat k times. In that case we have n possibilities in the first step, $n-1$ possibilities in the second step and so forth. In this case our experiment has $n(n-1)(n-2)\dots(n-k+1) = \frac{n!}{(n-k)!}$ possible outcomes after k steps.

DEFINITION 2.6 (Permutations). A permutation $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ of the set $\{1, \dots, n\}$ is a one-to-one map (or function) of $\{1, \dots, n\}$ to itself. The set of all such permutations is denoted S_n.

We often use the following short hand notation for a permutation $\sigma \in S_n$:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & \dots & n \\ \sigma(1) & \sigma(2) & \sigma(3) & \dots & \sigma(n) \end{pmatrix}.$$

For example, the permutation

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 3 & 2 \end{pmatrix}$$

satisfies $\sigma(1) = 4, \sigma(2) = 1, \sigma(3) = 3, \sigma(4) = 2$.

EXAMPLE 2.7 (Permutations). How many permutations of $\{1, \dots, n\}$ are there? We have n choices for $\sigma(1)$, then $n-1$ choices for $\sigma(2)$ and so on. This is the same as ordered sampling without replacement and we get $|S_n| = n!$.

EXAMPLE 2.8 (Card shuffling). We shuffle a deck of 52 cards. How likely is it that, after shuffling, the order of the cards has been exactly reversed? A natural choice for the associated probability space is S_{52} with the uniform distribution, where the outcome $\sigma \in S_{52}$ means that card number $\sigma(i)$ is at position i in the shuffled deck. Thus,

$$\mathbb{P}(\text{order exactly reversed}) = \frac{1}{|S_{52}|} = \frac{1}{52!} \approx 1.2398 \times 10^{-68}.$$

2.4 Unordered sampling without replacement

EXAMPLE 2.9 (Adjusting for overcounting). Suppose you have n friends but can only invite k ($\leq n$) of them for dinner. How many choices of distinct dinner parties do you have? If you simply sample k friends out of n without replacement, then some outcomes result in the same collection of friends coming to dinner because you choose the same friends in a different order. The number of possible orderings is $k!$ and, thus, the number of different dinner constellations is $n(n-1)(n-2)\dots(n-k+1)/k!$.

DEFINITION 2.10 (Binomial coefficient). We define the binomial coefficient 'n choose k' as

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}$$

EXAMPLE 2.11 (Counting subsets). How many subsets of size k does $\{1, 2, \ldots, n\}$ have? This is the same problem as choosing k out of n friends without caring about the order. Thus, the answer is $\binom{n}{k}$.

THEOREM 2.12 (Binomial theorem). Let x, y be real numbers and $n \in \mathbb{N}$. Then

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

Proof. We write $(x+y)^n$ as an n-fold product and multiply out

$$\underbrace{(x+y)\ldots(x+y)}_{n \text{ times}} = \sum_{z_1 \in \{x,y\}} \cdots \sum_{x_n \in \{x,y\}} z_1 \ldots z_n = \sum_{z \in \{x,y\}^n} z_1 \ldots z_n \,. \tag{2.1}$$

For each x_i we can either have $x_i = x$ or $x_i = y$. Since multiplication is commutative many monomials from the sum over $z = (z_1, \ldots, z_n)$ from (2.1) are identical. In fact, what matters is the number of times x appears in the product. Thus, we find

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

where $C(n,k)$ is the number of $z \in \{x,y\}^n$ with exactly k appearances of x. This exactly the same as choosing k elements from $\{1, \ldots, n\}$ without caring about the order. Thinking back to Example 2.9 shows the theorem. □

2.5 Unordered sampling with replacement

EXAMPLE 2.13 (Distributing coins). Suppose we have n friends lined up (order matters) and want to distribute k coins among them. The coins all have equal value and, thus, each friend only cares about the number of coins he/she receives. How many possible outcomes of the distribution process are there? We can think of separating our friends by walls, symbolised by "|". The coins between these walls have the symbol "∘". For example with $n = 4$ and $k = 6$ we need 3 separating walls and the configuration

$$\circ\,\circ\,\circ|\,|\,\circ\,|\,\circ\,\circ$$

means that the first friend gets 3 coins, the second none, the third 1 coin and the forth gets 2. Any sequence of $n-1$ walls and k coins is a valid configuration. How many such sequences are there? This is the same problem as counting the number of subsets of size $n-1$ (position of the walls) out of $n+k-1$ (total number of symbols), i.e. $\binom{n+k-1}{n-1}$.

EXAMPLE 2.14 (Unordered sampling with replacement). Suppose we have a set of n elements $A = \{a_1, \ldots, a_n\}$. We sample k times without caring about the order and allowing for repetition. In particular, we only keep track of the number $x_i \geq 0$ of how often a_i was drawn. Thus, each outcome is represented as $(x_1, x_2, \ldots, x_n)$ such that $x_1 + x_2 + \cdots + x_n = k$. This is the same problem as distributing k coins among n friends and we find $\binom{n+k-1}{n-1}$ possible outcomes.

3 Random variables and their distributions

Here we discuss the content of **Section 3 - 4**, except for Sections 3.2.3, 3.2.4 and 4.1.2, from "Introduction to probability, statistics, and random processes".

3.1 Basic definitions

DEFINITION 3.1 (Random variables). A random variable is a function on the sample space S with values in $\mathbb{R}$, i.e. $X : S \to \mathbb{R}$.

EXAMPLE 3.2 (Row sum). The probability space associated to rolling two dice is $S = \{1, 2, \dots, 6\}^2$ with the uniform distribution. The row sum $X : S \to \mathbb{R}$ defined through $X((s_1, s_2)) = s_1 + s_2$ is a random variable. In particular, the event $A =$"The row sum is 5" can be expressed in terms of X via

$$A = \{s \in S : X(s) = 5\} = \{(s_1, s_2) \in S : x_1 + x_2 = 5\} = \{(1, 4), (4, 1), (2, 3), (3, 2)\} = \{X = 5\}\,.$$

Events that can be expressed in terms of random variables are often written in a short hand fashion, namely we write $\{X \in A\}$ instead of $\{s \in S : X(s) \in A\}$ for some $A \subset \mathbb{R}$ and a random variable $X : S \to \mathbb{R}$. Multiples, sums, products, limits and functions of random variables are again random variables, i.e. αX, $X+Y$, XY, $\lim_{n\to\infty} X_n$ and $f(X)$ are random variables for $\alpha \in \mathbb{R}$, random variables X, Y, X_n and all (reasonable) functions f.

EXAMPLE 3.3 (Coin toss experiment). The experiment of tossing a coin n times is modelled by $S = \{0, 1\}^n$ (1 for heads and 0 for tails) with the uniform distribution. The discrete random variable that projects onto the i-th entry $X_i : S \to \{0, 1\}, (x_1, \dots, x_n) \mapsto x_i$ encodes the result of the i-th coin toss. It takes two values $R_{X_i} = \{0, 1\}$. Both values are equally likely in this experiment, i.e. $\mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = 1/2$. We check this as follows:

$$\mathbb{P}(X_i = y) = \frac{|\{(x_1, \dots, x_n) \in S : x_i = y\}|}{|S|} = \frac{2^{n-1}}{2^n} = \frac{1}{2}$$

for $y = 0, 1$. The total number of heads in the experiment is $X = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^n X_i$ with range $R_X = \{0, 1, \dots, n\}$.

We will now associate to a random variable $X : S \to \mathbb{R}$ a probability distribution on its range R_X.

DEFINITION 3.4 (Distributions). We call the distribution $\mathbb{P}_X$ on R_X given by $\mathbb{P}_X(A) = \mathbb{P}(X \in A)$ for $A \subset R_X$ the distribution of X and write $X \sim \mathbb{P}_X$ for 'X has distribution $\mathbb{P}_X$'.

PROPOSITION 3.5 (Distribution of random variables). Let $(S, \mathbb{P})$ be a probability space and $X : S \to \mathbb{R}$ a random variable. Then $(R_X, \mathbb{P}_X)$ is a probability space.

Proof. We check the three defining properties of a probability space. First, we have the non-negativity $\mathbb{P}_X(A) = \mathbb{P}(X \in A) \geq 0$ because $\mathbb{P}$ assigns non-negative values. Second, we compute

$$\mathbb{P}_X(R_X) = \mathbb{P}(X \in R_X) = \mathbb{P}(S) = 1\,.$$

To check additivity, let $B_1, B_2, \dots$ be disjoint subsets of R_X. Then

$$\mathbb{P}_X\big(\cup_i B_i\big) = \mathbb{P}\big(X \in \cup_i B_i\big) = \mathbb{P}\big(\cup_i \{X \in B_i\}\big) = \sum_i \mathbb{P}(X \in B_i) = \sum_i \mathbb{P}_X(B_i)\,,$$

proving that $\mathbb{P}_X$ is a probability distribution. □

EXAMPLE 3.6 (Distribution of row sum). Let X be the row sum from Example 3.2. Then $R_X = \{2, 3, \dots, 12\}$ and we determine the distribution of X by computing its value on each outcome $x \in R_X$, namely

$$\begin{aligned}
\mathbb{P}_X(\{2\}) &= \mathbb{P}(X = 2) = \mathbb{P}(\{(1, 1)\}) = \frac{1}{36}\,,\\
\mathbb{P}_X(\{3\}) &= \mathbb{P}(X = 3) = \mathbb{P}(\{(1, 2), (2, 1)\}) = \frac{1}{18}\,,\\
\mathbb{P}_X(\{4\}) &= \mathbb{P}(X = 3) = \mathbb{P}(\{(1, 3), (3, 1), (2, 2)\}) = \frac{1}{12}\,,
\end{aligned}$$

and so on.

3.2 Discrete random variables

DEFINITION 3.7 (Discrete random variables). If $R_X = \text{Range}(X)$ is countable (finite or infinite) we say that X is a discrete random variable.

For discrete random variables with values in R_X knowing $\mathbb{P}_X$ is equivalent to knowing $P_X(x) := \mathbb{P}_X(\{x\})$. The function $P_X : R_X \to [0,1]$ is called the *probability mass function* (PMF) of X. It has the normalization property $\sum_{x \in R_X} P_X(x) = 1$ and satisfies $\mathbb{P}_X(A) = \sum_{x \in A} P_X(x)$.

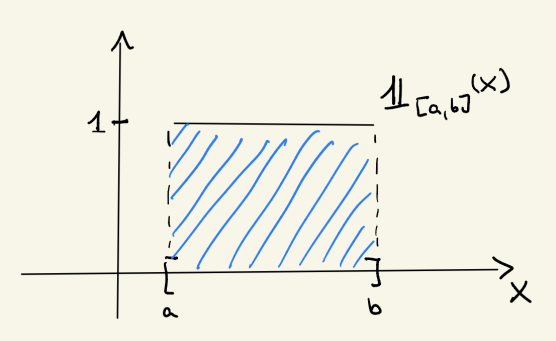
DEFINITION 3.8 (Bernoulli distribution). The distribution $\mathbb{P}$ on $\{0,1\}$ with $\mathbb{P}(\{0\}) = 1-p$ and $\mathbb{P}(\{1\}) = p$ for some $p \in (0,1)$ is called Bernoulli distribution and is denoted $\mathbb{P} = \text{Bernoulli(p)}$. A random variable X with $X \sim \text{Bernoulli(p)}$ is called Bernoulli random variable.

EXAMPLE 3.9 (Coin toss is Bernoulli). The individual coin tosses X_i from Example 3.3 are all Bernoulli(p)-distributed with $p = \frac{1}{2}$.

EXAMPLE 3.10 (Indicator random variable). For any event A with non-trivial probability on some probability space $(S, \mathbb{P})$ the indicator random variable $\mathbb{1}_A$ defined by

$$\mathbb{1}_A(x) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

is Bernoulli distributed, namely $\mathbb{1}_A \sim \text{Bern(p)}$ with $p = \mathbb{P}(A)$.



EXAMPLE 3.11 (Successes of several Bernoulli experiments). Let $X_1, \ldots, X_n$ be the result of independently performed Bernoulli experiments with fixed parameter $p \in (0,1)$, i.e. $X_i \sim \text{Bernoulli(p)}$. What is the probability of k successes, i.e. what is $\mathbb{P}(\sum_{i=1}^n X_i = k)$? To answer this question we write the event as a disjoint union and take its probability

$$\mathbb{P}\left(\sum_{i=1}^n X_i = k\right) = \mathbb{P}\left(\bigcup_x \{X = x\}\right) = \sum_x \mathbb{P}(X = x)\,.$$

Here the union and sum on the right hand side are over all $x \in \{0,1\}^n$ with $\sum_{i=1}^n x_i = k$ and $X = (X_1, \ldots, X_n)$. Since x encodes k successes and $n-k$ failures that are all independent we get $\mathbb{P}(X = x) = p^k(1-p)^{n-k}$, independent of x. Together with $\#\{x : \sum x_i = k\} = \binom{n}{k}$ we see that $\mathbb{P}(\sum X_i = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

DEFINITION 3.12 (Binomial distribution). We call the distribution $\mathbb{P} = \text{Binomial}(n,p)$ on $\{0, \ldots, n\}$ defined through the PMF

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

the Binomial distribution.

EXAMPLE 3.13 (Counting rare events). We model the requests arriving at a server and want to know how many requests to expect in a given time interval $I = [0,t]$. We assume that requests are made rarely (so that their total number remains finite) and independently of each other. We divide the time interval

I into n disjoint subintervals I_i of length t/n and assume that at most one request is made in I_i with probability $p_n = \alpha t/n$, proportional to the subinterval length. The request in I_i is then modelled by a Bernoulli random variable $X_i \sim \text{Bernoulli}(p_n)$ and we are interested in the total number of requests, i.e. the distribution of $\Sigma_n = \sum_{i=1}^n X_i$ as n tends to infinity. We know that $\Sigma_n \sim \text{Binomial}(n, p_n)$ and therefore

$$\mathbb{P}(\Sigma_n = 0) = (1-p_n)^n = \left(1 - \frac{\alpha t}{n}\right)^n \to \mathrm{e}^{-\alpha t}, \qquad n \to \infty .$$

The last convergence is seen from

$$\left(1 - \frac{x}{n}\right)^n = \mathrm{e}^{n \log(1-\frac{x}{n})} = \mathrm{e}^{-x+O(n^{-1})} \to \mathrm{e}^{-x}, \qquad n \to \infty$$

for every $x > 0$, where we used $\log(1+y) = y + O(y^2)$ as $y \to 0$. Furthermore, we have

$$\frac{\mathbb{P}(\Sigma_n = k)}{\mathbb{P}(\Sigma_n = k-1)} = \frac{\binom{n}{k} p_n^k (1-p_n)^{n-k}}{\binom{n}{k-1} p_n^{k-1} (1-p_n)^{n-k+1}} = \frac{(n-k+1)p_n}{k(1-p_n)} \to \frac{\alpha t}{k}, \qquad n \to \infty .$$

Thus, we conclude $\mathbb{P}(\Sigma_n = k) = \frac{(\alpha t)^k}{k!} \mathrm{e}^{-\alpha t}$.

Definition 3.14 (Poisson distribution). For $\lambda > 0$ we call the distribution $\mathbb{P} = \text{Poisson}(\lambda)$ on non-negative integers defined through the PMF

$$\mathbb{P}(\{k\}) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}, \qquad k \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$$

the Poisson distribution.

3.3 Continuous random variables

Definition 3.15 (Continuous distribution). A probability distribution $\mathbb{P}$ on an interval $S \subset \mathbb{R}$ is called a continuous distributions if there is a function $f : \mathbb{R} \to \mathbb{R}$ such that

$$\mathbb{P}([a,b]) = \int_a^b f(x) \mathrm{d}x,$$

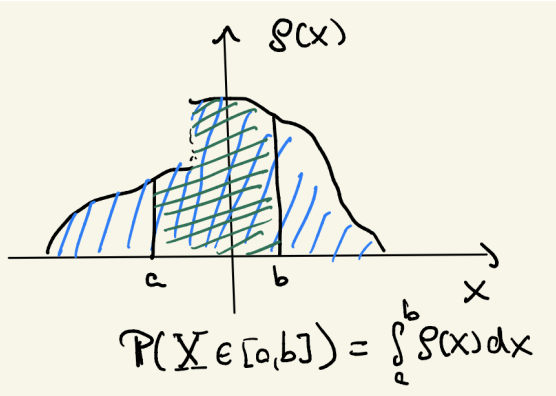
holds for all $a \le b$. In this case f is called probability density function (PDF).

The PDF has the following properties

- f has non-negative values, i.e. $f(x) \ge 0$.
- f is normalized by $\int_{-\infty}^{\infty} f(x) \mathrm{d}x = 1$.

Random variables X whose distributions $\mathbb{P}_X$ on $S = R_X$ are continuous distributions are called *continuous random variables*. We write f_X for the corresponding PDF. Put differently, a continuous random variable X satisfies

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



EXAMPLE 3.16 (Uniform distribution). Let $\mathbb{P}$ be the uniform distribution on the interval $[a, b]$. By definition

$$\mathbb{P}([s,t]) = \frac{t-s}{b-a} = \int_s^t \frac{1}{b-a}\mathrm{d}x$$

for all $a \le s \le t \le b$. Therefore $\mathbb{P}$ is a continuous distribution with constant PDF $f(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$.

EXAMPLE 3.17 (Waiting times). Let us go back to our example of counting rate events in Example 3.13. How long do we have to wait for the first request to arrive at the server? Let T_n be the time of the first request in our discretised model. Then

$$\mathbb{P}(T_n > t) = \mathbb{P}(\Sigma_n = 0) \to \mathrm{e}^{-\alpha t}\,, \qquad n \to \infty\,.$$

DEFINITION 3.18 (Exponential distribution). For $\alpha > 0$ the distribution Exponential(α) on $(0, \infty)$ with PDF

$$f(x) = \alpha\,\mathrm{e}^{-\alpha x}\mathbb{1}_{(0,\infty)}(x)\,,$$

is called exponential distribution.

Exponential random variables $T \sim$ Exponential(α) model waiting times, e.g. of arrivals of server requests, customers at a supermarket or decay events in radioactive materials.

DEFINITION 3.19 (Gaussian distribution). For $v > 0$ and $a \in \mathbb{R}$ the distribution $N(a, v)$ on $\mathbb{R}$ with PDF

$$f(x) = \frac{1}{\sqrt{2\pi v}}\,\mathrm{e}^{-\frac{(x-a)^2}{2v}} \tag{3.1}$$

is called Gaussian (or normal) distribution with mean a and variance v. The distribution $N(0, 1)$ is called standard normal distribution.

Let us now show that the f from (3.1) is a probability density, i.e. that $\int f(x)\mathrm{d}x = 1$. First we change coordinates to $y = (x-a)/\sqrt{2v}$ and find

$$\int f(x)\mathrm{d}x = \frac{1}{\sqrt{2\pi v}}\int_{-\infty}^{\infty}\mathrm{e}^{-\frac{(x-a)^2}{2v}}\mathrm{d}x = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}\mathrm{e}^{-y^2}\mathrm{d}y\,.$$

Then we multiply two copies of the same integral and rewrite

$$\left(\int_{-\infty}^{\infty}\mathrm{e}^{-y^2}\mathrm{d}y\right)^2 = \int\int \mathrm{e}^{-x^2-y^2}\mathrm{d}x\mathrm{d}y = 4\int_0^{\infty}\int_0^{\infty}\mathrm{e}^{-x^2-y^2}\mathrm{d}x\mathrm{d}y\,.$$

Now we perform the following substitution in the x-integral:

$$r := \sqrt{x^2+y^2}\,, \qquad \frac{\mathrm{d}r}{\mathrm{d}x} = \frac{x}{\sqrt{x^2+y^2}} = \frac{\sqrt{r^2-y^2}}{r}\,.$$

Thus, we find

$$\begin{aligned}
\int_0^{\infty}\int_0^{\infty}\mathrm{e}^{-x^2-y^2}\mathrm{d}x\mathrm{d}y &= \int_0^{\infty}\int_0^{\infty}\frac{r\,\mathrm{e}^{-r^2}}{\sqrt{r^2-y^2}}\mathbb{1}(r>y)\mathrm{d}r\mathrm{d}y \\
&= \int_0^{\infty} r\,\mathrm{e}^{-r^2}\int_0^{r}\frac{1}{\sqrt{r^2-y^2}}\mathrm{d}y\mathrm{d}r \\
&= \int_0^{\infty} r\,\mathrm{e}^{-r^2}\mathrm{d}r\int_0^{1}\frac{1}{\sqrt{1-y^2}}\mathrm{d}y\,,
\end{aligned}$$

where we scaled the y-variable for the last step. We are left with the task of computing two integrals, namely

$$\int_0^{\infty}\mathrm{e}^{-r^2}r\,\mathrm{d}r = -\frac{1}{2}\int_0^{\infty}\frac{\mathrm{d}}{\mathrm{d}r}\mathrm{e}^{-r^2}\,\mathrm{d}r = \frac{1}{2}\,,$$

and

$$\int_0^1 \frac{1}{\sqrt{1-y^2}}\mathrm{d}y = \int_0^{\pi/2} \frac{\cos\varphi}{\sqrt{1-(\sin\varphi)^2}}\mathrm{d}\varphi = \frac{\pi}{2}\,.$$

For the second integral we used the substitution $y = \sin\varphi$. Altogether, we showed

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

From the fact that (3.1) is the density of a probability distribution, i.e. from its normalisation, we conclude the following simple formula for Gaussian integrals:

$$\int \mathrm{e}^{-\frac{\alpha}{2}x^2+\beta x}\mathrm{d}x = \sqrt{\frac{2\pi}{\alpha}}\,\mathrm{e}^{\frac{\beta^2}{2\alpha}} \int \frac{\mathrm{e}^{-\frac{\alpha}{2}(x-\beta/\alpha)^2}}{\sqrt{2\pi\alpha}}\mathrm{d}x = \sqrt{\frac{2\pi}{\alpha}}\,\mathrm{e}^{\frac{\beta^2}{2\alpha}}\,. \tag{3.2}$$

Remark 3.20 (Universality of the Gaussian distribution). The Gaussian distribution is arguably the most important distribution in probability. Its ubiquitous appearance in nature is due to the fact that the fluctuation of large sums of independent random variables are asymptotically Gaussian under very mild assumptions. In subsequent course you will learn about the central limit theorem, asserting that quite generically $\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mathbb{E}X_i) \to N(0,v)$ as $n\to\infty$ in some sense if all X_i are outcomes of the same repeated independent experiment.

The so called chi-squared distribution plays an important role in statistics. A special case is the χ^2_1-distribution which is the distribution of X^2, where $X \sim N(0,1)$. Let us compute the PDF of X^2. Clearly f_{X^2} can have non-zero values only on non-negative numbers. To read off the PDF f_{X^2} we have to write $\mathbb{P}[X^2 \in [a,b]]$ for $0 \le a < b$ as an integral over the interval $[a,b]$. This is done using the substitution rule for integrals, i.e. via

$$\mathbb{P}[X^2 \in [a,b]] = \mathbb{P}[X \in [\sqrt{a},\sqrt{b}]] + \mathbb{P}[X \in [-\sqrt{b},-\sqrt{a}]] = \frac{2}{\sqrt{2\pi}}\int_{\sqrt{a}}^{\sqrt{b}} \mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_a^b \frac{\mathrm{e}^{-\frac{x}{2}}}{\sqrt{x}}\mathrm{d}x\,.$$

We conclude that the PDF for the χ^2_1-distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}}\frac{\mathrm{e}^{-\frac{x}{2}}}{\sqrt{x}}\,\mathbb{1}_{(0,\infty)}(x)\,.$$

More generally, we can compute the PDF for any function $g(X)$ of a continuous random variable X in terms of the PDF of X itself.

Lemma 3.21 (Transformation of density). Let X be a continuous random variable with density f_X and $g:\mathbb{R}\to\mathbb{R}$ continuously differentiable with either $g'(x) > 0$ or $g'(x) < 0$ for all x. Then $g(X)$ has density

$$f_{g(X)}(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}\,.$$

Proof. Let $a < b$. The we use the substitution rule for integration to compute

$$\mathbb{P}(g(X)\in[a,b]) = \mathbb{P}(X \in g^{-1}[[a,b]]) = \int f_X(x)\mathbb{1}_{g^{-1}[[a,b]]}(x)\mathrm{d}x = \int_a^b f_X(g^{-1}(y))\frac{1}{|g'(g^{-1}(y))|}\mathrm{d}y\,.$$

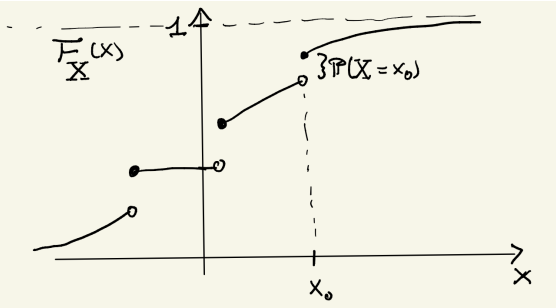
For this we read off the density of $g(X)$. The absolute value is only needed if g is monotonically decreasing. □

3.4 Cumulative distribution function

DEFINITION 3.22 (Cumulative distribution function). The cumulative distribution function (CDF) of a real random variable X is defined as $F_X(x) := \mathbb{P}(X \leq x)$.

PROPOSITION 3.23 (Properties of CDF). The CDF $F = F_X$ of a real random variable X has the following properties:

1. Monotonicity: $F(x) \leq F(y)$ for $x \leq y$
2. Right continuity: $F(x) = \lim_{y \downarrow x} F(y)$
3. Limits at infinity: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$



The following proof is for the interested reader.

Proof. * To show monotonicity we see that

$$F(y) = \mathbb{P}(X \leq y) = \mathbb{P}(X \leq x) + \mathbb{P}(x < X \leq y) \geq \mathbb{P}(X \leq x) = F(x)\,.$$

To show right continuity we pick a monotonically decreasing sequence ε_n of positive numbers such that $\lim_{n \to \infty} \varepsilon_n = 0$. Then

$$\begin{aligned}
\mathbb{P}(X \leq x + \varepsilon_n) &= \mathbb{P}\big(\cap_{k=1}^n \{X \leq x + \varepsilon_k\}\big) \\
&= 1 - \mathbb{P}\big(\cup_{k=1}^n \{X > x + \varepsilon_k\}\big) \\
&= 1 - \mathbb{P}\big(\cup_{k=2}^n \{\varepsilon_k < X - x \leq \varepsilon_{k-1}\} \cup \{X > x + \varepsilon_1\}\big) \\
&= 1 - \sum_{k=2}^{n} \mathbb{P}\big(\varepsilon_k < X - x \leq \varepsilon_{k-1}\big) + \mathbb{P}\big(X > x + \varepsilon_1\big)\,.
\end{aligned}$$

We can take the limit on he right hand side by (1.1) and get

$$\begin{aligned}
\lim_{n \to \infty} \mathbb{P}(X \leq x + \varepsilon_n) &= 1 - \sum_{k=2}^{\infty} \mathbb{P}\big(\varepsilon_k < X - x \leq \varepsilon_{k-1}\big) + \mathbb{P}\big(X > x + \varepsilon_1\big) \\
&= 1 - \mathbb{P}\big(\cup_{k=2}^{\infty} \{\varepsilon_k < X - x \leq \varepsilon_{k-1}\} \cup \{X > x + \varepsilon_1\}\big) \\
&= 1 - \mathbb{P}\big(X > x\big) = \mathbb{P}\big(X \leq x\big)\,.
\end{aligned}$$

□

EXAMPLE 3.24 (CDF for Bernoulli distribution). We compute the CDF for a Bernoulli(p) random variable X. For $x < 0$ we have $F_X(x) = \mathbb{P}(X \leq x) = 0$ since X only takes the values 0 and 1. For $x \in [0, 1)$ we get

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X = 0) = 1 - p$$

and for $x \geq 1$ we have

$$F_X(x) = \mathbb{P}(X \in \{0, 1\}) = 1\,.$$

Altogether, the CDF of X is

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0\,, \\ 1 - p & \text{for } 0 \leq x < 1\,, \\ 1 & \text{for } x \geq 1\,. \end{cases}$$

A discrete random variable X has a piecewise constant CDF F_X. At a jump point x the hight of the jump corresponds to the probability of X having the value X, i.e. $\mathbb{P}(X = x) = F_X(x) - \lim_{\varepsilon \downarrow 0} F_X(x-\varepsilon)$.

EXAMPLE 3.25 (CDF for Exponential distribution). We compute the CDF for a $\mathrm{Exp}(\lambda)$ random variable X. For $x < 0$ we have $F_X(x) = \mathbb{P}(X \leq x) = 0$ since X only takes positive values. For $x \geq 0$ we get

$$\mathbb{P}(X \leq x) = \int_0^x f_X(y)\mathrm{d}y = \lambda \int_0^x \mathrm{e}^{-\lambda y}\mathrm{d}y = 1 - \mathrm{e}^{-\lambda x}\,.$$

Altogether, we have

$$F_X(x) = (1 - \mathrm{e}^{-\lambda x})\mathbb{1}_{[0,\infty)}(x)\,.$$

Note that $F_X'(x) = \lambda \mathrm{e}^{-\lambda x} = f_X(x)$ for $x > 0$.

For continuous random variables X with differentiable CDF the following relation between PDF and CDF holds:

$$f_X(x) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \int_x^{x+\varepsilon} \rho(y)\mathrm{d}y = \lim_{\varepsilon \downarrow 0} \frac{\mathbb{P}([x, x+\varepsilon])}{\varepsilon} = \lim_{\varepsilon \downarrow 0} \frac{F_X(x+\varepsilon) - F_X(x)}{\varepsilon} = F_X'(x)\,.$$

LEMMA 3.26 (Classification of CDF). Let $F : \mathbb{R} \to [0,1]$ be a function that satisfies Properties 1.,2. and 3. from Proposition 3.23. Then there is a random variable X on the probability space $S = [0,1]$ with the uniform distribution such that $F = F_X$ is its CDF.

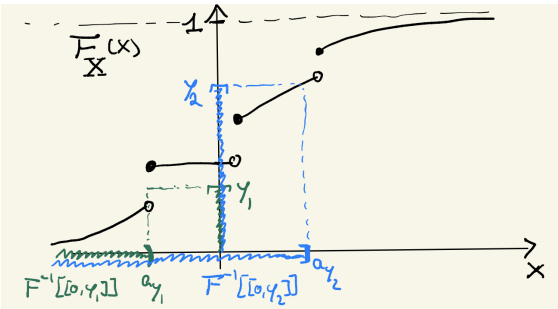
This lemma shows that there can also be mixed distributions that are neither completely discrete nor completely continuous. The corresponding CDF may have jumps without being piecewise constant.

The following proof is for the interested reader.

Proof. * Since F is monotonously increasing, the preimage $F^{-1}[[0,y]]$ is an interval for all $y \in [0,1]$. Since F is right continuous and by Property 3 we even have for all $y \in (0,1)$ that

$$F^{-1}[[0,y]] = \{x \in \mathbb{R} : F(x) \leq y\} = \begin{cases} (-\infty, a_y) & \text{if } y \notin F[\mathbb{R}] \\ (-\infty, a_y] & \text{if } y \in F[\mathbb{R}] \end{cases},$$

for some $a_y \in \mathbb{R}$. **Check that this is the case!**



We set $X(y) := a_y$. Now we prove that $F_X = F$. Indeed,

$$F_X(x) = \mathbb{P}(X \leq x) = |\{y \in (0,1) : a_y \leq x\}| = |\{y \in (0,1) : y \leq F(x)\}| = F(x)\,.$$

Here, the second to last identity follows from the inclusions

$$\{y : y \leq F(x)\} \subset \{y : a_y \leq x\} \subset \{y : y \leq F(x)\}\,,$$

where the first is clear from the definition of a_y and the second holds because $y \leq F(a_y)$. □

4 Expectation and variance

Here we discuss the content of **Sections 3.2.3, 3.2.4 and 4.1.2** from "Introduction to probability, statistics, and random processes".

4.1 Expectation

DEFINITION 4.1 (Expectation). For a discrete random variable X with PMF $P_X : \Sigma \to [0,1]$ we define its expectation through

$$\mathbb{E}X = \sum_{x \in R_X} x\, P_X(x)\,.$$

For a continuous random variable X with PDF f_X we define its expectation through

$$\mathbb{E}X = \int x\, f_X(x) \mathbb{1}_{R_X}(x) \mathrm{d}x = \int_{R_X} x\, f_X(x)(x) \mathrm{d}x\,.$$

For a function $g : R_X \to \mathbb{R}$ and a random variable X with PMF P_X (or PDF f_X) we have

$$\mathbb{E}g(X) = \sum_{x \in R_X} g(x) P_X(x) \qquad \Big(\text{ or } \quad \mathbb{E}g(X) = \int_{R_X} g(x) f_X(x) \mathrm{d}x \Big)\,. \tag{4.1}$$

Indeed, in case of a discrete random variable X and $g : R_X \to \mathbb{R}$ we can compute

$$\begin{aligned}
\mathbb{E}g(X) &= \sum_y y\, \mathbb{P}(g(X) = y) \\
&= \sum_y y\, \mathbb{P}(X \in g^{-1}[\{y\}]) \\
&= \sum_y g(x) \sum_{x : g(x) = y} \mathbb{P}(X = x) \\
&= \sum_x g(x)\, \mathbb{P}(X = x)\,.
\end{aligned}$$

The calculation for continuous random variables is more involved but similar. The formula (4.1) is often called law of the unconscious statistician (LOTUS).

In particular, we get the following relation between probabilities and expectation

$$\mathbb{E}\mathbb{1}_{X \in A} = \mathbb{E}\mathbb{1}_A(X) = \mathbb{P}(X \in A)\,. \tag{4.2}$$

For discrete random variables this follows from

$$\mathbb{E}\,\mathbb{1}_A(X) = \sum_x \mathbb{1}_A(x) \mathbb{P}(X = x) = \sum_{x \in A} \mathbb{P}(X = x)\,.$$

EXAMPLE 4.2 (Binomial random variable). Let $X \sim \text{Binomial}(n, p)$. Then

$$\begin{aligned}
\mathbb{E}\,X &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-1-k)!} p^k (1-p)^{n-1-k} \\
&= np\,.
\end{aligned}$$

EXAMPLE 4.3 (LOTUS for continuous uniform distribution). Let X be uniformly distributed on $[0, \pi]$. Then by LOTUS we have

$$\mathbb{E}\,\sin(X) = \frac{1}{\pi} \int_0^\pi \sin(x) \mathrm{d}x = -\frac{1}{\pi} \cos(x)\Big|_0^\pi = \frac{2}{\pi}\,.$$

The following is a very helpful property of expectations. We state it without proof.

PROPOSITION 4.4 (Linearity of expectation). The expectation is linear, i.e. for random variables X, Y and a real number α we get

$$\mathbb{E}\alpha X = \alpha \mathbb{E}X\,, \qquad \mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y\,.$$

4.2 Variance and higher order moments

DEFINITION 4.5 (Variance). For a real random variable X we define its variance through

$$\operatorname{Var} X = \mathbb{E}(X - \mathbb{E}X)^2 .$$

The quantity $\sqrt{\operatorname{Var} X}$ is called the deviation of X.

The deviation of X is a rough measure of the typical size of X.

LEMMA 4.6 (Properties of variance). The variance of a real random variable X can be computed via

$$\operatorname{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2 .$$

For any $\alpha, \mu \in \mathbb{R}$ the variance satisfies the scaling law $\operatorname{Var}(\alpha X + \mu) = \alpha^2 \operatorname{Var} X$.

Proof. The formula for the variance follows from multiplying out the square and using linearity, i.e. by

$$\operatorname{Var} X = \mathbb{E}(X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2) = \mathbb{E}X^2 - (\mathbb{E}X)^2 .$$

For the scaling rule we realise that $X - \mathbb{E}X$ stays the same if we add a constant μ and scales linearly in the multiplying constant $\alpha > 0$. □

EXAMPLE 4.7 (Uniform random variable). Let X be uniformly distributed on $[0, 1]$. To determine the variance of X we compute

$$\mathbb{E}X = \int_0^1 x \, \mathrm{d}x = \frac{1}{2} , \qquad \mathbb{E}X^2 = \int_0^1 x^2 \, \mathrm{d}x = \frac{1}{3} .$$

Thus, $\operatorname{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = 1/12$.

EXAMPLE 4.8 (Gaussian distribution). We show that $\mathbb{E}X = a$ and $\operatorname{Var} X = v$ for a random variable with $X \sim N(a, v)$. Indeed,

$$\mathbb{E}X = \int x f_X(x) \mathrm{d}x = a \int f_X(x) \mathrm{d}x + \int \frac{x - a}{\sqrt{2\pi v}} \, \mathrm{e}^{-\frac{(x-a)^2}{2v}} \mathrm{d}x = a + \int \frac{x}{\sqrt{2\pi v}} \, \mathrm{e}^{-\frac{x^2}{2v}} \mathrm{d}x ,$$

where we used the normalisation of f_X and substituted $x \to x + a$. The second summand on the right hand side vanishes because the integrand is an odd function of x. For the variance we find

$$\operatorname{Var} X = \int (x - a)^2 f_X(x) \mathrm{d}x = v \int \frac{x^2}{\sqrt{2\pi}} \, \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x ,$$

where we substituted $x \to \sqrt{v}(x + a)$. To determine the remaining integral we introduce an auxiliary parameter α and compute

$$\int \frac{x^2}{\sqrt{2\pi}} \, \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x = -\sqrt{\frac{2}{\pi}} \frac{\mathrm{d}}{\mathrm{d}\alpha} \int \mathrm{e}^{-\alpha \frac{x^2}{2}} \mathrm{d}x \bigg|_{\alpha=1} = -\frac{\mathrm{d}}{\mathrm{d}\alpha} \frac{2}{\sqrt{\alpha}} \bigg|_{\alpha=1} = 1 ,$$

where we used (3.2) in the second identity.

Up to now we only considered $\mathbb{E}X$ and $\mathbb{E}X^2$ for random variables X. But sometimes we want to know the expectation of higher order monomials. These higher order moments are particularly useful when controlling large deviations of random variables since these are given more weight.

DEFINITION 4.9 (Moments). The k-th moment of a random variable is defined to be $\mathbb{E}X^k$.

EXAMPLE 4.10 (Moments of exponential distribution). Let X be a random variable with $X \sim \mathrm{Exp}(\lambda)$. Then its expectation is

$$\mathbb{E}X = \lambda \int_0^\infty x\,\mathrm{e}^{-\lambda x}\mathrm{d}x = -\lambda \frac{\mathrm{d}}{\mathrm{d}\lambda} \int_0^\infty \mathrm{e}^{-\lambda x}\mathrm{d}x = \lambda \frac{\mathrm{d}}{\mathrm{d}\lambda} \frac{\mathrm{e}^{-\lambda x}}{\lambda}\bigg|_0^\infty = -\lambda \frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{1}{\lambda} = \frac{1}{\lambda}\,.$$

The second moment is

$$\mathbb{E}X^2 = \lambda \int_0^\infty x^2\,\mathrm{e}^{-\lambda x}\mathrm{d}x = \lambda \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \int_0^\infty \mathrm{e}^{-\lambda x}\mathrm{d}x = \lambda \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\frac{1}{\lambda} = -\lambda \frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{1}{\lambda^2} = \frac{2}{\lambda^2}\,.$$

More generally, we can compute the k-th moment

$$\mathbb{E}X^k = \lambda \int_0^\infty x^k\,\mathrm{e}^{-\lambda x}\mathrm{d}x = (-1)^k\lambda \frac{\mathrm{d}^k}{\mathrm{d}\lambda^k} \int_0^\infty \mathrm{e}^{-\lambda x}\mathrm{d}x = (-1)^k\lambda \frac{\mathrm{d}^k}{\mathrm{d}\lambda^k}\frac{1}{\lambda} = \frac{k!}{\lambda^k}\,.$$

5 Joint distributions and independence

5.1 Joint distributions

DEFINITION 5.1 (Discrete joint distribution). Let X, Y be two discrete random variables on the same probability space $(S, \mathbb{P})$. The function $P_{X,Y} : R_X \times R_Y \to [0,1]$ defined through

$$P_{X,Y}(x,y) := \mathbb{P}(X = x, Y = y)$$

is called joint probability mass function (joint PMF). The assignment

$$\mathbb{P}_{X,Y}(A) := \sum_{(x,y)\in A} P_{X,Y}(x,y)$$

for $A \subset R_X \times R_Y$ is a probability distribution on $R_X \times R_Y$ and is called the joint distribution of X, Y.

Joint distributions provide information about the dependence of random variables.

EXAMPLE 5.2 (Two coin tosses). Let $S = \{0,1\}^2$ with the uniform distribution be the probability space associated with two coin tosses. Let $X : S \to \{0,1\}, X(s_1,s_2) := s_1$ and $Y : S \to \{0,1,2\}, X(s_1,s_2) := s_1 + s_2$ be the result of the first toss and the number of heads, respectively. We compute the joint PMF for X, Y as

$$\mathbb{P}(X=0,Y=0) = \frac{1}{4}\,, \quad \mathbb{P}(X=0,Y=1) = \frac{1}{4}\,, \quad \mathbb{P}(X=0,Y=2) = 0\,,$$
$$\mathbb{P}(X=1,Y=0) = 0\,, \quad \mathbb{P}(X=1,Y=1) = \frac{1}{4}\,, \quad \mathbb{P}(X=1,Y=2) = \frac{1}{4}\,.$$

This can be summarised in the following table:

	$Y=0$	$Y=1$	$Y=2$
$X=0$	1/4	1/4	0
$X=1$	0	1/4	1/4

Summing up the row and column values gives the marginal distribution of X and Y, respectively. Indeed, $\mathbb{P}(X=0) = \mathbb{P}(X=1) = \frac{1}{2}$, as well as $\mathbb{P}(Y=0) = \mathbb{P}(Y=2) = 1/4$ and $\mathbb{P}(Y=1) = 1/2$. Thus, $X \sim \mathrm{Bernoulli}(1/2)$ and $Y \sim \mathrm{Binomial}(1/2, 2)$, as expected.

DEFINITION 5.3 (Continuous joint distribution). Let X, Y be two random variables. They are called jointly continuous if there is a function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ such that

$$\mathbb{P}(X \in [a,b], Y \in [c,d]) = \int_a^b \int_c^d f_{X,Y}(x,y)\mathrm{d}y\mathrm{d}x$$

holds for any $a < b$ and $c < d$. In this case $f_{X,Y}$ is called the joint probability density function (joint PDF) of X and Y.

EXAMPLE 5.4 (Multivariate Gaussian distribution). Two random variables are called jointly Gaussian if they are jointly continuous with joint PDF

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{\det A}} \exp\left(-\frac{1}{2}(x-m_1, y-m_2) A^{-1} \begin{pmatrix} x-m_1 \\ y-m_2 \end{pmatrix} \right), \tag{5.1}$$

where $m = (m_1, m_2) \in \mathbb{R}^2$, $\det A = a_{11}a_{22} - a_{12}a_{21}$ and

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

is a positive definite matrix (A matrix A is positive definite if it is symmetric, $A = A^T$ and has positive eigenvalues). In particular, $a_{12} = a_{21}$. In this case we write $(X,Y) \sim N(m, A)$ and we say that m is the mean and A the covariance matrix of the vector (X,Y).

The joint PDF of X and Y is used to compute expectations of observables that depend on both random variables. This is done through the formula

$$\mathbb{E}\, g(X,Y) = \int\int g(x,y) f_{X,Y}(x,y) \mathrm{d}x\mathrm{d}y\,,$$

for some function $g : \mathbb{R}^2 \to \mathbb{R}$.

In particular, the first marginal distribution (distribution of X) is determined by integrating the joint PDF over the variable corresponding to Y, i.e.

$$\mathbb{P}(X \in [a,b]) = \mathbb{E}\,\mathbb{1}_{[a,b]}(X) = \int\int \mathbb{1}_{[a,b]}(x) f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = \int_a^b f_X(x)\mathrm{d}x\,,$$

where

$$f_X(x) = \int f_{X,Y}(x,y)\mathrm{d}y\,.$$

EXAMPLE 5.5. Let X, Y be jointly Gaussian with $(X,Y) \sim N(0,A)$ and

$$A = \frac{1}{3}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

More explicitly X and Y have joint PDF

$$f_{X,Y}(x,y) = \frac{\sqrt{3}}{2\pi} \mathrm{e}^{-x^2-y^2+xy}\,.$$

We compute $\mathbb{E}g(X,Y)$ for $g(x,y) = xy$ via

$$\begin{aligned}
\mathbb{E}XY &= \frac{\sqrt{3}}{2\pi}\int\int xy\,\mathrm{e}^{-x^2-y^2+xy}\mathrm{d}x\mathrm{d}y \\
&\overset{(1)}{=} \frac{\sqrt{3}}{2\pi}\int\int \frac{\mathrm{d}}{\mathrm{d}a}\mathrm{e}^{-x^2-y^2+axy}\mathrm{d}x\mathrm{d}y\Big|_{a=1} \\
&= \frac{\sqrt{3}}{2\pi}\frac{\mathrm{d}}{\mathrm{d}a}\int \mathrm{e}^{-y^2}\int \mathrm{e}^{-x^2+axy}\mathrm{d}x\mathrm{d}y\Big|_{a=1} \\
&\overset{(2)}{=} \frac{\sqrt{3}}{2\pi}\frac{\mathrm{d}}{\mathrm{d}a}\int \mathrm{e}^{-y^2}\sqrt{\pi}\,\mathrm{e}^{\frac{1}{4}a^2y^2}\mathrm{d}y\Big|_{a=1} \\
&= \frac{\sqrt{3}}{2\sqrt{\pi}}\frac{\mathrm{d}}{\mathrm{d}a}\int \mathrm{e}^{-\frac{1}{2}(2-\frac{a^2}{2})y^2}\mathrm{d}y\Big|_{a=1} \\
&\overset{(2)}{=} \frac{\sqrt{3}}{2\sqrt{\pi}}\frac{\mathrm{d}}{\mathrm{d}a}\sqrt{\frac{2\pi}{2-\frac{a^2}{2}}}\Big|_{a=1} \\
&= \sqrt{3}\frac{\mathrm{d}}{\mathrm{d}a}\sqrt{\frac{1}{4-a^2}}\Big|_{a=1} \\
&= \frac{\sqrt{3}}{2}\frac{2a}{(4-a^2)^{3/2}}\Big|_{a=1} = \frac{1}{3}\,.
\end{aligned}$$

Here, we introduced the auxiliary parameter a in (1) to simplify performing the integral and used (3.2) in (2).

5.2 Independence

DEFINITION 5.6 (Independence of two random variables). Two random variables X, Y are called independent if the events $\{X \in A\}$ and $\{Y \in B\}$ are independent for all possible A, B, i.e. if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)\,.$$

THEOREM 5.7 (Product rule). For two independent random variables the following product rule holds.

1. Two discrete random variables X, Y are independent if and only if their joint PMF factorises, i.e. if

$$P_{X,Y}(x, y) = P_X(x)P_Y(y)\,, \qquad x \in R_X\,,\ y \in R_Y\,. \tag{5.2}$$

2. Two jointly continuous random variables X, Y are independent if and only if their joint PDF factorises, i.e. if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)\,, \qquad x, y \in \mathbb{R}\,.$$

Proof. We perform the proof in the discrete case. The continuous case is analogous. We first show that (5.2) implies independence. Indeed,

$$\mathbb{P}(X \in A, Y \in B) = \sum_{x \in R_X, y \in R_y} P_{X,Y}(x, y) = \sum_{x \in R_X, y \in R_y} P_X(x)P_Y(y) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)\,,$$

and, thus, X and Y are independent. On the other hand independence of X and Y implies

$$P_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) = P_X(x)P_Y(y)\,,$$

i.e. the joint PMF factorises. □

DEFINITION 5.8 (i.i.d.). We use the abbreviation i.i.d. for independent and identically distributed random variables, i.e. we say that X, Y are i.i.d. if they are independent and have all the same distribution $\mathbb{P}_X = \mathbb{P}_Y$.

EXAMPLE 5.9 (Rolling a dice several times). Rolling a dice twice is modelled by two i.i.d. random variables X_1, X_2 with uniform distribution on $\{1, 2, 3, 4, 5, 6\}$.

Now we introduce an important operation on two functions, namely the convolution. For two non-negative functions $f, g : \mathbb{R} \to \mathbb{R}$ we define the convolution $f \star g : \mathbb{R} \to \mathbb{R}$ via

$$(f \star g)(x) := \int_{\mathbb{R}} f(y)g(x - y)\mathrm{d}y\,.$$

Note that $f \star g = g \star f$.

LEMMA 5.10 (Convolution rule). Let X and Y be independent continuous random variables with PDFs f_X and f_Y, respectively. Then $X + Y$ has PDF

$$f_{X+Y}(z) = \int f_X(x)f_Y(z - x)\mathrm{d}x = f_X \star f_Y(z)\,.$$

Proof. We determine the PDF of $X + Y$ by computing the probability

$$\mathbb{P}(X + Y \in I) = \int\int f_X(x)f_Y(y)\mathbb{1}(x + y \in I)\mathrm{d}x\mathrm{d}y = \int\int f_X(x - y)f_Y(y)\mathbb{1}(x \in I)\mathrm{d}y\mathrm{d}x\,,$$

by shifting the variables $x \to x - y$. □

REMARK 5.11. For discrete random variables with values in $\mathbb{Z}$ a similar rule holds.

EXAMPLE 5.12. Let X, Y be i.i.d. Exponential(α)-distributed random variables. Then the PDF of $X + Y$ is

$$\begin{aligned} f_{X+Y}(z) &= \alpha^2 \int \mathrm{e}^{-\alpha x} \mathbb{1}_{(0,\infty)}(x) \mathrm{e}^{-\alpha(z-x)} \mathbb{1}_{(0,\infty)}(z-x) \mathrm{d}x \\ &= \alpha^2 \mathrm{e}^{-\alpha z} \int_0^z \mathrm{d}x = \alpha^2 z \, \mathrm{e}^{-\alpha z}. \end{aligned}$$

DEFINITION 5.13 (Covariance). For two random variables X and Y we define their covariance as

$$\mathrm{Cov(X, Y)} := \mathbb{E}((\mathrm{X} - \mathbb{E}\mathrm{X})(\mathrm{Y} - \mathbb{E}\mathrm{Y})) = \mathbb{E}(\mathrm{XY}) - (\mathbb{E}\mathrm{X})(\mathbb{E}\mathrm{Y}) \, .$$

If $\operatorname{Var} X, \operatorname{Var} Y > 0$, then we define the correlation of X and Y as

$$\mathrm{Corr(X, Y)} := \frac{\mathrm{Cov(X, Y)}}{\sqrt{\mathrm{Var\,X}\,\mathrm{Var\,Y}}} \, .$$

In case $\mathrm{Cov(X, Y)} = 0$ we say that X and Y are uncorrelated.

PROPOSITION 5.14 (Property of covariance). For random variables X, Y, Z and numbers $\alpha, \mu \in \mathbb{R}$ we have the following:

1. Symmetry: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
2. Shift invariance: $\mathrm{Cov}(X + \mu, Y) = \mathrm{Cov}(X, Y)$
3. Bilinearity:

$$\begin{aligned} \mathrm{Cov}(\alpha X + Z, Y) &= \alpha \mathrm{Cov}(X, Y) + \mathrm{Cov}(Z, Y) \\ \mathrm{Cov}(Y, \alpha X + Z) &= \alpha \mathrm{Cov}(Y, X) + \mathrm{Cov}(Y, Z) \end{aligned}$$

4. Relation between variance and covariance:

$$\mathrm{Cov}(X, X) = \operatorname{Var} X \, .$$

5. If X, Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Proof. All properties 1 to 4 follow immediately from the definition of Cov and the linearity of $\mathbb{E}$. Property 5 follows from the product rule Theorem 5.7 because

$$\mathrm{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = (\mathbb{E}(X - \mathbb{E}X))(\mathbb{E}(Y - \mathbb{E}Y)) = 0 \, ,$$

when X, Y are independent. □

COROLLARY 5.15. If X and Y are independent random variables, then $\operatorname{Var}(X + Y) = \operatorname{Var} X + \operatorname{Var} Y$.

Proof. We compute

$$\operatorname{Var}(X + Y) = \mathrm{Cov}(X + Y, X + Y) = \operatorname{Var} X + \operatorname{Var} Y + 2\,\mathrm{Cov}(X, Y) = \operatorname{Var} X + \operatorname{Var} Y \, .$$

□

6 Statistics

6.1 Testing

In statistical testing we check the evidence against the *null hypothesis* $\mathscr{H}_0$ (unproblematic, normal or expected situation). To formulate the null hypothesis and its alternative we need a model for the randomness underlying our data.

Definition 6.1 (Statistical model). A statistical model is a family $(\mathbb{P}_\theta)_{\theta\in\Theta}$ of probability distributions on some fixed probability space Σ (from which our samples are drawn).

Remark 6.2 (Independent observations). Here we will only discuss statistics of n independent observations $x_1,\dots,x_n$ of the same experiment. Then the statistical model encodes the distribution $\mathbb{P}_\theta$ of one single experiment.

The parameter space Θ of our statistical model $\mathbb{P}_\theta$ is split into two disjoint sets $\Theta = \Theta_0 \cup \Theta_a$. Our (acceptable) null hypothesis $\mathscr{H}_0$ is that $\theta \in \Theta_0$, while our problematic alternative $\mathscr{H}_a$ is $\theta \in \Theta_a$. Given a data sample $x_1,\dots,x_n \in \Sigma$ we will either reject $\mathscr{H}_0$ or not.

Decision	$\mathscr{H}_0$ true	$\mathscr{H}_0$ false
$\mathscr{H}_0$ not rejected	correct	type II error (false negative)
$\mathscr{H}_0$ rejected	type I error (false positive)	correct

We want to avoid type I errors and therefore we fix a small significance level (or probability) $\alpha \in (0,1)$ of making such mistake, usually something like $\alpha = 0.05$. Then we have to fix a decision rule that depending on the observed data $x_1,\dots,x_n$ either rejects $\mathscr{H}_0$ or not and that is compatible with our choice of significance level α.

Example 6.3 (Gaussian model). We consider the statistical model $(N(\theta,v))_{\theta\in\mathbb{R}}$ and test the null hypothesis $\Theta_0 = \{\theta_0\}$ against the alternative $\Theta_a = \mathbb{R}\setminus\{\theta_0\}$. Given a sample $X_1,\dots,X_n$ we design a test as follows

$$\text{Reject } \mathscr{H}_0 \text{ if } |\overline{X}-\theta_0| > \varepsilon\,, \qquad \text{Do not reject } \mathscr{H}_0 \text{ if } |\overline{X}-\theta_0| \le \varepsilon\,,$$

where $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$. What choices of ε are compatible with the choice of significance level $\alpha = 0.05$? To find out we compute the probability of a type I error

$$\mathbb{P}_{\theta_0}(|\overline{X}-\theta_0|>\varepsilon) = \frac{1}{\sqrt{2\pi v}}\int_{-\varepsilon}^{\varepsilon} \mathrm{e}^{-\frac{x^2}{2v}}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{-\frac{\varepsilon}{\sqrt{v}}}^{\frac{\varepsilon}{\sqrt{v}}} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x\,.$$

In this computation we used that $\overline{X} \sim N(\theta_0, v)$ under the null hypothesis $\mathscr{H}_0$. This stems from the fact that for $Y_1,\dots,Y_n$ independent Gaussian random variables with $Y_i \sim N(m_i,v_i)$ we have $\sum_{i=1}^n Y_i \sim N(\sum_{i=1}^n m_i, \sum_{i=1}^n v_i)$. Thus, any choice $\varepsilon < c_\alpha\sqrt{v}$ with c_α defined by the identity

$$\frac{1}{\sqrt{2\pi}}\int_{-c_\alpha}^{c_\alpha} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x = \alpha$$

is compatible with significance level $\alpha = 0.05$. Of course be optimal choice is $\varepsilon = c_\alpha\sqrt{v}$, because that gives the highest chance of not having to reject $\mathscr{H}_0$.

6.2 Estimators

Here we consider a statistical model $(\mathbb{P}_\theta)_{\theta\in\Theta}$ and want to get a good estimate for the true value of θ from our data $x_1,\dots,x_n$. A simple principle is the following:

Definition 6.4 (Maximum-likelihood estimator). If $\mathbb{P}_\theta$ is continuous and has pdf ρ_θ (discrete and has pmf p_θ), then

$$T(x) := \operatorname{argmax}_{\theta\in\Theta}\prod_{i=1}^n \rho_\theta(x_i) \qquad \left(T(x) := \operatorname{argmax}_{\theta\in\Theta}\prod_{i=1}^n p_\theta(x_i)\right)$$

is called the maximum-likelihood estimator (MLE) for θ, provided a unique maximiser exists.

EXAMPLE 6.5 (MLE for Bernoulli). Suppose we observe a repeated Bernoulli experiment $x_1, \dots, x_n \in \{0, 1\}$ and want to estimate its parameter, i.e. our statistical model is $(\mathrm{Bern}(\theta))_{\theta\in[0,1]}$. Then we have to solve the following maximisation problem

$$T(x) = \mathrm{argmax}_{\theta\in[0,1]}\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}.$$

Differentiating this function and denoting $k = \sum_i x_i$ we find $k(1-\theta) - (n-k)\theta = 0$ as a condition for the maximiser; or equivalently $\theta = k/n$. Thus, the maximum-likelihood estimator is $T(x) = \frac{1}{n}\sum_{i=1}^n x_i$.

EXAMPLE 6.6 (MLE for Gaussian). Suppose we observe repeated Gaussian numbers $x_1, \dots, x_n \in \mathbb{R}$ in the statistical model is $(N(\theta, v))_{\theta\in\mathbb{R}}$. Then for the MLE we have to solve the following maximisation problem

$$T(x) = \mathrm{argmax}_{\theta\in\mathbb{R}}\frac{\mathrm{e}^{-\frac{1}{2v}\sum_i(x_i-\theta)^2}}{(2\pi v)^{n/2}}$$

The maximum is just the maximum of the exponent, i.e. the minimum of $h(\theta) = \frac{1}{2}\sum_i(x_i-\theta)^2$. To determine the minimum we differentiate and find

$$h'(\theta) = \sum_i(\theta - x_i) = n\theta - \sum_i x_i\,.$$

Therefore, $T(x) = \frac{1}{n}\sum_{i=1}^n x_i$ is again the MLE.