Lecture notes

MA 2631
Probability

Stephan Sturm

Version (rough draft)
October 14, 2014

Contents

List of Figures

List of Tables

Chapter 1

Combinatorial Analysis

1.1 Counting

1.1 Example. A communication system consists of four antennas, two of which are defective. The whole system is working unless two neighboring antennas are defective. What is the probability that the whole system works?

There is no difference between the antennas, except that some are working and some not...

Solution. We line up all the possible configurations. Let 1 denote therefore a working antenna and 0 a defective one. The result is presented in Table 1.1. In three of six cases the system is working. Thus the probability that the system is working is

$$\frac{3}{6} = \frac{1}{2} = 50\% = 0.5.$$

□

1	1	0	0
1	0	1	0
1	0	0	1
0	1	1	0
0	1	0	1
0	0	1	1

Table 1.1: List of all configurations. Defective configurations are in red.

Here the counting was easy, but if you would try to calculate the probability that a system with twenty antennas and seven of them defective is working, this would take a lot of time. Thus we will have to develop cleverer

ways of counting. The subfield of mathematics concerned with methods of clever counting is called *combinatorics*.

As a first step, we will have to learn how to combine different experiments:

1.2 Example. Assume you have two dice. What are all the possible outcomes when rolling the dice?

Solution. Let us list systematically all the pairs which can appear. E.g., $(3,4)$ will stand here for the outcome that the first die shows a 3 and the second one a 4. Then all the outcomes are listed in Table 1.2. Thus we have in total $6 \cdot 6 = 36$ possible outcomes. □

Wait, are $(3,4)$ and $(4,3)$ not the same thing? Actually not, as long as we can discern the two dice...

$$\underbrace{\left.\begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array}\right\} \mathbf{6}}_{\mathbf{6}}$$

Table 1.2: List of all possible outcomes when rolling two dice.

More general we have the following:

> **Principle of Counting**: If we have two experiments, the first with n possible outcomes and the second with m possible outcomes, then there are all together $n \cdot m$ possible outcomes.

One can generalize this even further.

> **Generalized Principle of Counting**: If we have r experiments, the first with n_1 possible outcomes, the second with n_2 possible outcomes, and so on (up to the r^{th} with n_r possible outcomes), then the total number of possible outcomes is
>
> $$n_1 \cdot n_2 \cdot \ldots \cdot n_r = \prod_{j=1}^{r} n_j.$$

The product sign $\prod$ is just a useful abbreviation for large products - as the sum symbol $\sum$ is for sums.

1.3 Example. Consider 7-digit license plates where the first three places have to be filled with letters and the four others with numbers.

a) How many different license plates do exist?

b) And what happens if every letter and every number appears at most once on the plate?

Solution.

a) Of course, by the generalized principle of counting we have

$$26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 26^3 \cdot 10^4 = 175,760,000.$$

b) The same procedure gives us now

$$26 \cdot 25 \cdot 24 \cdot 10 \cdot 9 \cdot 8 \cdot 7 = 78,624,000.$$

□

1.2 Permutations

1.4 Example. How many ways are there to arrange the letters a, b, and c?

Solution. We just try again to list all of them, see Table 1.3. It turns out that we have six possibilities. □

a	b	c
a	c	b
b	a	c
b	c	a
c	a	b
c	b	a

Table 1.3: List of all arrangements of the letters a, b and c.

Rearrangement possibilities are known in mathematics as *permutations*. We will next try to find a way to make the counting of permutations efficient. Analyzing Example 1.4, we observe the following:

- There are three possibilities for the first letter:

$$\underline{a}\,\square\square\square \qquad \underline{b}\,\square\square\square \qquad \underline{c}\,\square\square\square$$

- Once having fixed the first letter, say b, there are only two possibilities for the second letter:

$$\underline{b}\,\underline{a}\,\square\square \qquad \underline{b}\,\underline{c}\,\square\square$$

- Having now the first two letters fixed, say b and a, there is only one remaining letter for the last place

$$\underline{b}\,\underline{a}\,\underline{c}\,\square$$

Thus we have in total

$$3 \cdot 2 \cdot 1 = 6$$

possibilities, as we had already found it first by naive counting. In general, if we have n objects, we have

The factorial – denoted by ! – will be used often.

$$n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1 = \prod_{j=1}^{n} j = n!$$

possibilities (or permutations).

1.5 Example. There are n male and m female students in the probability class. We assume that at the end of the term they will all have different scores.

a) How many different rankings are possible?

b) If male and female students are ranked separately (as in Olympics), how many different rankings would exist?

Solution.

a) As we have in total $(n+m)$ students, it follows that we have $(n+m)!$ permutations.

b) In this case we have $n!$ possible rankings for the male students and $m!$ rankings for the female students. Thus, by the principle of counting, we have in total $n! \cdot m!$ permutations.

Note that $n! \cdot m! \neq (n \cdot m)!$ as, e.g., $(2 \cdot 2)! = 24 > 4 = 2! \cdot 2!$

□

1.6 Example. How many words can be formed with the letters from the following word?

PEPPER

So are we looking for *anagrams?* How about CONSTANTINOPLE?

Solution. It seems that we have a problem now. The letter **P** appears here three times and the letter **E** twice. Interchanging thus, e.g, just two of **P**s will not create an additional word. What we can do is the following:

- Let's first do as the different **P**s and **E**s would be indeed different letters and count all their arrangements. Let's introduce subscripts to mark the distinction:

  $$\mathbf{P_1E_1P_2P_3E_2R}$$

  Now we have six different letters and thus 6! permutations.

- Next we recognize that the three **P**s ($\mathbf{P_1}$, $\mathbf{P_2}$ and $\mathbf{P_3}$) can be arranged in 3! ways.

- Similarly we have 2! ways to arrange the two **E**s.

- Thus there are in total

  $$\frac{6!}{3! \cdot 2!} = \frac{6 \cdot 5 \cdot 4}{2} = 60$$

  possible arrangements.

The easiest way to see the last step is to apply the principle of counting backwards: The sought result multiplied with 3! and 2! has to be 6!.

□

1.3 Combinations

1.7 Example. Assume we have five items, **A**, **B**, **C**, **D**, and **E**. How many different groups of three can we form out of them?

Solution. Oh, this is easy, we can just list them again, see Table 1.4. □

ABC	**ABD**	**ABE**
ACD	**ACE**	**ADE**
BCD	**BCE**	**BDE**
CDE		

Table 1.4: All three letter groups made out of **ABCDE**.

Again, we should rather look for a way of counting which is easily expandable to large numbers of items and groups. Proceeding as before shows us

- There are 5 ways to pick the 1^{st} item
- There are 4 ways to pick the 2^{nd} item
- There are 3 ways to pick the 3^{rd} item

Thus we have $5 \cdot 4 \cdot 3 = 60$ different groups. Note however that here the order of the elements plays a role. To discard the effect of ordering, we have to divide by the number of permutations which can be made from three objects in a group, thus $3 \cdot 2 \cdot 1 = 6$. In total we get

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

ways to choose three elements out of five.

1.8 Definition. In general there are by the indicated system of counting

$$\begin{aligned} & \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k \cdot \ldots \cdot 2 \cdot 1} \\ = & \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1) \cdot (n-k) \cdot \ldots \cdot 2 \cdot 1}{(k \cdot \ldots \cdot \cdots 2 \cdot 1) \cdot ((n-k) \cdot \ldots \cdot 2 \cdot 1)} \\ = & \frac{n!}{k! \cdot (n-k)!} \end{aligned}$$

Note that $0! = 1$ by convention. This makes sense in the *choose* context: There is exactly one way to pick 0 elements out of n.

ways to choose k elements out of n. We denote this by the *binomial coefficient* (or *choose*) notation $\binom{n}{k}$ (which is declared for $0 \leq k \leq n$),

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

1.9 Lemma. *For $1 \leq k \leq n-1$ it holds that*

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

In general one defines $\binom{n}{k} = 0$ if $k < 0$ or $k > n$.

Proof. We remember that $\binom{n}{k}$ is the number of ways we can pick k elements out of n. Now we fix an arbitrary element out of the n and denote it by ★. Note now the following:

Number of groups of k elements that do contain ★: $\binom{n-1}{k-1}$

Number of groups of k elements that do NOT contain ★: $\binom{n-1}{k}$

Number of groups of k elements: $\binom{n}{k}$

It is clear that we combine all groups of k elements that do contain ★ with all that do not contain ★, we will just get the number of all groups of k elements, proving the lemma, see also Figure 1.1. □

Alternatively this can be proved by direct calculations...

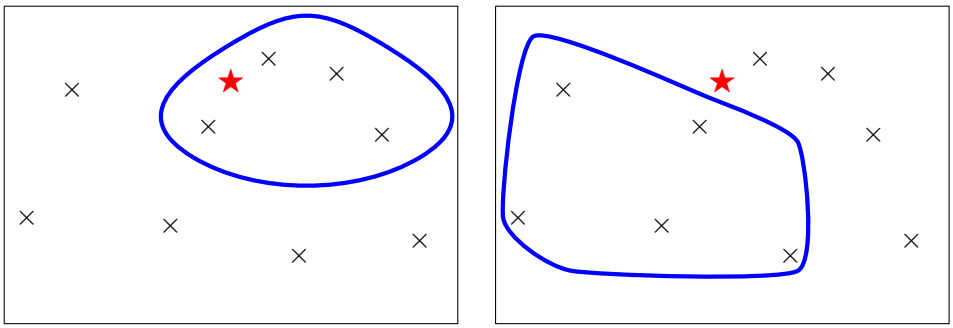


Figure 1.1: Proof of Lemma 1.9: Left: Choosing k elements including ★ is effectively picking $k-1$ element out of $n-1$. Right: Choosing k elements *not* including ★ is effectively picking k element out of $n-1$.

The relationship of Lemma 1.9 can be illustrated via *Pascal's triangle* as illustrated in Figure 1.2.

1.10 Theorem (The Binomial Theorem)**.** *Let $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$, then it holds that*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}. \tag{1.1}$$

We use the convention that the set of natural numbers $\mathbb{N}$ contains 0.

Proof. We prove the theorem by induction over n. Let's check for the first values of n

$$n = 0: \qquad (x+y)^0 = 1 = \binom{0}{0} x^0 y^0, \quad \checkmark$$

$$n = 1: \qquad (x+y)^1 = x + y = \binom{1}{0} x^1 y^0 + \binom{1}{1} x^0 y^1. \quad \checkmark$$

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

Figure 1.2: *Pascal's triangle*: every binomial coefficient is the sum of the two above.

Now we assume that (1.1) is true for $n-1$. Then it follows that

Induction is the steamroller among the mathematical techniques of proof.

$$\begin{aligned}
(x+y)^n &= (x+y)\cdot(x+y)^{n-1} \overset{Ass.}{=} (x+y)\cdot\left(\sum_{k=0}^{n-1}\binom{n-1}{k}x^k y^{n-k-1}\right) \\
&= \sum_{k=0}^{n-1}\binom{n-1}{k}x^{k+1}y^{n-k-1} + \sum_{k=0}^{n-1}\binom{n-1}{k}x^k y^{n-k} \\
&\overset{Subst.}{=} \sum_{j=1}^{n}\binom{n-1}{j-1}x^j y^{n-j} + \sum_{j=0}^{n-1}\binom{n-1}{j}x^j y^{n-j} \\
&= x^n + \sum_{j=1}^{n-1}\binom{n-1}{j-1}x^j y^{n-j} + \sum_{j=1}^{n-1}\binom{n-1}{j}x^j y^{n-j} + y^n \\
&= x^n + \sum_{j=1}^{n-1}\left(\binom{n-1}{j-1} + \binom{n-1}{j}\right)x^j y^{n-j} + y^n \\
&\overset{Lem.\ 1.9}{=} x^n + \sum_{j=1}^{n-1}\binom{n}{j}x^j y^{n-j} + y^n \\
&= \sum_{j=0}^{n}\binom{n}{j}x^j y^{n-j},
\end{aligned}$$

We substitute in the left sum $j = k+1$ and in the right one $j = k$

thus it is also true for n. It follows that the (1.1) is true for every natural number n. □

We are going to turn back to our initial motivating example, Example 1.1. Is there a general way to calculate the number of set-ups of working and defective antennas?

1.11 Example (Continuation of Example 1.1)**.** We assume that we have n antennas, k of which are defective.

a) How many line-ups of these antennas are possible?

b) In how many of them is the communication system working?

Solution.

a) The answer to the first question is not hard, as this means just assigning the defective antennas to k spots (the remaining are then automatically filled with functional antennas). And we know that there are exactly $\binom{n}{k}$ ways to do this.

b) Here we have to change a bit the way how we think. We have in total $n-k$ working antennas. We just line them up and then try to add the defective antennas. We can put any defective antenna in any spot between two working antennas, or at each end of the line, which gives $n-k+1$ possibilities, see Figure 1.3. And at any such place we can put at most one antenna, otherwise we would have two neighboring defective antennas. Thus we have to choose just k spots out of those $n-k+1$ to find all possible functioning arrangements – and there are $\binom{n-k+1}{k}$ of them.

Note that a positive solution requires of course $n-k+1 \geq k$ or, equivalently $k \leq \frac{n+1}{2}$.

□

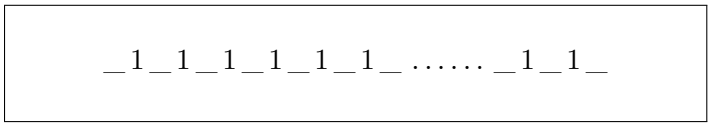


Figure 1.3: Antenna arrangements. 1 stands for a working antenna and _ stands for a place at which we can put a defective one.

Of course we can use this result to calculate the probability that the communication system is working, it is

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

Chapter 2

The Axioms of Probability

2.1 Sample Spaces and Events

2.1 Definition. A *sample space* is the collection of all possible outcomes of an experiment. We will denote a sample space by Ω.

We will always assume that $\Omega \neq \emptyset$.

2.2 Example.

a) A simple coin flip:
$$\Omega = \{H, T\}.$$

b) Two consecutive coin flips:
$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}.$$

c) The lifetime of a light bulb (in hours):
$$\Omega = \{x \in \mathbb{R} : x \geq 0\} = [0, \infty).$$

□

2.3 Definition. An *event* is a subset of the sample space, i.e., a collection of possible outcomes of the experiment.

2.4 Example. For the examples of Example 2.2:

a) $E = \{H\}$ is the event that a head is thrown.

b) $E = \{(H, T), (H, H)\}$ is the event that a head comes up in the first toss.

c) $E = \{x \in \mathbb{R} : x \geq 5\}$ is the event that the light bulb lasts for at least five hours.

□

Note that $\emptyset$ and Ω are always possible events in Ω. For any two events E, F of a sample space we can define the following operations:

- $E \cup F$ (*union*): either E or F (or both of them) occur.
- $E \cap F$ (*intersection*): both events, E and F, occur.
- $E^c = \Omega \setminus E$ (*complement*): E does not occur.

2.5 Example.

a) Rolling two dice:

$$\begin{aligned}
\Omega &= \{(i,j) : 1 \leq i,j \leq 6\},\\
E &= \{(1,j) : 1 \leq i \leq 6\} = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}\\
&= \text{"first die a one"},\\
F &= \{(i,j) : i+j = 5\} = \{(1,4),(4,1),(2,3),(3,2)\}\\
&= \text{"sum of all pips of both dice is five"},
\end{aligned}$$

$$\begin{aligned}
E \cup F &= \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(4,1),(2,3),(3,2)\},\\
E \cap F &= \{(1,4)\},\\
E^c &= \{(i,j) : 2 \leq i \leq 6, 1 \leq j \leq 6\}.
\end{aligned}$$

b) Lifetime of a light bulb:

$$\begin{aligned}
\Omega &= \{x \in \mathbb{R} : x \geq 0\},\\
E &= \{x \in \mathbb{R} : x \geq 3\} = [3,\infty),\\
&= \text{"light bulb lasts for at least three years"},\\
F &= \{x \in \mathbb{R} : x < 5\} = [0,5)\\
&= \text{"light bulb lasts for less than five years"},
\end{aligned}$$

$$\begin{aligned}
E \cup F &= \{x \in \mathbb{R} : x \geq 0\} = [0,\infty) = \Omega,\\
E \cap F &= \{x \in \mathbb{R} : 3 \leq x < 5\} = [3,5),\\
E^c &= \{x \in \mathbb{R} : 0 \leq x < 3\} = [0,3).
\end{aligned}$$

□

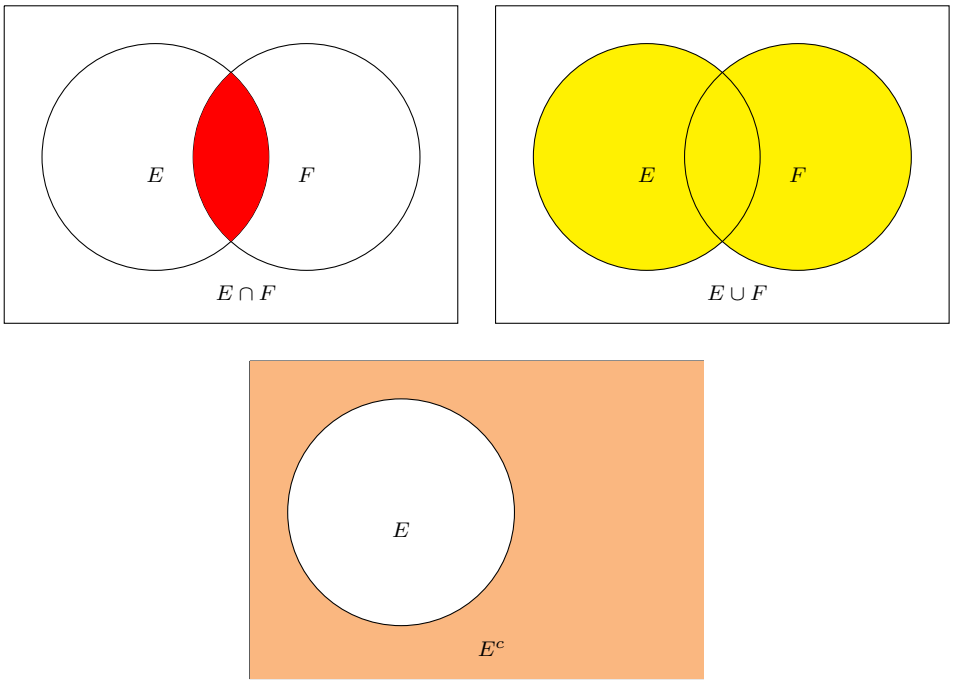


Figure 2.1: *Venn diagrams* representing the basic operations of events.

2.6 Remark.

a) Events can be conveniently represented in *Venn diagrams*, see Figure 2.1.

b) $\cup$, $\cap$, and ^{c} obey some rules.

- Commutative laws for $\cup$ and $\cap$:

$$E \cap F = F \cap E, \qquad E \cup F = F \cup E.$$

- Associative laws for $\cup$ and $\cap$:

$$(E \cap F) \cap G = E \cap (F \cap G), \qquad (E \cup F) \cup G = E \cup (F \cup G).$$

- Distributive laws for $\cup$ and $\cap$:

$$E \cap (F \cup G) = (E \cap F) \cup (E \cap G), \qquad E \cup (F \cap G) = (E \cup F) \cap (E \cup G).$$

This can be visualized in a Venn diagram, see Figure 2.2.

– Involution for ^{c}:

$$(E^c)^c = E.$$

c) By associativity, we can define unions and intersections also for more than two events, even for countably many. If E_1, E_2, ... is a sequence of events in a sample space Ω, then we write their union (resp. intersection)

$$\bigcup_{j=1}^{\infty} E_j, \qquad \bigcap_{j=1}^{\infty} E_j.$$

d) We have $\emptyset^c = \Omega$ and $\Omega^c = \emptyset$.

e) If $E \cap F = \emptyset$ we call E and F *disjoint* (or *mutually exclusive*).

f) We write $E \subseteq F$ if all the outcomes in E are also contained in F and $E \supseteq F$ if all the elements of F are also contained in E. $E \subseteq F$ and $E \supseteq F$ imply together $E = F$.

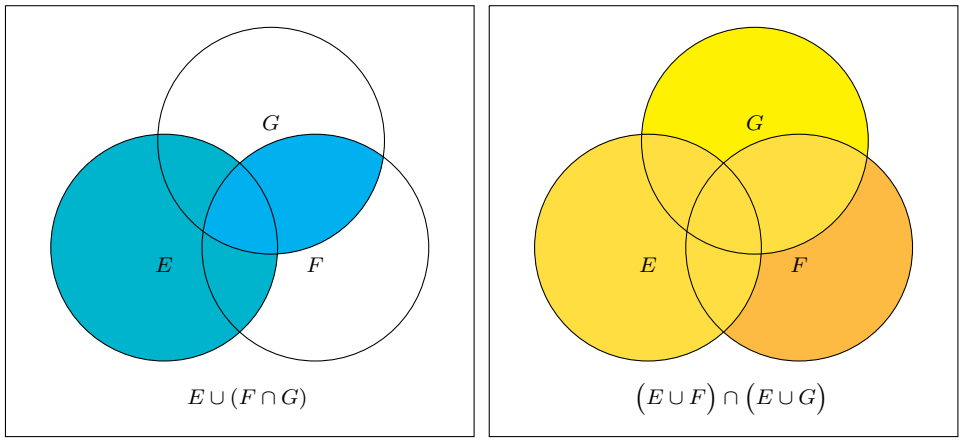


Figure 2.2: *Venn diagram* illustrating the second distributive law.

2.7 Proposition (De Morgan laws)**.** *For multiple (or even countably many; $n = \infty$) events, the* De Morgan laws *hold:*

a)

$$\left(\bigcup_{j=1}^{n} E_j\right)^c = \bigcap_{j=1}^{n} E_j^c.$$

b)

$$\Big(\bigcap_{j=1}^{n} E_j\Big)^c = \bigcup_{j=1}^{n} E_j^c.$$

Proof. a) We proceed by equivalence transformations: for any $\omega \in \Omega$ we have

$$\begin{aligned}
\omega \in \Big(\bigcup_{j=1}^{n} E_j\Big)^c &\iff \omega \notin \bigcup_{j=1}^{n} E_j \\
&\iff \text{for all } j,\ 1 \le j \le n,\ \omega \text{ is not in any of the } E_j \\
&\iff \text{for all } j,\ 1 \le j \le n,\ \omega \text{ is contained in every } E_j^c \\
&\iff \omega \in \bigcap_{j=1}^{n} E_j^c.
\end{aligned}$$

b) Using the result from part a) we conclude

$$\Big(\bigcap_{j=1}^{n} E_j\Big)^c = \Big(\bigcap_{j=1}^{n} \big(E_j^c\big)^c\Big)^c \overset{a)}{=} \Big(\Big(\bigcup_{j=1}^{n} E_j^c\Big)^c\Big)^c = \bigcup_{j=1}^{n} E_j^c.$$

□

2.2 The Axioms of Probability

We want to assign now probabilities to events. The following axioms capture our intuition of probability.

2.8 Definition. Consider an experiment with sample space Ω. A probability is a function that assigns to every event in Ω areal number such that

The idea that one can also describe probability by axioms goes back to *Kolmogorov* (1933).

i) $0 \le \mathbb{P}[E] \le 1$ for every $E \subseteq \Omega$.

ii) $\mathbb{P}[\Omega] = 1$.

iii) If E_1, E_2, ... are disjoint events (i.e., $E_i \cap E_j = \emptyset$ for all $i \neq j$), then

$$\mathbb{P}\Big[\bigcup_{j=1}^{\infty} E_j\Big] = \sum_{j=1}^{\infty} \mathbb{P}[E_j].$$

2.9 Example.

a) Tossing a fair coin. We have $\Omega = \{H, T\}$ and

$$\mathbb{P}[\{H\}] = \mathbb{P}[\{T\}] = \frac{1}{2}.$$

By a *fair coin* we mean a coin where both, *heads* and *tails* have the same probability to land on the top.

b) Let again $\Omega = \{H, T\}$, but now

$$\mathbb{P}[\{H\}] = \frac{2}{3}, \qquad \mathbb{P}[\{T\}] = \frac{1}{3}.$$

This models tossing a biased coin.

c) Rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$. For a fair die we have

$$\mathbb{P}[\{1\}] = \mathbb{P}[\{2\}] = \mathbb{P}[\{3\}] = \mathbb{P}[\{4\}] = \mathbb{P}[\{5\}] = \mathbb{P}[\{6\}] = \frac{1}{6}.$$

If we are very exact, we note that we use here a property that will only be proved in Proposition 2.10 b)

It follows then that

$$\mathbb{P}[\{2, 4, 6\}] = \mathbb{P}[\{2\}] + \mathbb{P}[\{4\}] + \mathbb{P}[\{6\}] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Thus the probability to throw an even number on top is 50%. □

We can derive some further properties of probabilities directly from the axioms.

2.10 Proposition. *Let Ω be a sample space, then*

a) The empty set $\emptyset$ has zero probability,

$$\mathbb{P}[\emptyset] = 0.$$

b) For any $n \in \mathbb{N}$ and events E_1, E_2, …, E_n it holds that

$$\mathbb{P}\Big[\bigcup_{j=1}^{n} E_j\Big] = \sum_{j=1}^{n} \mathbb{P}[E_j].$$

c) For every event $E \subseteq \Omega$

$$\mathbb{P}[E^c] = 1 - \mathbb{P}[E].$$

d) If $E \subseteq F$ for two events $E, F \subseteq \Omega$, then

$$\mathbb{P}[E] \leq \mathbb{P}[F].$$

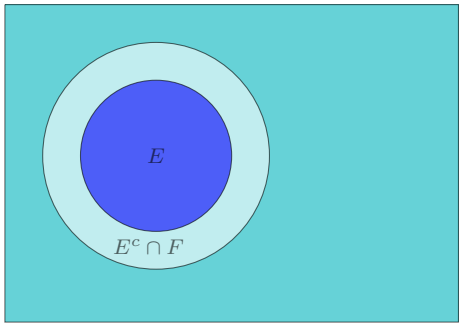


Figure 2.3: The decompositions in part d) of Proposition 2.10.

e) For $E, F \subseteq \Omega$ we have

$$\mathbb{P}[E \cup F] = \mathbb{P}[E] + \mathbb{P}[F] - \mathbb{P}[E \cap F].$$

Proof. a) Setting $E_j = \emptyset$ in Definition 2.8 iii) yields

$$\mathbb{P}[\emptyset] = \mathbb{P}\Big[\bigcup_{j=1}^{\infty} \emptyset\Big] = \sum_{j=1}^{\infty} \mathbb{P}[\emptyset]. \tag{2.1}$$

As $\mathbb{P}[\emptyset]$ is just a real number between 0 and 1, $\mathbb{P}[\emptyset] = a \in [0,1]$, it follows that $\mathbb{P}[\emptyset] = 0$ as for any $a > 0$ the right hand of (2.1) would diverge, while the left hand is smaller than one.

b) Similarily, setting $E_k = \emptyset$ for all $k \geq n+1$ yields

$$\begin{aligned}\mathbb{P}\Big[\bigcup_{j=1}^{n} E_j\Big] &= \mathbb{P}\Big[\Big(\bigcup_{j=1}^{n} E_j\Big) \cup \Big(\bigcup_{j=n+1}^{\infty} E_j\Big)\Big] = \mathbb{P}\Big[\bigcup_{j=1}^{\infty} E_j\Big] \\ &= \sum_{j=1}^{\infty} \mathbb{P}[E_j] = \sum_{j=1}^{n} \mathbb{P}[E_j] + \sum_{j=n+1}^{\infty} \underbrace{\mathbb{P}[E_j]}_{=0} = \sum_{j=1}^{\infty} \mathbb{P}[E_j].\end{aligned}$$

c) We have $E \cup E^c = \Omega$ and $E \cap E^c = \emptyset$, thus

$$1 = \mathbb{P}[\Omega] = \mathbb{P}[E \cup E^c] \overset{b)}{=} \mathbb{P}[E] + \mathbb{P}[E^c],$$

implying

$$\mathbb{P}[E^c] = 1 - \mathbb{P}[E].$$

d) Note that

$$F = E \cup (E^c \cap F) \qquad \text{and} \qquad E \cap (E^c \cap F) = \emptyset,$$

and thus

$$\mathbb{P}[F] = \mathbb{P}\Big[E \cup (E^c \cap F)\Big] \overset{b)}{=} \mathbb{P}[E] + \underbrace{\mathbb{P}\big[E^c \cap F\big]}_{\geq 0} \geq \mathbb{P}[E].$$

For an illustration see Figure 2.3.

e) We note first that

$$F = (E \cap F) \cup (E^c \cap F) \qquad \text{and} \qquad (E \cap F) \cup (E^c \cap F) = \emptyset,$$

and thus

$$\mathbb{P}[F] \overset{b)}{=} \mathbb{P}[E \cap F] + \mathbb{P}\big[E^c \cap F\big]. \tag{2.2}$$

Moreover, as

$$E \cup F = E \cup (E^c \cup F) \qquad \text{and} \qquad E \cap (E^c \cup F) = \emptyset,$$

we have by (2.2)

$$\begin{aligned} \mathbb{P}[E \cup F] \;&=\; \mathbb{P}\Big[E \cup (E^c \cap F)\Big] \overset{b)}{=} \mathbb{P}[E] + \mathbb{P}\big[E^c \cap F\big] \\ &\overset{(2.2)}{=} \mathbb{P}[E] + \mathbb{P}[F] - \mathbb{P}[E \cap F]. \end{aligned}$$

For an illustrattion see Figure 2.4.

□

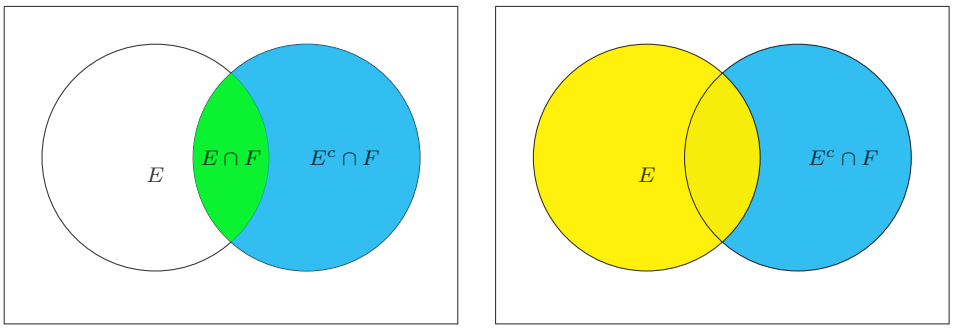


Figure 2.4: The decompositions in part e) of Proposition 2.10.

Using the De Morgan laws (Proposition 2.7) we can extend part e) of the previous Proposition to multiple events.

2.11 Proposition (Inclusion-Exclusion Principle)**.** *For events E_1, E_2, …, E_n in a sample space Ω it holds that*

$$\begin{aligned}\mathbb{P}[E_1 \cup \cdots \cup E_n] = &\sum_{j=1}^{n} \mathbb{P}[E_j] \\ &- \sum_{1 \le j_1 < j_2 \le n} \mathbb{P}[E_{j_1} \cap E_{j_2}] \\ &+ \cdots \\ &+ (-1)^{r-1} \sum_{1 \le j_1 < \cdots < j_r \le n} \mathbb{P}[E_{j_1} \cap \cdots \cap E_{j_r}] \\ &+ \cdots \\ &+ (-1)^{n-1} \mathbb{P}[E_1 \cap \cdots \cap E_n],\end{aligned}$$

or, in more compact form,

$$\mathbb{P}\Big[\bigcup_{j=1}^{n} E_j\Big] = \sum_{r=1}^{n} (-1)^{r-1} \sum_{1 \le j_1 < \cdots < j_r \le n} \mathbb{P}\Big[\bigcap_{j=1}^{r} E_{j_r}\Big].$$

Proof. The proof by induction over n is quite straightforward, but also quite tedious and thus omitted, an illustration for the case $n = 3$ is given in Figure 2.5. □

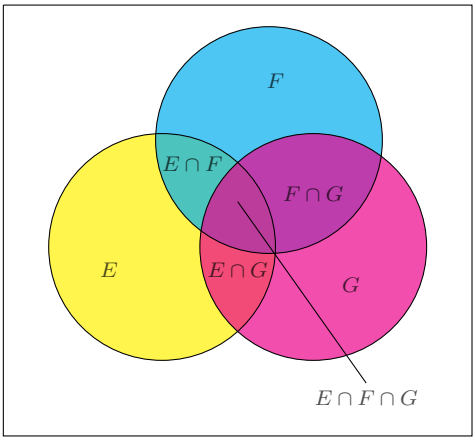


Figure 2.5: The *Inclusion-Exclusion Principle* for $n = 3$.

2.3 Interpretation

From a purely mathematical viewpoint, the interpretation of probability is not important. However, it helps to understand better what we are doing.

At least we hope that our mathematical theory is the model for something happening in reality, no?

There are two main interpretations of probability:

1. *Frequentist interpretation*: The probability of an event is the average percentage of outcomes of an experiment if repeated under identical conditions very often (e.g., "The probability to get a '6' when rolling a die is $\frac{1}{6}$".)
2. *Subjective interpretation*: The probability of an event is a measure of somebody's believe that the event will happen (e.g., "I think that the chance that the *Red Sox* will win tomorrow is 30%".)

For more about interpretations of probability and a philosophical take of our subject, we refer to wikipedia or the Stanford Encyclopedia of Philosophy:
https://en.wikipedia.org/wiki/Probability_interpretations
https://plato.stanford.edu/entries/probability-interpret/

2.4 Laplacian Sample Spaces

In many experiments, we have the case that every outcome is equally likely (e.g., when rolling a fair die or flipping a fair coin). In the case of a sample space $\Omega = \{1, 2, \dots, N\}$ this means that

$$\mathbb{P}[\{1\}] = \mathbb{P}[\{2\}] = \cdots = \mathbb{P}[\{N\}]].$$

Clearly, it follows from the axioms of probability (Definition 2.8 c)) that

$$1 = \mathbb{P}[\Omega] = \mathbb{P}[\{1\}] + \mathbb{P}[\{2\}] + \cdots + \mathbb{P}[\{N\}] = N \cdot \mathbb{P}[\{1\}],$$

and thus

$$\mathbb{P}[\{1\}] = \mathbb{P}[\{2\}] = \cdots = \mathbb{P}[\{N\}] = \frac{1}{N}.$$

Laplacian sample spaces are named after *Pierre-Simon Laplace* (1749-1827) who made important contributions to the development of probability theory.

Thus in a frequentist interpretation, denoting by $|E|$ the number of elements of the event $E \subseteq \Omega$, we have that

$$\mathbb{P}[E] = \frac{|E|}{|\Omega|} = \frac{\text{number of outcomes in E}}{\text{number of outcomes in } \Omega}$$

$$\sim \frac{\text{"number of successful outcomes"}}{\text{"number of all outcomes"}}.$$

Sample spaces with equally likely outcomes are called *Laplacian sample space*.

2.12 Example. Rolling two dice, what is the probability that the sum of them is seven?

Of course, we mean by *the sum of two dice* more precisely *the sum of the pips on the top of both dice...*

Solution. We start by modeling sample space and event, i.e.,

$$\Omega = \{(i,j) \,:\, 1 \leq i,j \leq 6\},$$
$$E = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}.$$

Thus we have

$$\mathbb{P}[E] = \frac{6}{36} = \frac{1}{6}.$$

□

2.13 Example. An urn contains eleven balls, six of which are white and five black. We draw three balls, what is the probability that one of them is white and two are black?

For some (rather historical than morbid) reasons the boxes containing balls so ubiquitous in probability are called *urns*.

Solution.
Solution 1: We model the sample space as all ordered draws (hence triplets). There are

$$11 \cdot 10 \cdot 9 = 990$$

possible outcomes. Now we count through the different possible configurations:

- First ball white, second and third ball black:

$$6 \cdot 5 \cdot 4 = 120.$$

- First ball black, second ball white, and third ball black:

$$5 \cdot 6 \cdot 4 = 120.$$

- First and second ball black, third ball white:

$$5 \cdot 4 \cdot 6 = 120.$$

Thus we have for the event in question

$$\mathbb{P}[E] = \frac{120+120+120}{990} = \frac{360}{990} = \frac{4}{11} \approx 36.16\%.$$

Solution 2: Another way to solve the problem is the following. There are in total $\binom{11}{3}$ possible outcomes (as we draw three balls out of eleven). Moreover we have

- $\binom{6}{1}$ possibilities to draw the one white ball out of six,
- $\binom{5}{2}$ possibilities to draw two black balls out of five.

Thus we have again

$$\mathbb{P}[E] = \frac{\binom{6}{1} \cdot \binom{5}{2}}{\binom{11}{3}} = \frac{60}{165} = \frac{4}{11} \approx 36.16\%.$$

□

2.14 Example. How likely is it that (at least) two of 16 students in class have the same birthday?

Let's forget about leap years for a moment...

Solution. The basic idea is to calculate first the probability of the reverse event, i.e., how likely it is that all 16 students have birthdays at different days. Thus we have

$$\frac{365 \cdot 364 \cdot 363 \cdots (365 - 16 + 1)}{365 \cdot 365 \cdot 365 \cdots 365} \approx 0.7164.$$

Thus the probability that at least two of the students have the same birthday is by Proposition 2.10 c) approximately

$$1 - 0.7164 = 28.36\%.$$

□

The situation above is often called the *birthday paradox* as a relatively small number of people suffice to get a high probability that at least two of them have at the same day birthday (at least small compared to the number of 365 days per year). As we have seen that this stems from the fact that we have actually to calculate the probability that *all* people have birthdays at different days which is given by a product of decreasing factors. One gets that already 23 people suffice to get a chance of more than 50% and 57 people to get a probability of over 99%. More information about the birthday paradox can be found on wikipedia or wolfram alpha:

https://en.wikipedia.org/wiki/Birthday_problem

https://www.wolframalpha.com/input/?i=birthday+problem+calculator&a=FSelect_**BirthdayProblem--&f2=23&f=BirthdayProblemWithLeapYear.n_23

Chapter 3

Conditional Probability and Independence

3.1 Conditional Probability

3.1 Example. Consider again the experiment of rolling two dice and the following events.

$$\begin{aligned} E &= \text{“the sum of the pips on top of the two dice is 6”} \\ &= \{(1,5),(2,4),(3,3),(4,2),(5,1)\}, \\ F &= \text{“the first die two pips on the top”} \\ &= \{(2,1),(2,2),(2,3),(2,4),(2.5),(2,6)\}. \end{aligned}$$

Note that we are sometimes a bit sloppy and do not write down the sample space explicitly. In the case that you have troubles following an example, it is recommended to do the work and write down Ω explicitly.

We have clearly

$$\mathbb{P}[E] = \frac{5}{36} \qquad \text{and} \qquad \mathbb{P}[F] = \frac{6}{36} = \frac{1}{6}.$$

If we assume now that we have already the information that the first die showed a 2 (while the second die is still rolling), what is now the probability that the sum is 6?

As the only possible outcome of “the sum of the pips on top of the two dice is 6” where the first die shows two pips on top is $(2,4)$, we have now

$$\frac{\mathbb{P}\big[\{(2,4)\}\big]}{\mathbb{P}\Big[\{(2,1),(2,2),(2,3),(2,4),(2.5),(2,6)\}\Big]} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{6}{36} = \frac{1}{6}.$$

We note that this is bigger than $\frac{5}{36}$, thus our chances to get a sum of six have improved as we learned the outcome of the first die. □

The crucial lesson from the above example is that if we know already something about the experiment, then we have only to consider the cases where this holds true. We call this the *conditional probability of E given F and denote it by* $\mathbb{P}[E \,|\, F]$.

3.2 Definition. The conditional probability of E given F (with $\mathbb{P}[F] > 0$) is defined as

$$\mathbb{P}[E \,|\, F] := \frac{\mathbb{P}[E \cap F]}{\mathbb{P}[F]}. \tag{3.1}$$

3.3 Example. A student is undecided if he takes *MA 2621 Probability for Applications* or *MA 2631 Probability*. He believes her chances to get an A are 75% in *MA 2621* and an A are 50% in *MA 2621*, but on the other hand he is more interested in the real stuff. Finally he decides to flip a coin. What is the chance that he ends up getting an A in *MA 2631*?

Solution. We denote the following events,

A	...	he gets an A,
M	...	he takes *MA 2631*.

We know that $\mathbb{P}[M] = 0.5$ and $\mathbb{P}[A \,|\, M] = 0.5$ and want to know $\mathbb{P}[A \cap M]$. This we can deuce from the definition of the conditional probability,

$$\mathbb{P}[A \cap M] = \mathbb{P}[M] \cdot \mathbb{P}[A \,|\, M] = 0.5 \cdot 0.5 = 0.25.$$

Thus his chances to end up with an A in *MA 2631* are 25%. □

3.4 Example. A student takes a one hour exam. The probability that she finishes after x hours ($0 \leq x < 1$) is $\frac{x}{2}$. If we know that after 45 minutes she is still working, what is the probability that she will need the full hour?

Solution. Let us introduce the following notation,

E_x	...	she finishes after x hours ($0 \leq x < 1$),
F	...	she needs the whole hour.

In this notation, it is our goal to calculate $\mathbb{P}[F \,|\, E_{0.75}^c]$. To do so we note first that

$$\mathbb{P}[F] = \mathbb{P}[E_1^c] = 1 - \mathbb{P}[E_1] = 1 - \frac{1}{2} = \frac{1}{2},$$

and

$$\mathbb{P}[F \cap E_x^c] = \mathbb{P}[F],$$

for every $x \in [0,1)$, as $F \subseteq E_x^c$. From this it follows by Definition 3.2

$$\begin{aligned}\mathbb{P}[F \mid E_{0.75}^c] &= \frac{\mathbb{P}[F \cap E_{0.75}^c]}{\mathbb{P}[E_{0.75}^c]} = \frac{\mathbb{P}[F]}{1 - \mathbb{P}[E_{0.75}]} = \frac{\frac{1}{2}}{1 - \frac{0.75}{2}} \\ &= \frac{\frac{1}{2}}{1 - \frac{3}{8}} = \frac{\frac{1}{2}}{\frac{5}{8}} = \frac{4}{5} = 80\%.\end{aligned}$$

□

We note that we can redefine conditional probability by rewriting (3.1) in the following way.

$$\mathbb{P}[E \cap F] = \mathbb{P}[E \mid F] \cdot \mathbb{P}[E].$$

Trying to generalize this to three events E_1, E_2, E_3 gives

$$\begin{aligned}\mathbb{P}[E_1 \cap E_2 \cap E_3] &= \mathbb{P}\Big[(E_1 \cap E_2) \cap E_3\Big] = \mathbb{P}[E_3 \mid E_1 \cap E_2] \cdot \mathbb{P}[E_1 \cap E_2] \\ &= \mathbb{P}[E_3 \mid E_1 \cap E_2] \cdot \mathbb{P}[E_2 \mid E_1] \cdot \mathbb{P}[E_1].\end{aligned}$$

In general we will prove the following proposition.

3.5 Proposition. *In general we have for events E_1, E_2, $\dots E_n$ (satisfying $\mathbb{P}[\bigcap_{j=1}^{n-1} E_j] > 0$)*

$$\mathbb{P}[E_1 \cap E_2 \cap \cdots \cap E_n] = \mathbb{P}[E_n \mid E_1 \cap E_2 \cap \cdots \cap E_{n-1}] \cdot \ldots \cdot \mathbb{P}[E_2 \mid E_1] \cdot \mathbb{P}[E_1],$$

or more compact,

$$\mathbb{P}\Big[\bigcap_{j=1}^{n} E_j\Big] = \prod_{k=1}^{n-1} \mathbb{P}\Big[E_{k+1} \,\Big|\, \bigcap_{j=1}^{k} E_j\Big] \cdot \mathbb{P}[E_1].$$

Proof. Expanding the conditional probabilities according to (3.1) and reducing the fractions yields

$$\begin{aligned}&\mathbb{P}[E_n \mid E_1 \cap E_2 \cap \cdots \cap E_{n-1}] \cdots \mathbb{P}[E_2 \mid E_1] \cdot \mathbb{P}[E_1] \\ &= \frac{\mathbb{P}[E_1 \cap E_2 \cap \cdots \cap E_n]}{\cancel{\mathbb{P}[E_1 \cap E_2 \cap \cdots \cap E_{n-1}]}} \cdot \frac{\cancel{\mathbb{P}[E_1 \cap E_2 \cap \cdots \cap E_{n-1}]}}{\cancel{\mathbb{P}[E_1 \cap E_2 \cap \cdots \cap E_{n-2}]}} \cdots \\ &\quad \cdots \frac{\cancel{\mathbb{P}[E_1 \cap E_2]}}{\cancel{\mathbb{P}[E_1]}} \cdot \cancel{\mathbb{P}[E_1]} \\ &= \mathbb{P}[E_1 \cap E_2 \cap \cdots \cap E_n].\end{aligned}$$

□

We note that the conditional probability is itself a probability.

3.6 Proposition. *Let C be an event in a sample space Ω such that we have $P[C] > 0$. Then*

$$\mathbb{Q}[\,\cdot\,] = \mathbb{P}[\,\cdot \mid C]$$

is a probability.

Proof. We have just to check that $\mathbb{Q}$ satisfies the axioms of probability.

i) We have for any event $E \subseteq \Omega$ by the definition of conditional probability, Definition 3.2,

$$\mathbb{Q}[E] = \mathbb{P}[E \mid C] = \frac{\mathbb{P}[E \cap C]}{\mathbb{P}[C]} \geq 0,$$

$$\mathbb{Q}[E] = \mathbb{P}[E \mid C] = \frac{\mathbb{P}[E \cap C]}{\mathbb{P}[C]} \overset{\text{Prop. 2.10 } d)}{\leq} \frac{\mathbb{P}[C]}{\mathbb{P}[C]} = 1.$$

ii) In a smilar way we have

$$\mathbb{Q}[\Omega] = \mathbb{P}[\Omega \mid C] = \frac{\mathbb{P}[\Omega \cap C]}{\mathbb{P}[C]} = \frac{\mathbb{P}[C]}{\mathbb{P}[C]} = 1.$$

iii) (similar to) *Homework.*

□

3.2 Bayes's formula

It is often very useful to invert conditional probabilities. It is easily calculated what the probability is that a guessing student gets one question right at multiple choice test. But can we say something about the probability that he was guessing, given that the answer was correct? The present section deals with inversion techniques for conditional probabilities that help to answer this kind of questions.

3.7 Proposition (Formula of the Total Probability)**.** *Let F be an event in Ω and $E_1, E_2, \ldots, E_n$ a family of disjoint events (i.e., $E_i \cap E_j$ for all $i \neq j$) that have positive probability (i.e., $P[E_j] > 0$ for all j) and have as union the whole sample space (i.e., $\bigcup_{j=1}^{n} E_j = \Omega$). Then it holds that*

$$\mathbb{P}[F] = \sum_{j=1}^{n} \mathbb{P}[F \mid E_j] \cdot \mathbb{P}[E_j]. \tag{3.2}$$

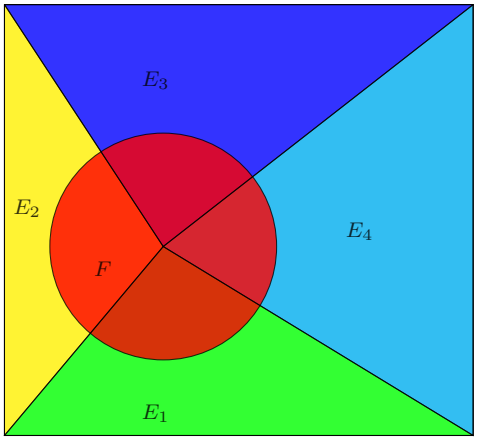


Figure 3.1: Illustration of the Formula of the Total Probability, (3.2).

Proof. By the distributive law (repeatedly applied) we have

$$F = F \cap \Omega = F \cap \left(\bigcup_{j=1}^{n} E_j \right) = \bigcup_{j=1}^{n} (F \cap E_j).$$

As this union is disjoint, we have by the definition of probability, Definition 2.8 iii), and the definition of conditional probability, Definition 3.2,

$$\mathbb{P}[F] = \mathbb{P}\left[\bigcup_{j=1}^{n} (F \cap E_j) \right] = \sum_{j=1}^{n} \mathbb{P}[F \cap E_j] = \sum_{j=1}^{n} \mathbb{P}[F \mid E_j] \cdot \mathbb{P}[E_j].$$

□

3.8 Example. We consider the transmission of signals which are either 1 or 0. We know that one third of all signals are 1s, and the probability that a 1 is wrongly transmitted is $\frac{1}{10}$, whereas the probability that a 0 is wrongly transmitted is $\frac{1}{5}$. How likely is it that a signal is transmitted wrongly?

Solution. We start by modeling the sample space,

$$\Omega = \{(0,0), (0,1), (1,0), (1,1)\}.$$

Here the first component of every pair (i, j) denotes the signal sent and the second component the signal received. E.g., $(1, 0)$ means that a 1 was sent,

but a 0 received. Next we define the events relevant to our problem,

$$\begin{aligned} F &= \text{“faulty transmission”} \\ &= \{(0,1),(1,0)\}, \\ S_0 &= \text{“0 was sent”} \\ &= \{(0,0),(0,1)\}, \\ S_1 &= \text{“1 was sent”} \\ &= \{(1,0),(1,1)\}. \end{aligned}$$

and note

$$S_0 \cup S_1 = \Omega \qquad \text{and} \qquad S_0 \cap S_1 = \emptyset.$$

Moreover, we know already that

$$\mathbb{P}[S_0] = \frac{2}{3}, \qquad \mathbb{P}[S_1] = \frac{1}{3}, \qquad \mathbb{P}[F \,|\, S_0] = \frac{1}{5}, \qquad \mathbb{P}[F \,|\, S_1] = \frac{1}{10}.$$

Using the formula of the total probability (3.2), we can calculate from this the probability of an error in the transmission,

$$\begin{aligned} \mathbb{P}[F] &= \mathbb{P}[F \,|\, S_0] \cdot \mathbb{P}[S_0] + \mathbb{P}[F \,|\, S_1] \cdot \mathbb{P}[S_1] \\ &= \frac{1}{5} \cdot \frac{2}{3} + \frac{1}{10} \cdot \frac{1}{3} = \frac{5}{30} = \frac{1}{6} \approx 16.67\%. \end{aligned}$$

□

3.9 Theorem (Bayes's Formula)**.** *Let F be an event in a sample space Ω and E_1, E_2, …, E_n a family of disjoint events (i.e., $E_i \cap E_j$ for all $i \neq j$) such that $\bigcup_{j=1}^{n} E_j = \Omega$ and $\mathbb{P}[E_j] > 0$ for every event E_j as well as $\mathbb{P}[F] > 0$. Then it holds for every event E_k that*

$$\mathbb{P}[E_k \,|\, F] = \frac{\mathbb{P}[F \,|\, E_k] \cdot \mathbb{P}[E_k]}{\sum_{j=1}^{n} \mathbb{P}[F \,|\, E_j] \cdot \mathbb{P}[E_j]}. \tag{3.3}$$

Proof. Note that we have from the definition of the conditional probability, Definition 3.2,

$$\mathbb{P}[E_k \,|\, F] \cdot \mathbb{P}[F] = \mathbb{P}[E_k \cap F] = \mathbb{P}[F \,|\, E_k] \cdot \mathbb{P}[E_k],$$

and by the formula of the total probability (3.2)

$$\mathbb{P}[F] = \sum_{j=1}^{n} \mathbb{P}[F \,|\, E_j] \cdot \mathbb{P}[E_j].$$

Putting these two results together yields

$$\mathbb{P}[E_k \,|\, F] = \frac{\mathbb{P}[E_k \cap F]}{\mathbb{P}[F]} = \frac{\mathbb{P}[F \,|\, E_k] \cdot \mathbb{P}[E_k]}{\sum_{j=1}^{n} \mathbb{P}[F \,|\, E_j] \cdot \mathbb{P}[E_j]}.$$

□

3.10 Example (Continuation of Example 3.8)**.** In the setting of the previous example, we ask if a 1 was eventually received, what is the probability that a 1 was indeed sent?

Solution. In addition to the events defined in Example 3.8, we define also

$$\begin{aligned} R_1 &= \text{“1 was received”} \\ &= \{(0,1),(1,1)\}. \end{aligned}$$

Our goal is to calculate $\mathbb{P}[S_1 \,|\, R_1]$ which by Bayes's Formula (3.3) is

$$\mathbb{P}[S_1 \,|\, R_1] = \frac{\mathbb{P}[R_1 \,|\, S_1] \cdot \mathbb{P}[S_1]}{\mathbb{P}[R_1 \,|\, S_0] \cdot \mathbb{P}[S_0] + \mathbb{P}[R_1 \,|\, S_1] \cdot \mathbb{P}[S_1]}.$$

Two of the four terms in the fraction are well known,

$$\mathbb{P}[S_1] = \frac{1}{3} \qquad \text{and} \qquad \mathbb{P}[S_0] = \frac{2}{3}.$$

The other two can easily calculated by observing that by Proposition 3.6

$$\begin{aligned} \mathbb{P}[R_1 \,|\, S_1] &= \mathbb{P}[F^c \,|\, S_1] = 1 - \mathbb{P}[F \,|S_1] = 1 - \frac{1}{10} = \frac{9}{10}, \\ \mathbb{P}[R_1 \,|\, S_0] &= \mathbb{P}[F \,|\, S_0] = \frac{1}{5}, \end{aligned}$$

whence

$$\mathbb{P}[S_1 \,|\, R_1] = \frac{\frac{9}{10} \cdot \frac{1}{3}}{\frac{1}{5} \cdot \frac{2}{3} + \frac{9}{10} \cdot \frac{1}{3}} = \frac{\frac{9}{30}}{\frac{13}{30}} = \frac{9}{13} \approx 69.23\%.$$

□

3.11 Example. A student is answering a multiple choice test with m possible answers, exactly one of them correct. He knows the correct answer with probability p. If the student answers a question correctly, what is the probability that he really knew the answer?

Solution. Let us denote the events

C ... the answer to the question is correct,
K ... the student knew the answer.

If he knows the answer, of course he gives the right answer and we have $\mathbb{P}[C \mid K] = 1$. If he guesses the answer, its chances to get the right answer is $\mathbb{P}[C \mid K^c] = \frac{1}{m}$. And, of course by the problem, $\mathbb{P}[K] = p$. Thus we get by Bayes's formula (3.3)

$$\begin{aligned}\mathbb{P}[K \mid C] &= \frac{\mathbb{P}[C \mid K] \cdot \mathbb{P}[K]}{\mathbb{P}[C \mid K] \cdot \mathbb{P}[K] + \mathbb{P}[C \mid K^c] \cdot \mathbb{P}[K^c]} \\ &= \frac{1 \cdot p}{1 \cdot p + \frac{1}{m} \cdot (1-p)} = \frac{mp}{mp - p + 1}.\end{aligned}$$

To illustrate this solution numerically, we look at the case that the multiple choice question has four answers, i.e. $m = 4$. In that case we get that for $p = 20\%$, 50% and 80% the probability that the student knew indeed the question is 50%, 80% and $\frac{16}{17} \approx 94.12\%$ respectively. □

3.3 Independence

Intuitively, an event E should be called independent of the event F, if the probability of E does not changed if we assume F, thus mathematically

$$\mathbb{P}[E \mid F] = \mathbb{P}[E]. \tag{3.4}$$

The notion of independence will be later hugely important, in Chapters 6 and 7.

To avoid the hassle with division by zero, let's reformulate this. As

$$\mathbb{P}[E \mid F] = \frac{\mathbb{P}[E \cap F]}{\mathbb{P}[F]}$$

by Definition 3.2, condition (3.4) can be written as

$$\mathbb{P}[E \cap F] = \mathbb{P}[E] \cdot \mathbb{P}[F]$$

(and here we may allow events to have probability zero). In particular we see that this expression is symmetric, hence if E is independent of F, then automatically F is also independent of E. Lets turn this into a proper mathematical definition.

3.12 Definition. Two elements E, F in a sample space Ω are called *independent* if

$$\mathbb{P}\big[E \cap F\big] = \mathbb{P}[E] \cdot \mathbb{P}[F].$$

We note that this is a purely mathematical definition and does not say anything about causality.

3.13 Example. We are rolling two dice and consider the following events.

A	...	the sum of both dice 6,
B	...	the sum of both dice 7,
C	...	the first die shows 4.

Which two of them are independent?

Solution. We note first that

$$\begin{aligned}
\mathbb{P}[A] &= \mathbb{P}\big[\{(1,5),(2,4),(3,3),(4,2),(5,1)\}\big] = \frac{5}{36},\\
\mathbb{P}[B] &= \mathbb{P}\big[\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}\big] = \frac{1}{6},\\
\mathbb{P}[C] &= \mathbb{P}\big[\{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\}\big] = \frac{1}{6}.
\end{aligned}$$

Now we check the three different cases:

$$\begin{aligned}
\mathbb{P}[A \cap B] &= \mathbb{P}[\emptyset] = 0 < \frac{5}{216} = \frac{5}{36} \cdot \frac{1}{6} = \mathbb{P}[A] \cdot \mathbb{P}[B],\\
\mathbb{P}[A \cap C] &= \mathbb{P}\big[\{(4,2)\}\big] = \frac{1}{36} = \frac{6}{216} > \frac{5}{216} = \frac{5}{36} \cdot \frac{1}{6} = \mathbb{P}[A] \cdot \mathbb{P}[C],\\
\mathbb{P}[B \cap C] &= \mathbb{P}\big[\{(4,3)\}\big] = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \mathbb{P}[B] \cdot \mathbb{P}[C].
\end{aligned}$$

Thus B and C are independent, while A and B as well as A and C are not. □

Next we intend to extend the notion of independence to more then three vents. Looking at for event we intend to require all of the following equalities

i)

$$\mathbb{P}[E \cap F \cap G] = \mathbb{P}[E] \cdot \mathbb{P}[F] \cdot \mathbb{P}[G]; \tag{3.5}$$

ii)

$$\begin{cases}
\mathbb{P}[E \cap F] &= \mathbb{P}[E] \cdot \mathbb{P}[F],\\
\mathbb{P}[E \cap G] &= \mathbb{P}[E] \cdot \mathbb{P}[G],\\
\mathbb{P}[F \cap G] &= \mathbb{P}[F] \cdot \mathbb{P}[G].
\end{cases} \tag{3.6}$$

In the case that ii) is satisfied (but not necessarily i)), we will say that E, F and G are *pairwise independent*. For more than three events we have the following general definition.

3.14 Definition. The events E_1, E_2, ..., E_n are called independent, if for every $1 \leq j_1 < \cdots < j_r \leq n$, $1 \leq r \leq n$, it holds that

$$P\big[E_{j_1} \cap E_{j_2} \cap \cdots \cap E_{j_r}\big] = \mathbb{P}\big[E_{j_1}\big] \cdot \ldots \cdot \mathbb{P}\big[E_{j_r}\big],$$

or, compactly written,

$$\mathbb{P}\Big[\bigcap_{k=1}^{r} E_{j_k}\Big] = \prod_{k=1}^{r} \mathbb{P}[E_{j_k}].$$

3.15 Example (Counterexamples for Independence Relations)**.** Concentrating on the case of three events, what are example where only (3.5) is satisfied and not (3.6), resp. only (3.6) and not (3.6)?

Solution.

- Let's give first an example where i) is satisfied, but not ii). Let A be an event in a sample space Ω such that $0 < \mathbb{P}[A] < 1$ and define

  $$E := A, \qquad F := A, \qquad G = \emptyset.$$

  Then we have clearly

  $$E \cap F = A, \qquad E \cap G = F \cap G = E \cap F \cap G = \emptyset,$$

  and it follows that

  $$\mathbb{P}[E \cap F \cap G] = \mathbb{P}[\emptyset] = 0 = \mathbb{P}[A] \cdot \mathbb{P}[A] \cdot \mathbb{P}[\emptyset] = \mathbb{P}[E] \cdot \mathbb{P}[F] \cdot \mathbb{P}[G].$$

  However, we note that

  $$\mathbb{P}[E \cap F] = \mathbb{P}[A] \qquad \text{while} \qquad \mathbb{P}[E] \cdot \mathbb{P}[F] = \mathbb{P}[A]^2.$$

  As for $a \in \mathbb{R}$ it holds that $a^2 = a$ implies $a = 0$ or $a = 1$, it follows that

  $$\mathbb{P}[E \cap F] \neq \mathbb{P}[E] \cdot \mathbb{P}[F].$$

  since we asssumed $0 < \mathbb{P}[A] < 1$.

Note that by this argument an event is only independent of itself if it has probability zero or one.

- To give an example where ii) is satisfied, but not i), we consider the sample space of two consecutive (fair) coin flips,

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}.$$

We define the events

E	...	coin shows head on first flip,
F	...	coin shows head on second flip,
G	...	coin shows head exactly at one of the flips.

Now we have clearly

$$\begin{aligned}
\mathbb{P}[E\cap F] &= \mathbb{P}\big[\{(H,H)\}\big] = \frac{1}{4} = \frac{1}{2}\cdot\frac{1}{2} \\
&= \mathbb{P}\big[\{(H,H),(H,T)\}\big]\cdot\mathbb{P}\big[\{(H,H),(T,H)\}\big] = \mathbb{P}[E]\cdot\mathbb{P}[F] \\
\mathbb{P}[E\cap G] &= \mathbb{P}\big[\{(H,T)\}\big] = \frac{1}{4} = \frac{1}{2}\cdot\frac{1}{2} \\
&= \mathbb{P}\big[\{(H,H),(H,T)\}\big]\cdot\mathbb{P}\big[\{(H,T),(T,H)\}\big] = \mathbb{P}[E]\cdot\mathbb{P}[G] \\
\mathbb{P}[F\cap G] &= \mathbb{P}\big[\{(T,H)\}\big] = \frac{1}{4} = \frac{1}{2}\cdot\frac{1}{2} \\
&= \mathbb{P}\big[\{(H,H),(T,H)\}\big]\cdot\mathbb{P}\big[\{(H,T),(T,H)\}\big] = \mathbb{P}[E]\cdot\mathbb{P}[F].
\end{aligned}$$

However, on the other hand side we have

$$\begin{aligned}
&\mathbb{P}[E\cap F\cap G] = \mathbb{P}[\emptyset] = 0 < \frac{1}{8} = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} \\
&= \mathbb{P}\big[\{(H,H),(H,T)\}\big]\cdot\mathbb{P}\big[\{(H,H),(T,H)\}\big]\cdot\mathbb{P}\big[\{(H,T),(T,H)\}\big] \\
&= \mathbb{P}[E]\cdot\mathbb{P}[F]\cdot\mathbb{P}[G].
\end{aligned}$$

□

3.16 Example. An electric network has 10 parallel switches (see Figure 3.2) and the probability that one of them is turned on is 20%, independently of the status of all the other switches. What is the probability that the whole system is working?

Solution. Denote

S_n	...	the n-th switch is turned on,
F	...	the whole electrical network is working.

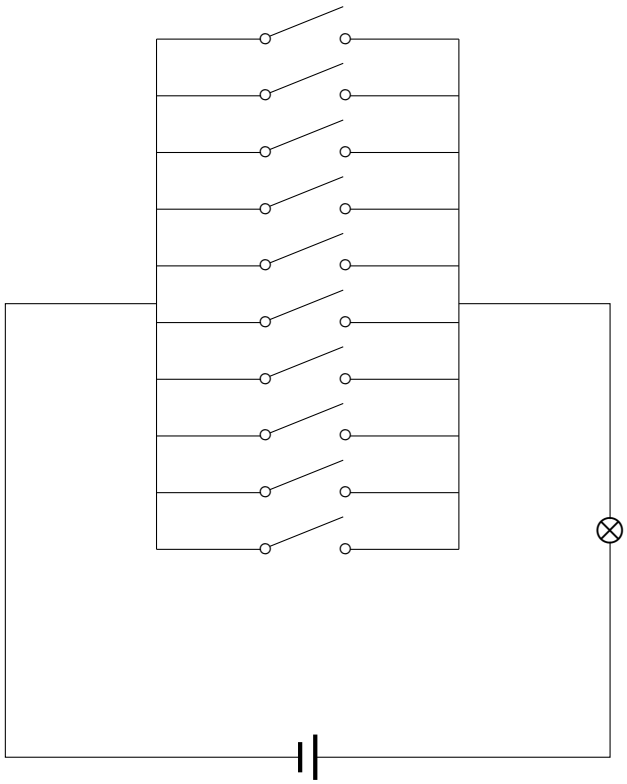


Figure 3.2: Electrical network of Example 3.16.

and make the key observation that the system is only not working if all switches are turned off. Thus

$$\mathbb{P}[F] = 1 - \mathbb{P}[F^c] = 1 - \mathbb{P}\left[\bigcap_{n=1}^{10} S_n^c\right] = 1 - \prod_{n=1}^{10} \mathbb{P}[S_n^c]$$
$$= 1 - \prod_{n=1}^{10} \Big(1 - \mathbb{P}[S_n]\Big) = 1 - \Big(1 - \mathbb{P}[S_1]\Big)^{10} = 1 - 0.8^{10} = 89.26\%.$$

Note that we make here use from the fact that from the independence of events follows also the independence of their complements.

□

Chapter 4

Discrete Random Variables

4.1 Discrete Random Variables and their Distributions

Often we are not so much interested in the outcome of an experiment itself, but only in some number measuring some quantity in the experiment. Therefore it is very convenient to work with the following.

4.1 Definition. A function $X : \Omega \to \mathbb{R}$ is called a *random variable*.

4.2 Example. We are tossing three fair coins. Denote by Y the number of heads. Thus Y can achieve the values 0, 1, 2, and 3. Our Laplacian sample space is

$$\Omega = \{(T,T,T),(H,T,T),(T,H,T),(T,T,H),(H,H,T),(T,H,H),\\ (H,T,H),(H,H,H)\}$$

and has in total 8 elements. We have thus

$$\begin{aligned}
\mathbb{P}[Y=0] &= \mathbb{P}[\{(T,T,T)\}] = \frac{1}{8},\\
\mathbb{P}[Y=1] &= \mathbb{P}[\{(H,T,T),(T,H,T),(T,T,H)\}] = \frac{3}{8},\\
\mathbb{P}[Y=2] &= \mathbb{P}[\{(H,H,T),(T,H,H),(H,T,H)\}] = \frac{3}{8},\\
\mathbb{P}[Y=3] &= \mathbb{P}[\{(H,H,H)\}] = \frac{1}{8}.
\end{aligned}$$

Note that we have used (and will use) the shorthand

$$\mathbb{P}[Y=j] = \mathbb{P}[\{Y=j\}] = \mathbb{P}[\{\omega : Y(\omega)=j\}].$$

It is clear to see how much easier the expression via the random variable Y is, compared to the listing of all events. Of course, the random variable obeys also certain rules, as all probabilities should sum up to one, i.e.,

$$\sum_{j=1}^{3} \mathbb{P}[Y = j] = \mathbb{P}\Big[\bigcup_{j=0}^{3} \{Y = j\}\Big] = \mathbb{P}[\Omega] = 1.$$

We can look at the *probability mass function* of Y,

$$p(i) = \mathbb{P}[Y = i],$$

and graph it, see Figure 4.1. □

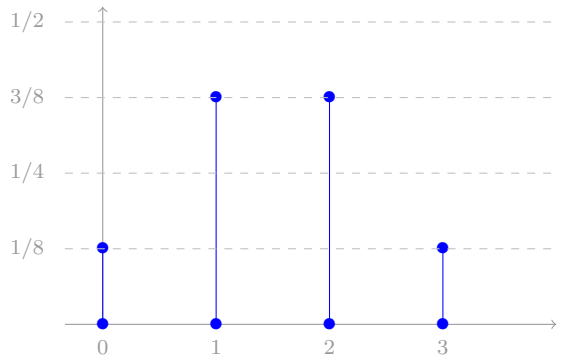


Figure 4.1: Probability mass function of Y, Example 4.2.

4.3 Definition. A random variable X is called *discrete*, if it takes at most countably many values $x_1, x_2, \ldots, x_j, \ldots$. Its *probability mass function* (abbreviated *pmf*) $p(x) = p_X(x) = \mathbb{P}[X = x]$ thus satisfies

If we have more then one random mass function, we write p_X instead of p to avoid confusion.

$$\begin{aligned} p(x_j) &\geq 0, \\ p(x) &= 0 \text{ for all } x \neq x_j, \\ \sum_{k=1}^{\infty} p(x_k) &= \sum_{k=1}^{\infty} \mathbb{P}[X = x_k] = \mathbb{P}\Big[\bigcup_{k=1}^{\infty} \{X = x_k\}\Big] = \mathbb{P}[\Omega] = 1. \end{aligned}$$

4.4 Example. Assume that the probability mass function of the random variable X is given by

$$p(j) = c \cdot \frac{\lambda^j}{j!}, \qquad j \in \mathbb{N},$$

for some positive constant λ.

a) What is c?

b) Find $\mathbb{P}[X = 0]$ and $\mathbb{P}[X > 2]$.

Solution.

a) As all the probabilities have to sum up to one, we conclude from

$$1 \stackrel{!}{=} \sum_{j=0}^{\infty} p(j) = c \cdot \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = c \cdot e^{\lambda},$$

that

$$c = e^{-\lambda}.$$

b) We find that

$$\begin{aligned}
\mathbb{P}[X = 0] &= e^{-\lambda} \frac{\lambda^0}{0!} = e^{-\lambda}, \\
\mathbb{P}[X > 2] &= 1 - \mathbb{P}[X = 0] - \mathbb{P}[X = 1] - \mathbb{P}[X = 2] \\
&= 1 - e^{-\lambda} - e^{-\lambda} \cdot \lambda - e^{-\lambda} \cdot \frac{\lambda^2}{2} = 1 - e^{-\lambda}\Big(1 + \lambda + \frac{\lambda^2}{2}\Big).
\end{aligned}$$

□

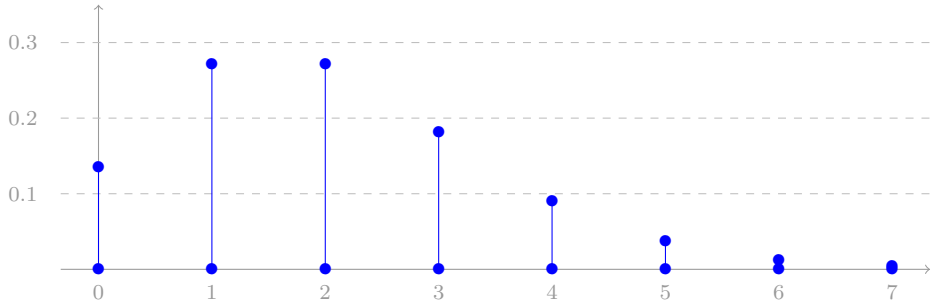


Figure 4.2: Probability mass function of X with parameter $\lambda = 2$, Example 4.4.

4.2 Expectation & Variance of Discrete Random Variables

The idea behind the expectation is to calculate the "average" outcome of a random variable, where the possible outcomes are weighted by their likelihood.

4.5 Definition. Let X be a discrete random variable with probability mass function p. Then the *expected value* (or the *expectation* or the *mean*) of X taking values $x_1, x_2, \ldots$ is defined to be

$$\mathbb{E}[X] = \sum_k x_k \cdot p(x_k).$$

When X takes infinitely many values, the series does not necessarily converge. For this class we assume that all appearing series converge to some real number.

4.6 Example. Consider the following game: Rolling a die, one wins the ten dollar per pip shown on the top side. What is the expected gain? How much would you pay to participate in this game?

Solution. Let X be the random variable which describes the gains from this game. Then we have

$$\begin{aligned}\mathbb{P}[X = 10] = \mathbb{P}[X = 20] = \mathbb{P}[X = 30] = \mathbb{P}[X = 40] \\ = \mathbb{P}[X = 50] = \mathbb{P}[X = 60] = \frac{1}{6},\end{aligned}$$

and thus

$$\begin{aligned}\mathbb{E}[X] &= 10 \cdot \frac{1}{6} + 20 \cdot \frac{1}{6} + 30 \cdot \frac{1}{6} + 40 \cdot \frac{1}{6} + 50 \cdot \frac{1}{6} + 60 \cdot \frac{1}{6} \\ &= (1 + 2 + 3 + 4 + 5 + 6)\frac{10}{6} = 21 \cdot \frac{5}{3} = 35.\end{aligned}$$

Thus if you can enter the game for less than \$ 35, it is on average profitable to participate in the game. □

4.7 Example. Calculate the expected value of the random variable described in Example 4.4.

Solution. We have

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

□

How can we calculate the expected value of some function of X, say $g(X)$ for some $g : \mathbb{R} \to \mathbb{R}$? Note that $g(X)$ is itself a random variable as $X : \omega \to \mathbb{R}$ and thus $g \circ X : \Omega \to \mathbb{R}$

4.8 Example. Let X be a random variable with

$$\mathbb{P}[X = -1] = 0.2, \qquad \mathbb{P}[X = 0] = 0.5, \qquad \mathbb{P}[X = 1] = 0.3.$$

What is $\mathbb{E}[X^2]$?

Solution. We define a new random variable $Y := X^2$ for which we have

$$\mathbb{P}[Y = 1] = \mathbb{P}[X^2 = 1] = \mathbb{P}[X = -1] + \mathbb{P}[X = 1] = 0.2 + 0.3 = 0.5,$$
$$\mathbb{P}[Y = 0] = \mathbb{P}[X^2 = 0] = \mathbb{P}[X = 0] = 0.5,$$

and thus

$$\mathbb{E}[X^2] = \mathbb{E}[Y] = 1 \cdot 0.5 + 0 \cdot 0.5 = 0.5.$$

□

In general we have the following proposition.

4.9 Proposition. *Assume that X is a discrete random with probability mass distribution p with $p(x_i) > 0$ for $x_1, x_2, \ldots$ and $g : \mathbb{R} \to \mathbb{R}$ a real-valued function. Then it holds that*

$$\mathbb{E}[g(X)] = \sum_i g(x_i) \cdot \mathbb{P}[X = x_i].$$

Proof. We note first that the difficulty hails from the fact that $g(x_i) = g(x_j)$ for $i \neq j$ is possible and thus $g(X)$ may take less values than X (e.g., when X takes the values -1 and 1 and $g(x) = x^2$ as in the example above). Denote all these possible values of $g(X)$ by $y_1, y_2, \ldots$ and note that in order to calculate $\mathbb{E}[g(X)]$, we have effectively to sum up about all possible values of $g(X)$, hence the y_j weighted with their respective probability. Moreover, we note that we have

$$\mathbb{P}[g(X) = y_j] = \sum_{x_i \,:\, g(x_i) = y_j} \mathbb{P}[X = x_i],$$

as we sum up on the right hand side just the probabilities of different cases which lead to the same event $\{g(X) = y_j\}$. Using this we may conclude that

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_j y_j \cdot \mathbb{P}[g(X) = y_j] = \sum_j y_j \cdot \sum_{x_i \,:\, g(x_i)=y_j} \mathbb{P}[X = x_i] \\ &= \sum_j \sum_{x_i \,:\, g(x_i)=y_j} y_j \cdot \mathbb{P}[X = x_i] = \sum_j \sum_{x_i \,:\, g(x_i)=y_j} g(x_i) \cdot \mathbb{P}[X = x_i] \\ &= \sum_i g(x_i) \cdot \mathbb{P}[X = x_i],\end{aligned}$$

by observing that in the last step we just sum up over all x_i. □

4.10 Corollary. *For a random variable X it holds that for any a, $b \in \mathbb{R}$*

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b. \tag{4.1}$$

This means that the expectation is linear*.*

Proof. By Proposition 4.9 we have

$$\begin{aligned}\mathbb{E}[aX + b] &= \sum_j (a \cdot x_j + b) \cdot \mathbb{P}[X = x_j] \\ &= a \cdot \underbrace{\sum_j x_j \mathbb{P}[X = x_j]}_{=\mathbb{E}[X]} + b \cdot \underbrace{\sum_j \mathbb{P}[X = x_j]}_{=1} = a\,\mathbb{E}[X] + b.\end{aligned}$$

□

The expected value of a random variable does not necessarily give much insight in the behavior of the random variable, as the following example illustrates.

4.11 Example. Let W, X and Y three random variables given by

$$\begin{aligned}&\mathbb{P}[W = 0] = 1, \\ &\mathbb{P}[X = 1] = \mathbb{P}[X = -1] = \frac{1}{2}, \\ &\mathbb{P}[Y = 1,000] = \mathbb{P}[Y = -1,000] = \frac{1}{2}.\end{aligned}$$

At least I would be vary of entering a game in which I could win or lose with equal likelihood $ 1,000...

It is straightforward to see that

$$\mathbb{E}[W] = \mathbb{E}[X] = \mathbb{E}[Y] = 0.$$

However, their actual behavior differs vastly. □

To measure the difference of the three random variables, one should try to measure the distance of the actual outcomes from the mean. One could just go to the expected value of the distance and calculate $\mathbb{E}\big[|X - \mathbb{E}[X]|\big]$, but this has the drawback that it contains the mathematically unpleasant modulus function $|\cdot|$ (it's even not differentiable!). Thus one settles for the quadratic distance from the expected value and calls it the *variance*.

4.12 Definition. The variance of a random variable X is defined as the expected quadratic variance from the expectation,

$$\mathbb{V}\mathrm{ar}[X] := \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big].$$

4.13 Example (Continuation of Example 4.11)**.** For the example given above we calculate

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[W] &= \mathbb{E}\Big[\big(W - \mathbb{E}[W]\big)^2\Big] = \mathbb{E}[W^2] = 0^2 \cdot 1 = 0,\\
\mathbb{V}\mathrm{ar}[X] &= \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] = \mathbb{E}[X^2] = 1^2 \cdot \frac{1}{2} + (-1)^2 \cdot \frac{1}{2} = 1,\\
\mathbb{V}\mathrm{ar}[Y] &= \mathbb{E}\Big[\big(Y - \mathbb{E}[Y]\big)^2\Big] = \mathbb{E}[Y^2] = 1,000^2 \cdot \frac{1}{2} + (-1,000)^2 \cdot \frac{1}{2}.\\
&= 1,000,000.
\end{aligned}$$

□

To calculate the variance actually, it is often easier to use the formula given by the following proposition instead of the definition.

Note we write short $\mathbb{E}[X]^2$ for $\big(\mathbb{E}[X]\big)^2$

4.14 Proposition. *It holds that*

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{4.2}$$

Proof. As the expected value is itself a real number, we set $\mu = \mathbb{E}[X]$ and then get by using Corollary 4.10

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X] \;&=\; \mathbb{E}\Big[\big(X - \mu\big)^2\Big] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - \mathbb{E}[2\mu X] + \mathbb{E}[\mu^2]\\
&\overset{\text{Cor. 4.10}}{=} \mathbb{E}[X^2] - 2\mu\,\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2.
\end{aligned}$$

□

Similarly we can prove the following proposition.

4.15 Proposition. *For real numbers a, $b \in \mathbb{R}$, we have*

$$\mathbb{V}\mathrm{ar}[aX + b] = a^2\, \mathbb{V}\mathrm{ar}[X].$$

Proof. By the definition of the variance and Corollary 4.10

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[aX + b] &= \mathbb{E}\Big[\big(aX + b - \mathbb{E}[aX + b]\big)^2\Big] \\
&\overset{\text{Cor. 4.10}}{=} \mathbb{E}\Big[\big(aX + b - a\,\mathbb{E}[X] - b\big)^2\Big] = a^2\, \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] \\
&= a^2\, \mathbb{V}\mathrm{ar}[X].
\end{aligned}$$

□

To get a measure of expectation of the mean that does not square multiplicative constants one considers the *standard deviation*,

$$\mathbb{SD}[X] := \sqrt{\mathbb{V}\mathrm{ar}[X]}.$$

4.3 Examples of Discrete Random Variables

We are going to discuss some important classes of discrete random variables.

4.3.1 Bernoulli Random Variables

A random variable is called Bernoulli distributed (or a Bernoulli random variable), if it has the probability mass distribution

$$\mathbb{P}[X = 1] = p, \qquad \mathbb{P}[X = 0] = 1 - p.$$

Thus we can understand it as the probabilities of getting head in a (possibly unfair, if $p \neq \frac{1}{2}$) coin toss. More generally we can think about the Bernoulli variable as an experiment which succeeds with probability p and fails with probability $1 - p$.

We calculate straightforwardly

$$\begin{aligned}
\mathbb{E}[X] &= 1 \cdot p + 0 \cdot (1 - p) = p, \\
\mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[X^2] - E[X]^2 = \Big(1^2 \cdot p + 0^2 \cdot (1 - p)\Big) - p^2 \\
&= p - p^2 = p \cdot (1 - p).
\end{aligned}$$

4.3.2 Binomial Random Variables

A binomially distributed random variable describes that in n independent repetitions of the same experiment with success probability p one has exact k successes. As we can have exactly $\binom{n}{k}$ possibilities how the k successes can be arranged, we have thus

$$\mathbb{P}[X = k] = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}, \qquad \text{for all } k \in \{0, \dots n\}.$$

We write usually $X \sim \mathcal{B}(n; p)$ to indicate that X is a binomial variable with n trials and success probability p.

First we check that this is indeed a probability mass function, i.e., that all probabilities sum up to one (the non-negativity being obvious). By the Binomial Theorem, Theorem 1.10 we have

$$\sum_{k=0}^{n} \mathbb{P}[X = k] = \sum_{k=0}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \big(p + (1-p)\big)^n = 1^n = 1.$$

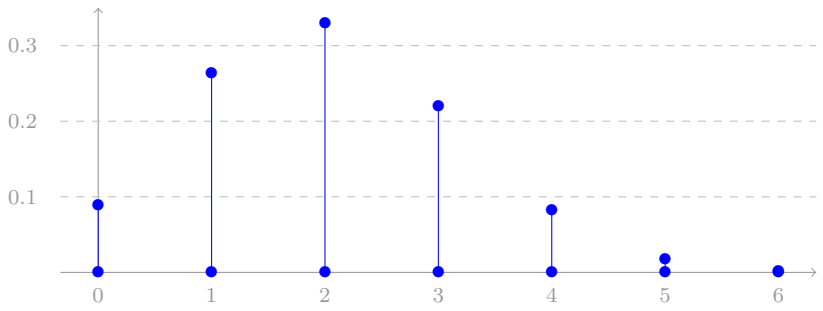


Figure 4.3: Probability mass function of a binomial random variable $X \sim \mathcal{B}(6; \frac{1}{3})$.

4.16 Example. If we roll seven dice, how likely is it to get exactly three 6s?

Solution. As every rolling die is an independent experiment with the same probabilities as the other, we observe that this is described by a binomial distribution. Moreover, to roll a six with one die is exactly $\frac{1}{6}$, thus $p = \frac{1}{6}$. Hence we get

$$P[X = 3] = \binom{7}{3} \cdot \left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^4 = 35 \cdot \frac{5^4}{6^7} \approx 7.81\%.$$

□

Next we calculate expected value and variance of a binomial random variable. For the expectation, we have

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{n} k \cdot \mathbb{P}[X=k] = \sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \\
&= \sum_{k=1}^{n} k \cdot \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k} \\
&= \sum_{k=1}^{n} n \cdot \frac{(n-1)!}{(k-1)! \cdot ((n-1)-(k-1))!} \cdot p^k \cdot (1-p)^{n-k} \\
&= \sum_{k=1}^{n} n \cdot \binom{n-1}{k-1} \cdot p^k \cdot (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} \cdot p^{k-1} \cdot (1-p)^{n-k} \\
&\overset{j=k-1}{=} np \underbrace{\sum_{j=0}^{n-1} \binom{n-1}{j} \cdot p^j \cdot (1-p)^{(n-1)-j}}_{=1} \\
&= np.
\end{aligned}$$

To calculate the variance, we remind ourselves of

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2,$$

and calculate $\mathbb{E}\big[X^2\big]$ by the same means as the expectation above (see homework). We finally end up with

$$\mathbb{V}\mathrm{ar}[X] = np(1-p).$$

4.3.3 Geometric Random Variables

A geometrically distributed random variable X describes the number of failures one has in an experiment with success probability p, $0 < p < 1$, before getting the first success. We write $X \sim Geom(p)$. Thus the probability mass distribution is described as

$$\mathbb{P}[X=k] = (1-p)^k p, \qquad \text{for } k \in \mathbb{N}.$$

Some people and books use a slightly different definition of a geometric random variable: k is not the number of unsuccessful trials, but the total number of trials.

One sees quickly that this is indeed a random variable as

$$\sum_{k=0}^{\infty} \mathbb{P}[X=k] = p\sum_{k=0}^{\infty}(1-p)^k = p\cdot\frac{1}{1-(1-p)} = 1,$$

by the summation formula for the geometric series.

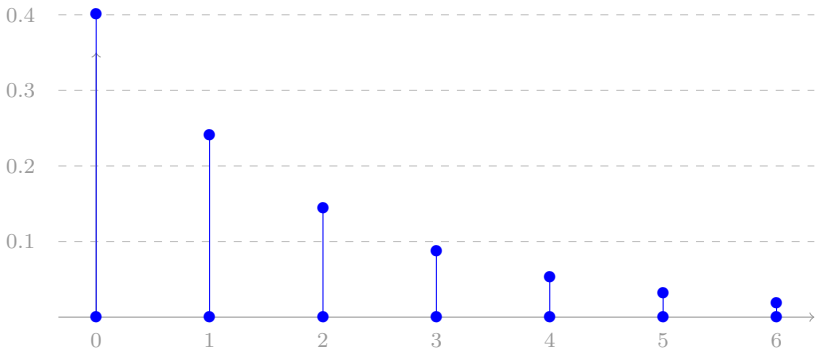


Figure 4.4: Probability mass function of a geometric random variable $X \sim Geom(0.4)$.

4.17 Example. Rolling one fair die, how likely is it that the first 6 appears on the fourth throw?

Solution. We have three failures before the success, thus

$$\mathbb{P}[X=3] = \left(1-\frac{1}{6}\right)^3 \cdot \frac{1}{6} = \frac{5^3}{6^4} \approx 9.65\%.$$

□

To calculate the expected value of a geometric random variable, we note that

$$\begin{aligned}
\mathbb{E}[X] &= p\sum_{k=0}^{\infty} k\cdot(1-p)^k = p\sum_{k=1}^{\infty}((k-1)+1)\cdot(1-p)^k \\
&= p\sum_{k=1}^{\infty}(k-1)\cdot(1-p)^k + p\sum_{k=1}^{\infty}(1-p)^k \\
&\overset{j=k-1}{=} (1-p)p\underbrace{\sum_{j=0}^{\infty} j\cdot(1-p)^j}_{=\mathbb{E}[X]} + p\Bigg(\underbrace{\sum_{k=0}^{\infty}(1-p)^k}_{=\frac{1}{p}} - 1\Bigg) \\
&= (1-p)\,\mathbb{E}[X] + (1-p).
\end{aligned}$$

Solving this equation for $\mathbb{E}[X]$ yields

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

Using the same method for the calculation of $\mathbb{E}[X^2]$, one can also get the variance. This yields

$$\mathbb{V}\mathrm{ar}(X) = \frac{1-p}{p^2}.$$

4.3.4 Poisson Random Variables

A random variable X is Poisson distributed with parameter $\lambda > 0$, shorthand $X \sim Poi(\lambda)$ if the probability mass function is given for

$$P[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad \text{for } k \in \mathbb{N}.$$

We recognize that this distribution appeared already, in Example 4.4 we showed that it is indeed a probability distribution and in Example 4.7 we showed that it has mean λ. In the same way one can also show that $\mathbb{V}\mathrm{ar}[X] = \lambda$. An illustration was given in Figure 4.2.

One can show (and we will show in Section 7.4) that the Poisson distribution is a good and simple approximation for the binomial distribution $\mathcal{B}(n;p)$ if n is large, p is small and $\lambda = n \cdot p$. It is often used to describe events that happen in a given (time) interval.

- Number of shoppers in a warehouse in a given time period
- Number of atoms of a radioactive material that disintegrate per hour
- Number of typographical errors in a book per page
- ...

4.18 Example. Assume that number of typographical errors in a book per page is Poisson distributed with parameter $\lambda = \frac{1}{2}$. How likely is it that there is at least one error on a randomly chosen page?

Solution. We have

$$\mathbb{P}[X \geq 1] = 1 - \mathbb{P}[X = 0] = 1 - e^{-\frac{1}{2}} \cdot \frac{\left(\frac{1}{2}\right)^0}{0!} = 1 - e^{-\frac{1}{2}} \approx 39.35\%.$$

□

4.4 The Cumulative Distribution Function

Another way to look at a random variable is to observe its *cumulative distribution function* (often abbreviated as *cdf*) defined by

$$F(x) = \sum_{\substack{y \le x \\ p(y)>0}} p(x) = \mathbb{P}\Big[\bigcup_{y \le x} \{X = y\}\Big] = \mathbb{P}[X \le x].$$

Again we will write F_X instead of F in the case that we have cdfs for different random variables.

For instance, for the random variable from Example 4.2 we have

$$F(x) = \begin{cases} 0 & \text{if } x < 0; \\ \frac{1}{8} & \text{if } 0 \le x < 1; \\ \frac{1}{2} & \text{if } 1 \le x < 2; \\ \frac{7}{8} & \text{if } 2 \le x < 3; \\ 1 & \text{if } x \ge 3. \end{cases}$$

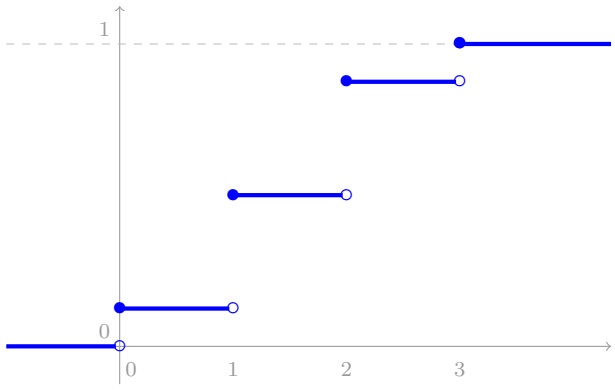


Figure 4.5: Cumulative distribution function of Y, Example 4.2.

4.19 Proposition (Properties of the cdf)**.** *If F ia a cumulative distribution function, then it holds that*

i) F is non-decreasing on $\mathbb{R}$,

ii) F is right-continuous, i.e., $\lim_{x \downarrow y} F(x) = F(y)$ for all $y \in \mathbb{R}$,

iii) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

Proof. No proof is given in this class. The interested reader is referred to Ross[1]. □

[1]Sheldon Ross. *A First Course in Probability*. 9th edition. Pearson, 2012. ISBN: 978-0-321-79477-2, Section 4.10.

Chapter 5

Continuous Random Variables

5.1 Continuous Random Variables and their Distributions

5.1 Example. A bus arrives every 10 minutes at a bus station. If you arrive (not knowing the bus schedule) at a random time at at the station, how long you will have to wait?

Solution. Let's model the waiting time by the random variable X. As you may arrive at every point between 0 and 10 minutes after the last bus left, the random variable may take as value every real number between 0 and 10. Thus there will be no way to write a probability mass function for this random variable. Thanks to the notion of the cumulative distribution function, the value $\mathbb{P}[X \leq x]$ is easily calculable, just $\frac{x}{10}$ for every x between 0 and 10. More precisely, we have

$$F(x) = \mathbb{P}[X \leq x] = \begin{cases} 0 & \text{if } x < 0; \\ \frac{x}{10} & \text{if } 0 \leq x < 10; \\ 1 & \text{if } x \geq 10. \end{cases}$$

□

Thus even when the random variable is not discrete, we can find a cumulative distribution function. But how to replace the probability mass function that was so handy for actual computations, e.g., of expectation

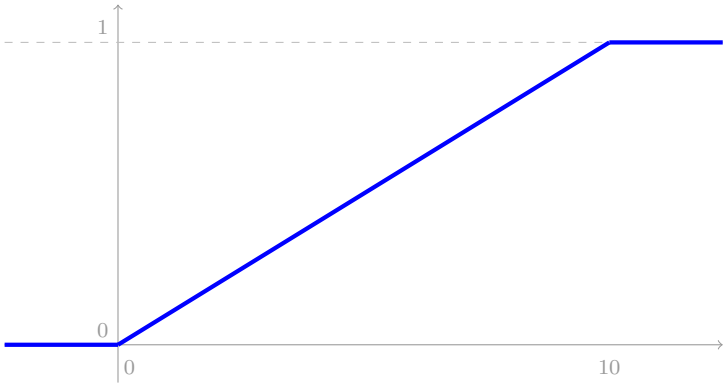


Figure 5.1: Cumulative distribution function of X, Example 5.1.

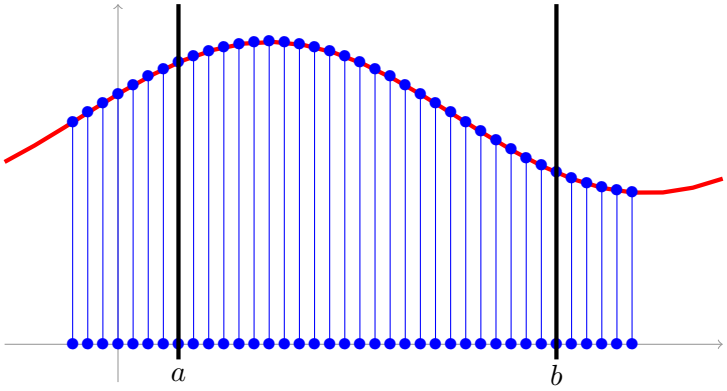


Figure 5.2: Calculating $\mathbb{P}[a < X \leq b]$ in the discrete and continuous case.

and variance? To see this, we can think what is happening if the probability mass distribution is very dense, as illustrated in Figure 5.2.

It looks like all the information is encoded in the function on which the probability mass points lie. This function seems to be the perfect analogue in the continuous setting that can replace the probability mass function.Thus, if we want to calculate the probability that X lies between a and b, we have in the discrete case

$$\mathbb{P}[a < X \leq b] = \sum_{a < x_j \leq b} p(x_j).$$

Note that $\mathbb{P}[a \leq X \leq b]$ is equal to $P[a < X < b]$ by the continuity of the integral.

In the continuous case (sums are becoming integrals), this should be now

$$\mathbb{P}[a < X \leq b] = \int_a^b f(x)\, dx.$$

For the cumulative distribution function, we would get then

$$F(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} f(y)\, dy.$$

This indicates (as F should be non-decreasing and tending to 1 as x goes to infinity, cf. Proposition 4.19) that we should have $f(x) \geq 0$ for all real x and $\int_{-\infty}^{\infty} f(x)\, dx = 1$. We will call f the *density* of X and call a random variable that admits a density *continuous*. We make this as basis for the following formal definition.

Again we will write f_X instead of f in the case that we have pdfs for different random variables.

5.2 Definition. Let $f : \mathbb{R} \to \mathbb{R}$ be a real-valued function such that $f(x) \geq 0$ and for all x and $\int_{-\infty}^{\infty} f(x)\, dx = 1$. We call X a *continuous* random variable with *probability density function* (or *pdf*) f if

$$\mathbb{P}[X \leq x] = F(x) = \int_{-\infty}^{x} f(y)\, dy.$$

Clearly, it follows from the definition that we have for any bounded interval $(a, b]$ that

Note that there are not only continuous and discrete random variables, these are just the two cases where computation is quite easily possible.

$$\begin{aligned}\mathbb{P}\big[X \in (a,b]\big] &= \mathbb{P}[a < X \leq b] = \mathbb{P}[X \leq b] - \mathbb{P}[X \leq a] = F(b) - F(a) \\ &= \int_{-\infty}^{b} f(y)\, dy - \int_{-\infty}^{a} f(y)\, dy = \int_{a}^{b} f(y)\, dy.\end{aligned}$$

5.3 Example. Let f be a real valued function given by

$$f(x) = \begin{cases} c \cdot (4x - 2x^2) & \text{if } 0 < x \leq 2; \\ 0 & \text{otherwise.} \end{cases}$$

for a real constant c.

a) What is c?

b) Calculate $\mathbb{P}[X > 1]$.

c) What is the cdf of X.

Solution.

a) As the density has to integrate up to 1, we have

$$\begin{aligned}1 \overset{!}{=} \int_{-\infty}^{\infty} f(x)\, dx &= \int_{0}^{2} c \cdot (4x - 2x^2)\, dx = c \cdot \left[2x^2 - \frac{2}{3}x^3\right]_0^2 \\ &= c \cdot \left(8 - \frac{16}{3}\right) = \frac{8}{3}\end{aligned}$$

and thus $c = \frac{3}{8}$.

b) We have

$$\mathbb{P}[X > 1] = \int_1^\infty f(x)\,dx = \int_1^2 \frac{3}{8} \cdot (4x - 2x^2)\,dx = \frac{3}{8} \cdot \left[2x^2 - \frac{2}{3}x^3\right]_1^2$$
$$= \frac{3}{8} \cdot \left(\frac{8}{3} - \frac{4}{3}\right) = \frac{1}{2}.$$

c) And here in the same way

$$F(x) = \mathbb{P}[X \le x] = \int_{-\infty}^x f(y)\,dy = \left\{ \begin{array}{l} 0 \\ \int_0^x \frac{3}{8} \cdot (4y - 2y^2)\,dy \\ \int_0^2 \frac{3}{8} \cdot (4y - 2y^2)\,dy \end{array} \right\}$$
$$= \left\{ \begin{array}{l} 0 \\ \frac{3}{8} \cdot \left[2y^2 - \frac{2}{3}y^3\right]_0^x \\ \frac{3}{8} \cdot \left[(2y^2 - \frac{2}{3}y^3)\right]_0^2 \end{array} \right\} = \left\{ \begin{array}{l} 0 \\ \frac{3}{8} \cdot \left(2x^2 - \frac{2}{3}x^3\right) \\ 1 \end{array} \right\}$$
$$= \left\{ \begin{array}{ll} 0 & \text{if } x \le 0; \\ \frac{3}{4}x^2 - \frac{1}{4}x^3 & \text{if } 0 < x \le 2; \\ 1 & \text{if } 0 < x \le 2. \end{array} \right.$$

□

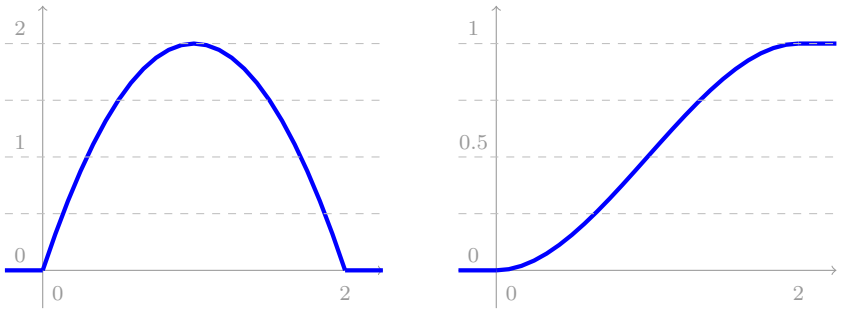


Figure 5.3: Density f and cumulative distribution function F of Example 5.3.

5.2 Expectation & Variance of Continuous Random Variables

To define the expected value of a continuous random variable, we are motivated again by the discrete case where $\mathbb{E}[X] = \sum_{x_j} x_j \cdot p(x_j)$. Replacing the probability mass distribution by the density and the sum by an integral yields the following definition.

5.4 Definition. Let X be a continuous random variable with probability density function f. Then the *expected value* (or the *expectation* or the *mean*) of X is defined to be

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

Of course the integral does not converge necessarily. For this class, we assume that all appearing integrals converge to some real number.

5.5 Example (Continuation of Example 5.3)**.** Calculating the expectation of the random variable X of the example above yields

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_0^2 x \cdot \frac{3}{8} \cdot (4x - 2x^2)\, dx = \frac{3}{8} \int_0^2 4x^2 - 2x^3\, dx \\ &= \frac{3}{8} \cdot \left[\frac{4}{3}x^3 - \frac{1}{2}x^4\right]_0^2 = \frac{3}{8} \cdot \left(\frac{32}{3} - 8\right) = 1.\end{aligned}$$

□

To derive the main properties of the expectation in the continuous case, we need the following proposition.

5.6 Proposition. *Let X be a random variable taking nonnegative values, i.e., $X \geq 0$. Then we have the following representation for the expectation,*

$$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}[X > x]\, dx.$$

Proof. As we have

$$\mathbb{P}[X > x] = \int_x^{\infty} f(y)\, dy,$$

it follows that

$$\begin{aligned}\int_0^{\infty} \mathbb{P}[X > x]\, dx &= \int_0^{\infty} \int_x^{\infty} f(y)\, dy\, dx = \int_0^{\infty} \int_0^{y} f(y)\, dx\, dy \\ &= \int_0^{\infty} y \cdot f(y)\, dy = \mathbb{E}[X].\end{aligned}$$

We note that in the second step we just used a re-parametrization of the integration area,

$$\{(x,y) : 0 < x \leq \infty,\, x < y < \infty\} = \{(x,y) : 0 < y \leq \infty,\, 0 < x < y\},$$

as seen in Figure 5.4. □

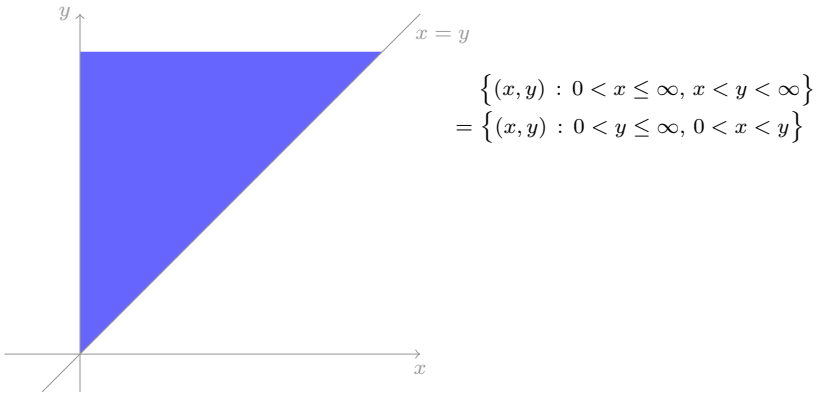


Figure 5.4: Two-dimensional integration area in proof of Proposition 5.6 .

5.7 Corollary. *For an arbitrary random variable X we have*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > x]\, dx - \int_0^\infty \mathbb{P}[X < -x]\, dx.$$

Proof. Homework. □

Similarily we have a formula for expectations of a function of a random variable.

5.8 Proposition. *Let X be a continuous random variable with density f and $g : \mathbb{R} \to \mathbb{R}$ a real valued function. Then we have*

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

Proof. Beyond the scope of this class. □

In particular it follows again that the expectation is linear (just by setting $g(x) = ax + b$).

5.9 Corollary. *The expectation of a continuous random variable X is linear, i.e. for $a, b \in \mathbb{R}$ we have*

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b. \tag{5.1}$$

We note that we may define the variance of a continuous random variable as in the discrete case.

5.10 Definition. The *variance* of a continuous random variable is defined as in the discrete case, namely

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big],$$

and the standard deviation by

$$\mathbb{SD}[X] = \sqrt{\mathbb{V}\mathrm{ar}[X]}.$$

5.11 Corollary. *For the variance of a continuous random variable we have*

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Proof. Exactly as in the discrete case, as the proof of Proposition 4.14 required only linearity which we have now, thanks to Corollary 5.9. □

5.12 Corollary. *For the variance of a continuous random variable X we have for real numbers a and b*

$$\mathbb{V}\mathrm{ar}[a \cdot X + b] = a^2\,\mathbb{V}\mathrm{ar}[X].$$

Proof. Identical to the discrete proof, Proposition 4.15. □

5.13 Example (Continuation of Examples 5.3 and 5.5)**.** The variance in the example above is given via

$$\begin{aligned}\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx = \int_0^2 x^2 \cdot \frac{3}{8} \cdot (4x - 2x^2)\,dx = \frac{3}{8}\int_0^2 4x^3 - 2x^4\,dx \\ &= \frac{3}{8} \cdot \left[x^4 - \frac{2}{5}x^5\right]_0^2 = \frac{3}{8} \cdot \left(16 - \frac{64}{5}\right) = \frac{6}{5}.\end{aligned}$$

by

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{6}{5} - 1^2 = \frac{1}{5}.$$

□

Moment Generating Functions

The quantities $\mathbb{E}[X^k]$ for $k \in \mathbb{N}$ are known as the *moments* of the random variable X. One way to compute them in wuite mechanical manner is with the help of the moment generating function, a technique particular popular with statisticians.

5.14 Definition. Let X be a random variable. The *moment generating function* m_X of X is given by

$$m_X(\lambda) := \mathbb{E}\big[e^{\lambda X}\big], \qquad \lambda in \mathbb{R}.$$

The function can be used to calculate the k-th moment of X by evaluating the k-th derivative of m_X at 0. We see this for $k = 1$ as for a continuous random variable with density f

$$\begin{aligned} m_X'(\lambda) &= \frac{\partial}{\partial\lambda}\,\mathbb{E}\big[e^{\lambda X}\big] = \frac{\partial}{\partial\lambda}\int_{-\infty}^{\infty} e^{\lambda X} f(x)\,dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial\lambda}\Big(e^{\lambda X} f(x)\Big)\,dx \\ &= \int_{-\infty}^{\infty} X e^{\lambda X} f(x)\,dx = \mathbb{E}\big[X e^{\lambda X}\big] \end{aligned}$$

and therefore

$$m_X'(0) = \mathbb{E}\big[X e^{0\cdot X}\big] = \mathbb{E}[X].$$

For general k we get iteratively

$$\begin{aligned} m_X^{(k)}(\lambda) &= \frac{\partial^k}{\partial\lambda^k}\,\mathbb{E}\big[e^{\lambda X}\big] = \frac{\partial^k}{\partial\lambda^k}\int_{-\infty}^{\infty} e^{\lambda X} f(x)\,dx = \int_{-\infty}^{\infty} \frac{\partial^k}{\partial\lambda^k}\Big(e^{\lambda X} f(x)\Big)\,dx \\ &= \int_{-\infty}^{\infty} X^k e^{\lambda X} f(x)\,dx = \mathbb{E}\big[X^k e^{\lambda X}\big] \end{aligned}$$

and therefore

$$m_X^{(k)}(0) = \mathbb{E}\big[X^k e^{0\cdot X}\big] = \mathbb{E}[X^k].$$

We note that this can be done in the same way for discrete random variable.

5.3 Examples of Continuous Random Variables

We are going to discuss some important classes of continuous random variables.

5.3.1 Uniform Random Variables

A random variable X is uniformly distributed on an interval $[a, b]$, $-\infty < a < b < \infty$, shorthand $X \sim \mathcal{U}([a, b])$, if its probability density function is given by

$$f(x) = \begin{cases} 0 & x < a; \\ \frac{1}{b-a} & a \le x \le b; \\ 0 & x > b. \end{cases}$$

It follows that the cumulative distribution function is

$$F(x) = \int_{-\infty}^{x} f(y)\, dy = \left\{ \begin{array}{l} 0 \\ \int_a^x \frac{1}{b-a}\, dy \\ 1 \end{array} \right\} = \begin{cases} 0 & \text{if } x < a; \\ \frac{x-a}{b-a} & \text{if } a \le x \le b; \\ 1 & \text{if } x > b. \end{cases}$$

Uniformly distributed random variables are modeling the case where every outcome in the interval $(a, b]$ is equally likely. It is straight forward to calculate expectation and variance,

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

and

$$\begin{aligned} \mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx = \int_a^b \frac{x^2}{b-a}\, dx = \left[\frac{x^3}{3(b-a)}\right]_a^b = \frac{b^3 - a^3}{3(b-a)} \\ &= \frac{a^2 + ab + b^2}{3} \end{aligned}$$

gives

$$\begin{aligned} \mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{a^2 + 2ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{a^2 + ab + b^2}{3} - \frac{a^2 + 2ab + b^2}{4} = \frac{a^2 - 2ab + b^2}{12} = \frac{(b-a)^2}{12}. \end{aligned}$$

5.15 Example (Continuation of Example 5.1)**.** Actually, we have already encountered a uniform random variable in the bus waiting Example 5.1 where we gave the cdf. We complement this now by the cdf and a calculation

of expectation, variance and standard deviation:

$$f(x) = \begin{cases} 0 & x < 0; \\ \frac{1}{10} & 0 \le x \le 10; \\ 0, & x > 10. \end{cases}$$

$$E[X] = \frac{0+10}{2} = 5,$$

$$\mathbb{V}\text{ar}[X] = \frac{(10-0)^2}{12} = \frac{25}{3} \approx 8.33, \mathbb{SD}[X] = \sqrt{\mathbb{V}\text{ar}[X]} \approx \sqrt{8.33} = \approx 2.89.$$

□

Figure 5.5: Probability density function of X, Examples 5.1 and 5.15.

5.3.2 Exponential Random Variables and Hazard Rates

A random variable X is called *exponentially distributed with parameter* $\lambda > 0$, shorthand $X \sim Exp(\lambda)$, if its probability density function is given by

$$f(x) = \begin{cases} 0 & \text{if } x < 0; \\ \lambda e^{-\lambda x} & \text{if } x \ge 0. \end{cases}$$

It follows that the cumulative distribution function is

$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \left\{ \begin{array}{l} 0 \\ \int_0^x \lambda e^{-\lambda y}\,dy \end{array} \right\} = \begin{cases} 0 & \text{if } x < 0; \\ 1 - e^{-\lambda x} & \text{if } x \ge 0. \end{cases}$$

Next we calculate expectation and variance of an exponential random variable. Using integration by parts we have

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_0^{\infty} \lambda x e^{-\lambda x}\,dx \\ &= \Big[-xe^{-\lambda x}\Big]_0^{\infty} - \int_0^{\infty} -e^{-\lambda x}\,dx = 0 - \Big[\frac{1}{\lambda}e^{-\lambda x}\Big]_0^{\infty} = -\Big(0 - \frac{1}{\lambda}\Big) = \frac{1}{\lambda}, \end{aligned}$$

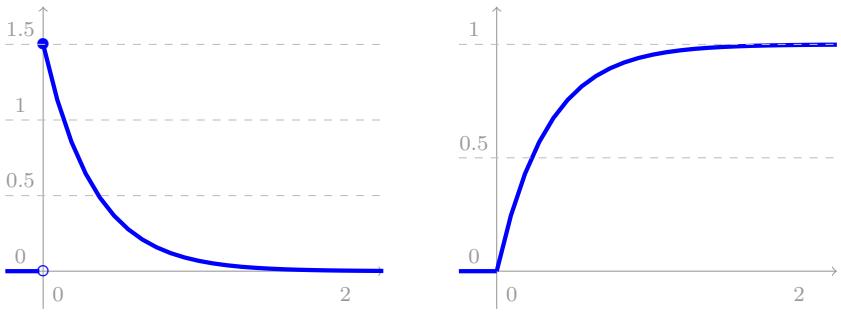


Figure 5.6: Probability density function f and cumulative distribution function F of an exponential random variable with $\lambda = \frac{3}{2}$.

as $\lim_{x\to\infty} xe^{-\lambda x} = 0$ (which can be shown, e.g., by l'Hospital's rule). In the same way, one calculates for the variance by integrating by parts twice (see homework).

Exponentially distributed random variables are used to model times up to the arrival of a random event (as the death of a light bulb or an electrical device, an earthquake,...). They are well fitted for modeling these events since they are *memoryless*: assuming the future live expectancy of a device described by an exponential distribution does not depend on the lifetime up to now. Expressed more mathematically, this means that a non-negative random variable satisfies

A memoryless random variable is sometimes also called *non-aging*

$$\mathbb{P}[X \geq t + s \,|X \geq t] = \mathbb{P}[X \geq s]$$

for all s, $t \geq 0$. For a prooof of this property see homework. Note also we have already seen in previous homework that also the geometric distribution is memoryless. Much more is true: One can show that the *geometric distribution* is the only memoryless distribution on the natural numbers (homework) and the exponential distribution is the only one on the nonnegative real numbers.

Survival Functions and Hazard Rates

Inspired from the notion of memoryless, one might look for anon-negative random variable X at the following quantity.

$$\begin{aligned}\mathbb{P}[X \leq t + h \,|X > t] &= \frac{\mathbb{P}[\{X \leq t + h\} \cap \{X > t\}]}{\mathbb{P}[X > t]} = \frac{\mathbb{P}[t < X \leq t + h\}]}{\mathbb{P}[X > t]} \\ &= \frac{F(s+h) - F(s)}{1 - F(s)} = \frac{F(s+h) - F(s)}{\bar{F}(s)}. \qquad (5.2)\end{aligned}$$

Here $\bar{F}(t) := 1 - F(t)$ is the *survival function*, the complement of the cdf often used in actuarial mathematics. Dividing (5.2) by h and sending it to zero yields

$$\begin{aligned}\lim_{h\downarrow 0} \frac{\mathbb{P}[X \le t+h \,|X > t]}{h} &= \lim_{h\downarrow 0} \frac{F(t+h) - F(t)}{h} \cdot \frac{1}{\bar{F}(t)} \\ &= \frac{F'(s)}{\bar{F}(s)} = \frac{f(s)}{\bar{F}(s)} =: \lambda(s).\end{aligned}$$

$\lambda(t)$ is called the *hazard rate* of X. In the case of an exponentially distributed random variable with parameter $\lambda > 0$, we have exactly

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)} = \frac{\lambda \cdot e^{-\lambda t}}{1 - \left(1 - e^{-\lambda t}\right)} = \frac{\lambda \cdot e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

if $t \ge 0$. Thus one can see the hazard rate as a generalization of the parameter of the exponential distribution, from a constant to a function. Thus, exprtessing a continuous random variable by its hazard rate is putting it into a form similar to that of an exponential random variable. Finally, for a continuous random variable you can recover the density (and the cdf) from the hazard rate.

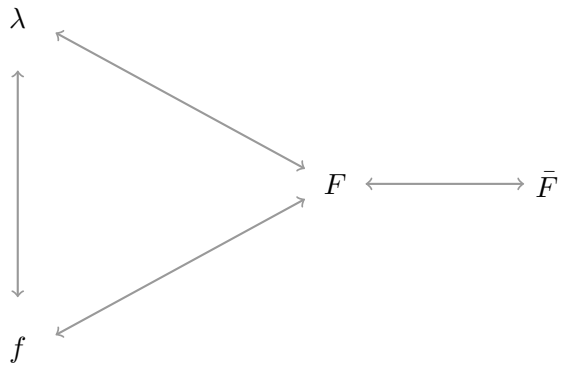


Figure 5.7: Equivalent specifications for continuous random variables.

Given the hazard rate

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)} = -\frac{\bar{F}'(t)}{\bar{F}(t)} = -\Big(\log\left(\bar{F}(t)\right)\Big)'$$

we have

$$\log\left(\bar{F}(t)\right) = -\int_0^t \lambda(s)\, ds + c,$$

for some real constant c and thus

$$\bar{F}(t) = e^c \cdot e^{-\int_0^t \lambda(s)\,ds}.$$

As $\bar{F}(0) = 1 - F(0) = 1$ by Proposition 4.19 iii), we have $c = 0$ and thus

$$\begin{aligned}
\bar{F}(t) &= e^{-\int_0^t \lambda(s)\,ds}, \\
F(t) &= 1 - e^{-\int_0^t \lambda(s)\,ds}, \\
f(t) &= \lambda(t) \cdot e^{-\int_0^t \lambda(s)\,ds}.
\end{aligned}$$

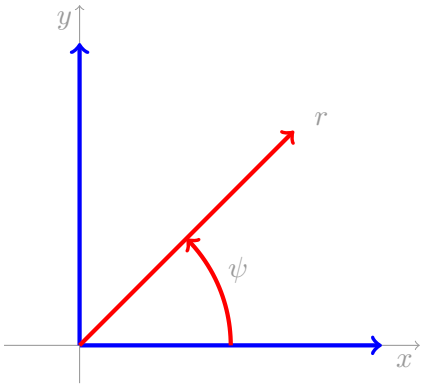
5.3.3 Normal Random Variables

We are looking for a random variable on the whole real line that has a density everywhere smooth and positive. As it has to integrate up to one, a reasonable choice would be a function of the type $c \cdot e^{-\frac{x^2}{2}}$. To find the constant c, we will have to calculate $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx$, which is not possible in a straightforward manner. However, one can calculate its square by turning it into a simpler two-dimensional integral,

$$\begin{aligned}
\left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx\right)^2 &= \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx \cdot \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx \\
&= \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx \cdot \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \cdot e^{-\frac{y^2}{2}}\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy.
\end{aligned}$$

Here we just used that we can rename the integration variable as we like and can interchange integrals. Now, the resulting integral can be solved by a transformation into polar coordinates, $x = r\cos\psi$, $x = r\cos\psi$, remembering that we have for the infinitesimal area segment $dx\,dy = r\,dr\,d\psi$. Then we have

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy &= \int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}}\,r\,dr\,d\psi = \int_0^{2\pi}\left[-e^{-\frac{r^2}{2}}\right]_0^{\infty} d\psi \\
&= \int_0^{2\pi} 1\,d\psi = 2\pi
\end{aligned}$$



$$x = x(r,\psi) = r\cos\psi$$
$$y = y(r,\psi) = r\sin\psi$$

$$\{-\infty < x < \infty, -\infty < x < \infty\}$$
$$= \{0 \le r < \infty, 0 \le \psi < 2\pi\}$$

$$dx\,dy = \begin{vmatrix} \frac{dx(r,\psi)}{dr} & \frac{dx(r,\psi)}{d\psi} \\ \frac{dy(r,\psi)}{dr} & \frac{dy(r,\psi)}{d\psi} \end{vmatrix} d\psi\,dr$$
$$= \begin{vmatrix} \cos\psi & -r\sin\psi \\ \sin\psi & r\cos\psi \end{vmatrix} d\psi\,dr = r\,d\psi\,dr$$

Figure 5.8: Change to polar coordinates.

Thus we get in total

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

and we can define the density

$$\varphi(x) := f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad \text{for all } x \in \mathbb{R}.$$

A random variable with this density is called standard normal distributed. Its cdf is

$$\Phi(x) := F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}}\,dy, \qquad \text{for all } x \in \mathbb{R},$$

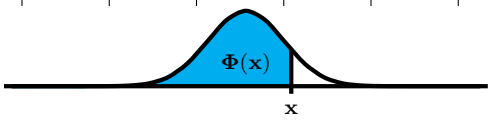
where the integral is not expressible by elementary functions. We will thus have to rely on numerical evaluation by a computer or, more classical, by a table (see Table 5.1).

We note that φ is symmetric, $\varphi(x) = \varphi(-x)$ Thus buy using the substitution $z = -y$ we have

$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)\,dy = -\int_{\infty}^{-x} \varphi(-z)\,dz = \int_{-x}^{\infty} \varphi(z)\,dz = 1 - \int_{-\infty}^{x} \varphi(z)\,dz$$
$$= 1 - \Phi(x)$$

Table of the cdf of the standard normal distribution $\Phi(x)$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000



For negative values of x we just use the identity $\Phi(x) = 1 - \Phi(-x)$.

Table 5.1: Table for the cumulative distribution function of a standard normal distribution

Next we calculate expectation and variance of a standard normal distributed random variable. In the same way as before

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} x\varphi(x)\,dx = \int_{-\infty}^{0} x\varphi(x)\,dx + \int_{0}^{\infty} x\varphi(x)\,dx \\ &= -\int_{\infty}^{0} (-y)\varphi(-y)\,dy + \int_{0}^{\infty} x\varphi(x)\,dx \\ &= -\int_{0}^{\infty} y\varphi(y)\,dy + \int_{0}^{\infty} x\varphi(x)\,dx = 0,\end{aligned}$$

and, using integration by parts,

$$\begin{aligned}\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \cdot x e^{-\frac{x^2}{2}}\,dx \\ &= \frac{1}{\sqrt{2\pi}} \Big[-x e^{-\frac{x^2}{2}}\Big]_{-\infty}^{\infty} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -e^{-\frac{x^2}{2}}\,dx = 0 + \frac{\sqrt{2\pi}}{\sqrt{2\pi}} = 1.\end{aligned}$$

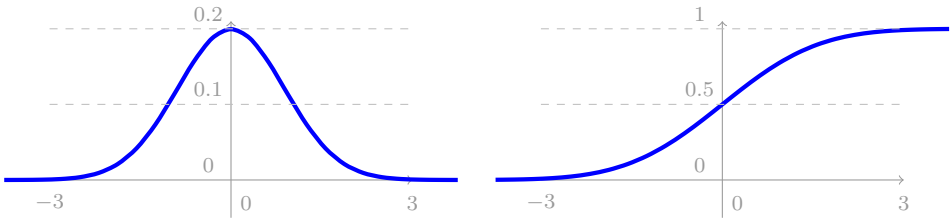


Figure 5.9: Probability density function φ and cumulative distribution function Φ of the standard normal distribution.

If X is a standard normal distribution, we can define a new random variable $Z := \sigma X + \mu$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$. Z is now a random variable with

$$\mathbb{E}[Z] = \mu, \qquad \mathbb{V}\mathrm{ar}[Z] = \sigma^2$$

by Corollaries 5.9 and 5.12. We say that Z is normally distributed with mean μ and variance σ^2 and write as shorthand $Z \sim \mathcal{N}(\mu, \sigma)$. To calculate the density and the cdf of Z, we will make use of the substitution $y = \frac{z-\mu}{\sigma}$ (and thus $dy = \frac{1}{\sigma}\,dz$) and get

$$\begin{aligned}F_Z(x) = \mathbb{P}[Z \le x] = \mathbb{P}[\sigma X + \mu \le x] = \mathbb{P}\Big[X \le \frac{x-\mu}{\sigma}\Big] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{y^2}{2}}\,dy \\ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{\left(\frac{z-\mu}{\sigma}\right)^2}{2}} \cdot \frac{1}{\sigma}\,dz &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(z-\mu)^2}{2\sigma^2}}\,dz,\end{aligned}$$

whence

$$f_Z(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}.$$

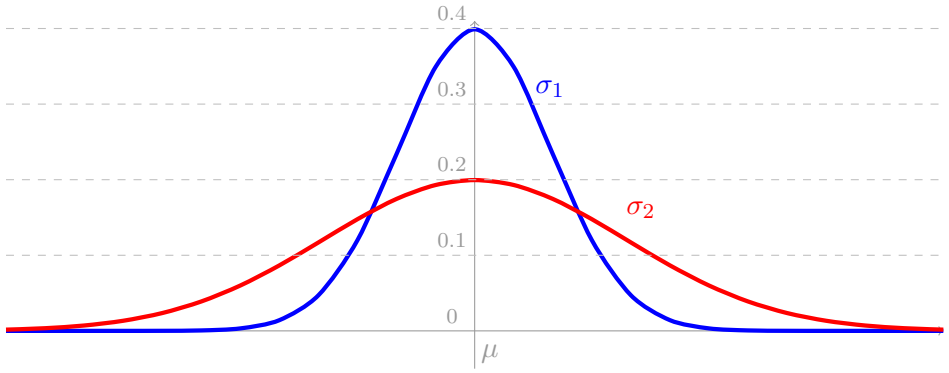


Figure 5.10: Two normal probability density functions with mean μ and $\sigma_2 > \sigma_1$.

The importance of the normal distributions stems from the fact that it appears often as a limit distribution (more in Chapter 7). In particular, the normal distribution is a good approximation of the binomial distribution if n is large. More precisely, we have for a $\mathcal{B}(n,p)$ distributed random variable X (with large n) that $\mu = \mathbb{E}[X] = np$ and $\sigma^2 = \mathbb{V}\mathrm{ar}[X] = np(1-p)$. If Z is now a standard normal variable, we have that $Y := \sqrt{np(1-p)}Z + np$ is a $\mathcal{N}(np, np(1-p))$ distributed random variable and we have

This is called the *de Moivre – Laplace theorem.*

$$\begin{aligned}\mathbb{P}[X \le x] \approx \mathbb{P}[Y \le x] &= \mathbb{P}[\sqrt{np(1-p)}Z + np \le x] = \mathbb{P}\left[Z \le \frac{x-np}{\sqrt{np(1-p)}}\right] \\ &= \Phi\left(\frac{x-np}{\sqrt{np(1-p)}}\right).\end{aligned}$$

Or, put differently by setting $y = \frac{x-np}{\sqrt{np(1-p)}}$,

$$\mathbb{P}\left[\underbrace{\frac{X - \overbrace{np}^{\mu}}{\sqrt{np(1-p)}}}_{\sigma} \le y\right] \approx \Phi(y).$$

As a rule of thumb one can say that this approximation is usually reasonable as long as $np(1-p) > 18$.

Thus X is getting normalized by moving the mean to zero and normalizing the variance to one.

5.16 Example. A test consists of 120 multiple-choice questions, each of which gives 4 choices, exactly one of them being correct.How likely is it to get more then 40 correct answers correct?

Solution. We know that the solution is given as a probability of a binomially $\mathcal{B}(n,p)$- distributed random variable, i.e.,

$$\mathbb{P}[X > 40] = \sum_{k=41}^{120} \mathbb{P}[X = k] = \sum_{k=41}^{120} \binom{120}{k} \Big(\frac{1}{4}\Big)^k \cdot \Big(\frac{3}{4}\Big)^{120-k}.$$

Thus a calculation of this is quite tedious (however it can be done by using a computer ... and yields 1.55%). It is much easier to use the normal approximation. Having $n \cdot p = 120 \cdot \frac{1}{4} = 30$ and $n \cdot p \cdot 1 - p = 120 \cdot \frac{1}{4} \cdot \frac{3}{4} = 22.5$, we conclude that for Z being a standard normal distribution

$$\begin{aligned}\mathbb{P}[X > 40] &= 1 - \mathbb{P}[X \leq 40] = 1 - \mathbb{P}\Big[\frac{X - np}{np(1-p)} \leq \frac{40 - np}{\sqrt{np(1-p)}}\Big] \\ &\approx 1 - \mathbb{P}\Big[Z \leq \frac{40 - 30}{\sqrt{22.5}}\Big] \approx 1 - \Phi\Big(\frac{10}{\sqrt{22.5}}\Big) = 1 - \Phi(2.108) \\ &\approx 1 - 0.9825 = 1.75\%.\end{aligned}$$

□

We note that this is not the unique way to calcualte the result. As X takes only interegers as values, we have

$$\begin{aligned}\mathbb{P}[X > 40] = \mathbb{P}[X \geq 41] &= 1 - \mathbb{P}\Big[\frac{X - np}{np(1-p)} < \frac{41 - np}{\sqrt{np(1-p)}}\Big] \\ &\approx 1 - \mathbb{P}\Big[Z < \frac{41 - 30}{\sqrt{22.5}}\Big] = 1 - \mathbb{P}\Big[Z \leq \frac{41 - 30}{\sqrt{22.5}}\Big] \\ &\approx 1 - \Phi\Big(\frac{11}{\sqrt{22.5}}\Big) = 1 - \Phi(2.319) \approx 1 - 0.9898 = 1.02\%.\end{aligned}$$

This comes from the fact that the normal distribution is continuous (and thus $\mathbb{P}[Z \leq z] = \mathbb{P}[Z < z]$ for all $z \in \mathbb{R}$) while the approximated binomial distribution is not. In practice one often opts for a *continuity correction* and chooses the midpoint of the interval. Applied to our example this yields

$$\mathbb{P}[X > 40.5] \approx \mathbb{P}\Big[Z > \frac{10.5}{\sqrt{22.5}}\Big] \approx 1 - \Phi(2.214) \approx 1 - 0.9866 = 1.34\%.$$

Continuity correction does not necessarily provides superior results. It works best when p close to 0.5.

Chapter 6

Jointly Distributed Random Variables

6.1 Joint Distribution of Random Variables

Until now, we have always considered one random variable alone. However, we are often interested in the study of multiple random variables at the same time, describing phenomena that occur together (say, the sum of five dice thrown and the value of the die with the highest number). To do this, we have to describe the *joint* distribution of several random variables. To keep things simple, we will stay with two random variables most of the time, the generalization to more is straightforward.

As we have described the cumulative distribution function of a random variable X as the probability that it is equal or smaller than a real number x,

$$F_X(x) = \mathbb{P}[X \leq x],$$

it suggests itself how this can be generalized to two random variables, X and Y. We look into the case where each one is less than or equal to a given real number.

6.1 Definition. The *joint cumulative distribution function* of the random variables X and Y is given by

$$F_{X,Y}(x, y) = \mathbb{P}[X \leq x,\, Y \leq y] = \mathbb{P}\big[\{X \leq x\} \cap \{Y \leq y\}\big].$$

We note that if we know $F_{X,Y}$, we can easily reconstruct F_X and F_Y,

$$\begin{aligned} F_X(x) &= \mathbb{P}[X \leq x] = \mathbb{P}[X \leq x,\, Y < \infty] = \lim_{b\to\infty} \mathbb{P}[X \leq x,\, Y \leq b] \\ &= \lim_{b\to\infty} F_{X,Y}(x,b) =: F_{X,Y}(x,\infty), \end{aligned}$$

where the last expression is just a shorthand for the limit in the rem before. In the same way we get $F_Y(y) = F_{X,Y}(\infty, y)$. We call F_X and F_Y the *marginal cumulative distribution functions* of the joint cumulative distribution function $F_{X,Y}$.

This way to define joint distribution appears at first sight a bit unhandy. Indeed, the calculation with joint distributions is more involved. As an example, consider that we could calculate $\mathbb{P}[X > x]$ easily by writing it as

$$\mathbb{P}[X > x] = 1 - \mathbb{P}[X \leq x] = 1 - F_X(x).$$

For the joint distribution it is not as easy to represent, say $\mathbb{P}[X > x,\, Y > y]$. But nevertheless we can calculate it in terms of the joint distribution function (and the marginals – being itself just particular instances of the joint distribution function as seen above),

$$\begin{aligned} & \mathbb{P}[X > x,\, Y > y] \\ = \;& \mathbb{P}\big[\{X > x\} \cap \{Y > y\}\big] = 1 - \mathbb{P}\Big[\big(\{X > x\} \cap \{Y > y\}\big)^c\Big] \\ \overset{\text{Prop. 2.7 b)}}{=} \;& 1 - \mathbb{P}\big[\{X > x\}^c \cup \{Y > y\}^c\big] \\ \overset{\text{Prop. 2.10 e)}}{=} \;& 1 - \mathbb{P}\big[\{X > x\}^c\big] - \mathbb{P}\big[\{Y > y\}^c\big] + \mathbb{P}\big[\{X > x\}^c \cap \{Y > y\}^c\big] \\ = \;& 1 - \mathbb{P}[X \leq x] - \mathbb{P}[Y \leq y] + \mathbb{P}[X \leq x,\, Y \leq y] \\ = \;& 1 - F_X(x) - F_Y(y) + F_{X,Y}(x,y). \end{aligned}$$

Here we used the inclusion-exclusion principle for the special case $n = 2$, see Proposition 2.10 e).

We do not want to restrict ourself to the joint cdf but want also to establish the analogues of probability mass and density functions. Starting with discrete random variables, we just set

$$p_{X,Y}(x,y) = \mathbb{P}[X = x,\, Y = y]$$

two define the *joint probability mass distribution*. We see directly that

$$\begin{aligned}
p_X(x) &= \mathbb{P}[X = x] = \mathbb{P}\Big[\{X = x\} \cap \Big(\bigcup_j \{Y = y_j\}\Big)\Big] \\
&= \sum_j \mathbb{P}[X = x,\, Y = y_j] = \sum_j p_{X,Y}(x, y_j) \\
p_Y(y) &= \mathbb{P}[Y = y] = \mathbb{P}\Big[\Big(\bigcup_j \{X = x_j\}\Big) \cap \{Y = y\}\Big] \\
&= \sum_j \mathbb{P}[X = x_j,\, Y = y] = \sum_j p_{X,Y}(x_j, y) \\
F_{X,Y}(x, y) &= \mathbb{P}[X \le x,\, Y \le y] = \sum_{\substack{j\,:\,x_j \le x \\ k\,:\,y_k \le y}} p_{X,Y}(x_j, y_k)
\end{aligned}$$

6.2 Example. Assume that we have an urn filled with 3 red, 4 white and 5 blue balls. We draw two balls (without replacement) and denote

X	...	number of red balls drawn,
Y	...	number of white balls drawn.

What is the joint probability mass function of X and Y?

Solution. Straightforward computations along the lines of Chapter 1 give

$$\begin{aligned}
p_{X,Y}(0,0) &= \frac{\binom{3}{0}\cdot\binom{4}{0}\cdot\binom{5}{2}}{\binom{12}{2}} = \frac{1\cdot 1\cdot 10}{66} = \frac{10}{66} = \frac{5}{33}, \\
p_{X,Y}(0,1) &= \frac{\binom{3}{0}\cdot\binom{4}{1}\cdot\binom{5}{1}}{\binom{12}{2}} = \frac{1\cdot 4\cdot 5}{66} = \frac{20}{66} = \frac{10}{33}, \\
p_{X,Y}(1,0) &= \frac{\binom{3}{1}\cdot\binom{4}{0}\cdot\binom{5}{1}}{\binom{12}{2}} = \frac{3\cdot 1\cdot 5}{66} = \frac{15}{66} = \frac{5}{22}, \\
p_{X,Y}(1,1) &= \frac{\binom{3}{1}\cdot\binom{4}{1}\cdot\binom{5}{0}}{\binom{12}{2}} = \frac{3\cdot 4\cdot 1}{66} = \frac{12}{66} = \frac{2}{11}, \\
p_{X,Y}(0,2) &= \frac{\binom{3}{0}\cdot\binom{4}{2}\cdot\binom{5}{0}}{\binom{12}{2}} = \frac{1\cdot 6\cdot 1}{66} = \frac{6}{66} = \frac{1}{11}, \\
p_{X,Y}(2,0) &= \frac{\binom{3}{2}\cdot\binom{4}{0}\cdot\binom{5}{0}}{\binom{12}{2}} = \frac{3\cdot 1\cdot 1}{66} = \frac{3}{66} = \frac{1}{22}.
\end{aligned}$$

We may check that indeed the sum of all probabilities gives 1. A very useful way to write the probabilities is in form of a quadratic array, see Table 6.1. This representation is very convenient as it allows for the straightforward calculation (and display) of the marginals. For an illustration of a joint distribution and the marginals see Figure 6.1. □

i \	0	1	2	$\mathbb{P}[X=i]$
0	$\frac{10}{66}$	$\frac{20}{66}$	$\frac{6}{66}$	$\frac{36}{66}$
1	$\frac{15}{66}$	$\frac{12}{66}$	0	$\frac{27}{66}$
2	$\frac{3}{66}$	0	0	$\frac{3}{66}$
$\mathbb{P}[Y=j]$	$\frac{28}{66}$	$\frac{32}{66}$	$\frac{6}{66}$	1

Table 6.1: Joint probability mass function of X and Y, Example 6.2, and its marginal mass distributions.

In the case of continuous random variable we want to define a *joint density function* $f_{X,Y}$ similarly to the joint probability mass functions in the discrete case. This means that we have to write the joint cdf as a double integral over the joint density,

$$F_{X,Y}(x,y) = \mathbb{P}[X \le x,\, Y \le y] = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(a,b)\,db\,da.$$

In the case that such a density exists that we say X and Y are jointly continuous random variables. We observe that this is always the case when $F_{X,Y}$ is continuously differentiable in both variables and we get

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

6.3 Example. Let's consider two random variables X and Y with joint density

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x}e^{-2y} & \text{if } 0 \le x < \infty, 0 \le y \le \infty; \\ 0 & \text{else.} \end{cases}$$

Calculate

a) $\mathbb{P}[X > 1,\, Y < 1]$,

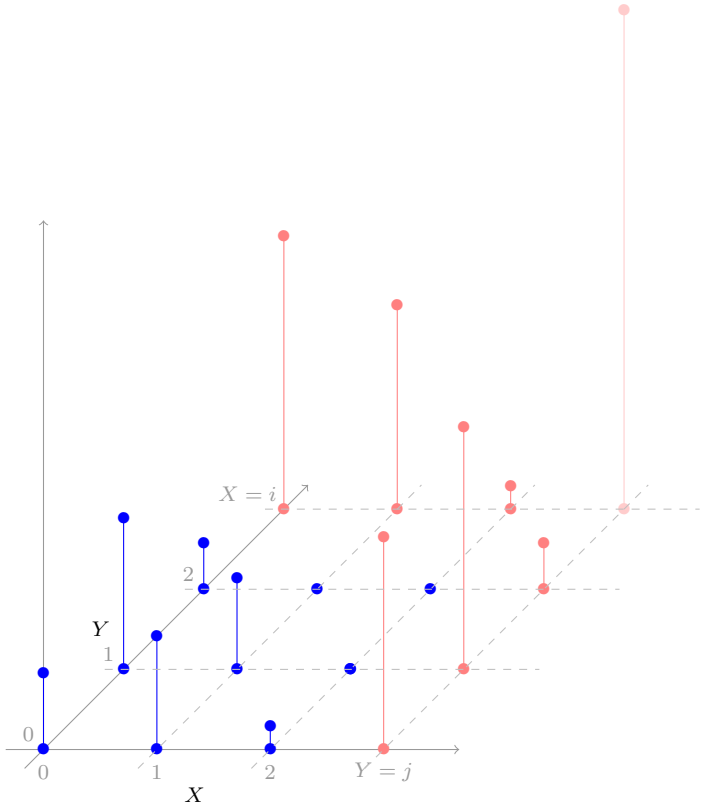


Figure 6.1: Joint probability mass function of X and Y, Example 6.2, (blue) and its marginals (red).

b) $\mathbb{P}[X < Y]$.

Solution.

a) Writing the probability as double integral and integrating it out stepwise gives

$$\begin{aligned}
\mathbb{P}[X > 1,\, Y < 1] &= \int_1^\infty \int_{-\infty}^1 f_{X,Y}(x,y)\,dy\,dx \\
&= \int_1^\infty \int_0^1 2e^{-x}e^{-2y}\,dy\,dx \\
&= \int_1^\infty e^{-x}\,dx \int_0^1 2e^{-2y}\,dy = \Big[-e^{-x}\Big]_1^\infty \cdot \Big[-e^{-2y}\Big]_0^1 \\
&= (0 + e^{-1}) \cdot (-e^{-2} + 1) = e^{-1} - e^{-3}.
\end{aligned}$$

b) Here it works similarly, however as the integration region is not parallel to the axis, we have to be careful about the order of integration,

$$\begin{aligned}
\mathbb{P}[X < Y] &= \iint_{\{(x,y)\in\mathbb{R}^2 \,:\, x<y\}} f_{X,Y}(x,y)\,dy\,dx \\
&= \int_0^\infty \int_0^y 2e^{-x}e^{-2y}\,dx\,dy = \int_0^\infty 2e^{-2y}\cdot\left[-e^{-x}\right]_0^y dy \\
&= \int_0^\infty 2e^{-2y}\cdot(1-e^{-y})\,dy = \int_0^\infty 2e^{-2y}\,dy - \int_0^\infty 2e^{-3y}\,dy \\
&= \left[-e^{-2y}\right]_0^\infty - \left[-\frac{2}{3}e^{-2y}\right]_0^\infty = 1 - \frac{2}{3} = \frac{1}{3}.
\end{aligned}$$

□

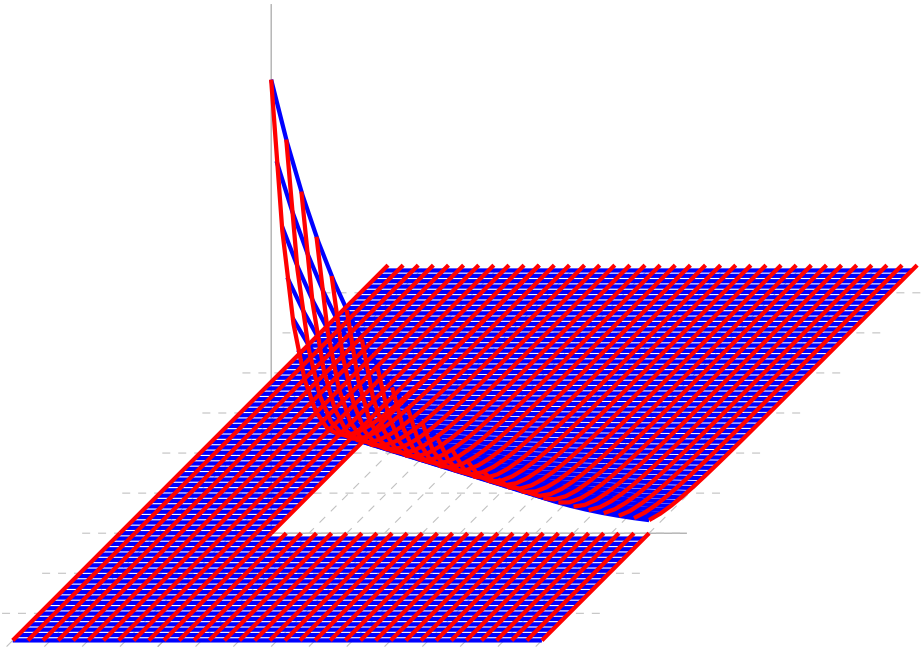


Figure 6.2: Joint probability density function of X and Y, Example 6.3.

6.2 Independent Random Variables

We have seen that when we have the joint distribution function of two random variables, we can easily calculate the marginal distributions. The

inverse is not true, just knowing the marginals does not specify uniquely a joint distribution, as the following example illustrates.

6.4 Example. Assume that the two random variables X and Y are jointly distributed with joint probability mass distribution

$$p_{X,Y}(0,0) = \frac{1}{4}, \quad p_{X,Y}(1,0) = \frac{1}{4}, \quad p_{X,Y}(0,1) = \frac{1}{4}, \quad p_{X,Y}(1,1) = \frac{1}{4},$$

and $\bar{X}$, $\bar{Y}$ distributed with joint probability mass distribution

$$p_{\bar{X},\bar{Y}}(0,0) = 0, \quad p_{\bar{X},\bar{Y}}(1,0) = \frac{1}{2}, \quad p_{\bar{X},\bar{Y}}(0,1) = \frac{1}{2}, \quad p_{\bar{X},\bar{Y}}(0,0) = 0.$$

Then of course the joint probability mass functions $p_{X,Y}$ and $p_{\bar{X},\bar{Y}}$. However the produce identical marginal mass distributions

$$\begin{aligned}
p_X(0) &= p_{X,Y}(0,0) + p_{X,Y}(0,1) = \frac{1}{2} = p_{\bar{X},\bar{Y}}(0,0) + p_{\bar{X},\bar{Y}}(0,1) = p_{\bar{X}}(0) \\
p_X(1) &= p_{X,Y}(1,0) + p_{X,Y}(1,1) = \frac{1}{2} = p_{\bar{X},\bar{Y}}(1,0) + p_{\bar{X},\bar{Y}}(1,1) = p_{\bar{X}}(1) \\
p_Y(0) &= p_{X,Y}(0,0) + p_{X,Y}(1,0) = \frac{1}{2} = p_{\bar{X},\bar{Y}}(0,0) + p_{\bar{X},\bar{Y}}(1,0) = p_{\bar{Y}}(0) \\
p_Y(1) &= p_{X,Y}(0,1) + p_{X,Y}(1,1) = \frac{1}{2} = p_{\bar{X},\bar{Y}}(0,1) + p_{\bar{X},\bar{Y}}(1,1) = p_{\bar{Y}}(1).
\end{aligned}$$

□

We are in particular interested in the case where the outcome of X does not affect the outcome of Y, thus we want to introduce a notion of independence for random variables. We remember that we introduced in Section 3.3 the notion of independence for events E, F, requiring that

$$\mathbb{P}[E \cap F] = \mathbb{P}[E] \cdot \mathbb{P}[F].$$

For that two random variables are independent, we want that all events we can describe with them are independent, i.e., the events $\{X \in A\}$, $\{Y \in B\}$, for all subsets A, B of the real line,

$$\mathbb{P}[\{X \in A\} \cap \{Y \in B\}] = \mathbb{P}[X \in A] \cdot \mathbb{P}[Y \in B].$$

for any choice of A, B sets in the real line. Fortunately one can simplify this, as it is already true when it holds for all intervals, or,

$$F_{X,Y}(x,y) = \mathbb{P}[X \leq x,\, Y \leq y] = \mathbb{P}[X \leq x] \cdot \mathbb{P}[Y \leq y] = F_X(x) \cdot F_Y(y),$$

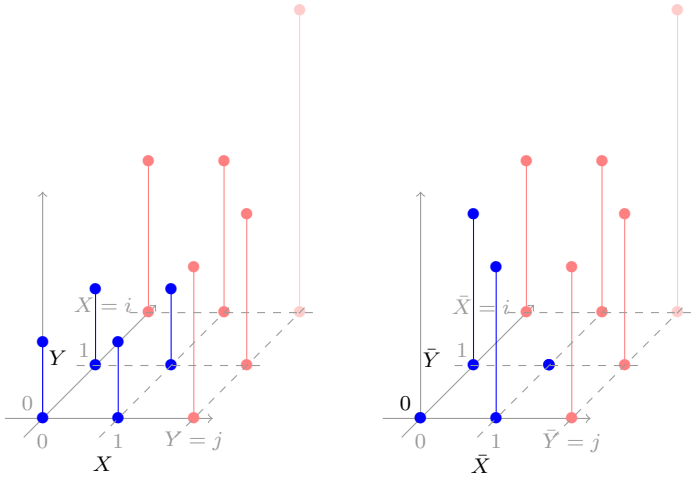


Figure 6.3: Joint probability mass function of X, Y and $\bar{X}$, $\bar{Y}$, Example 6.4, (blue) and their marginals (red). While the marginals are the same, the joint distributions are quite different.

for all x, $y \in \mathbb{R}$. One can translate this to the probability mass functions and densities to get as equivalent condition

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y),$$

or,

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y).$$

6.5 Example. Assume that the joint density of X and Y is given by

$$f_{X,Y}(x, y) = \begin{cases} 3x & \text{if } 0 \leq x < 1, 0 \leq y < x; \\ 0 & \text{else.} \end{cases}$$

Are X and Y independent?

Solution. We calculate first the marginals,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \left\{ \begin{array}{ll} \int_0^x 3x\,dy & \text{if } 0 \le x < 1; \\ 0 & \text{else,} \end{array} \right\}$$
$$= \left\{ \begin{array}{ll} [3xy]_{y=0}^{x} & \text{if } 0 \le x < 1; \\ 0 & \text{else,} \end{array} \right\} = \left\{ \begin{array}{ll} 3x^2 & \text{if } 0 \le x < 1; \\ 0 & \text{else,} \end{array} \right.$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \left\{ \begin{array}{ll} \int_y^1 3x\,dx & \text{if } 0 \le y < 1; \\ 0 & \text{else,} \end{array} \right\}$$
$$= \left\{ \begin{array}{ll} [\frac{3}{2}x^2]_{x=y}^{1} & \text{if } 0 \le y < 1; \\ 0 & \text{else,} \end{array} \right\} = \left\{ \begin{array}{ll} \frac{3}{2}(1-y^2) & \text{if } 0 \le y < 1; \\ 0 & \text{else.} \end{array} \right.$$

As we see that

$$f_X(x) \cdot f_Y(y) = \left\{ \begin{array}{ll} \frac{9}{2}x^2(1-y^2) & \text{if } 0 \le x < 1, 0 \le y < x; \\ 0 & \text{else} \end{array} \right\} \neq f_{X,Y}(x,y),$$

we can say that X and Y are not independent. □

6.3 Sums of Independent Random Variables

We know now a bit about the joint distribution of random variables. Can we say more about the sum of two (or more) random variables, in particular if they are independent?

We start with the case of discrete random variables and restrict for reasons of simplicity to the case where both random variables take values in the nonnegative integers. Then, of course, the sum can have only value n if one variable has value k, $0 \le k \le n$ and the other one $n-k$. Summing up over all possibilities and using independence gives the result,

$$p_{X+Y}(n) = \mathbb{P}[X+Y=n] = \sum_{k=0}^{n} \mathbb{P}[X=k,\, Y=n-k]$$
$$= \sum_{k=0}^{n} \mathbb{P}[X=k] \cdot \mathbb{P}[Y=n-k] = \sum_{k=0}^{n} p_X(k) \cdot p_Y(n-k).$$

We say that p_{X+Y} is the *convolution* of p_X and p_Y and write

$$(p_X * p_Y)(n) := p_{X+Y}(n) = \sum_{k=0}^{n} p_X(k) \cdot p_Y(n-k).$$

6.6 Example. Let X be a Poisson distributed random variable with parameter $\lambda > 0$ and Y Poisson distributed with parameter $\mu > 0$ and independent of X. What is the distribution of $X + Y$?

Solution. Using the convolution formula, rearranging and applying the binomial theorem gives us for every nonnegative integer n

$$\begin{aligned} p_{X+Y}(n) &= \sum_{k=0}^{n} p_X(k) \cdot p_Y(n-k) = \sum_{k=0}^{n} e^{-\lambda}\frac{\lambda^k}{k!}e^{-\mu}\frac{\mu^{n-k}}{(n-k)!} \\ &= e^{-\lambda}e^{-\mu}\frac{1}{n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda^k\mu^{n-k} = e^{-(\lambda+\mu)}\frac{1}{n!}\sum_{k=0}^{n}\binom{n}{k}\lambda^k\mu^{n-k} \\ &\overset{(1.1)}{=} e^{-(\lambda+\mu)}\frac{1}{n!}(\lambda+\mu)^n = e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^n}{n!}, \end{aligned}$$

and we see that the sum is itself a Poisson random variable, with parameter $\lambda + \mu$. □

Note that this is not the usual case. Most sums of two independent random variables of the same distribution family do NOT have a distribution of this family, see Example 6.7.

We want to extend now the argument to continuous random variables. To do so, we have first to work with the cdf .

$$\begin{aligned} F_{X+Y}(z) = \mathbb{P}[X+Y \le z] &= \iint_{\{(x,y)\in\mathbb{R}^2 \,:\, x+y\le z\}} f_{X,Y}(x,y)\,dy\,dx \\ &= \iint_{\{(x,y)\in\mathbb{R}^2 \,:\, y\le z-x\}} f_X(x)\cdot f_Y(y)\,dy\,dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_X(x)\cdot f_Y(y)\,dy\,dx \\ &= \int_{-\infty}^{\infty} f_X(x)\int_{-\infty}^{z-x} f_Y(y)\,dy\,dx = \int_{-\infty}^{\infty} f_X(x)\cdot F_Y(z-x)\,dx. \end{aligned}$$

Now differentiating with respect to z gives

$$\begin{aligned} f_{X+Y}(z) = F'_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(x)\cdot F'_Y(z-x)\,dx \\ &= \int_{-\infty}^{\infty} f_X(x)\cdot f_Y(z-x)\,dx. \end{aligned}$$

We call this again the convolution product of f_X and f_Y and write

$$(f_X * f_Y)(z) := f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)\cdot f_Y(z-x)\,dx.$$

6.7 Example. Let X and Y be two independent random variables, distributed uniformly on the interval $[0,1]$. Calculate the density of $X + Y$.

Solution. We note first that

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

and thus $f_Y(z-x) = 1$ if and only if $z \geq x \geq z-1$ (and otherwise zero). Thus we can conclude that

$$\begin{aligned} f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(x) \cdot f_Y(z-x)\,dx = \int_0^1 f_Y(z-x)\,dx \\ &= \left\{ \begin{array}{ll} 0 & z < 0; \\ \int_0^z 1\,dx & 0 \leq z \leq 1; \\ \int_{z-1}^1 1\,dx & 1 < z \leq 2; \\ 0 & z > 2 \end{array} \right\} = \left\{ \begin{array}{ll} 0 & z < 0; \\ z & 0 \leq z \leq 1; \\ 2-z & 1 < z \leq 2; \\ 0 & z > 2. \end{array} \right. \end{aligned}$$

Thus the convolution of two uniform densities (on$[0,1]$) gives a triangular density (on $[0,2]$), see Figure 6.4. □

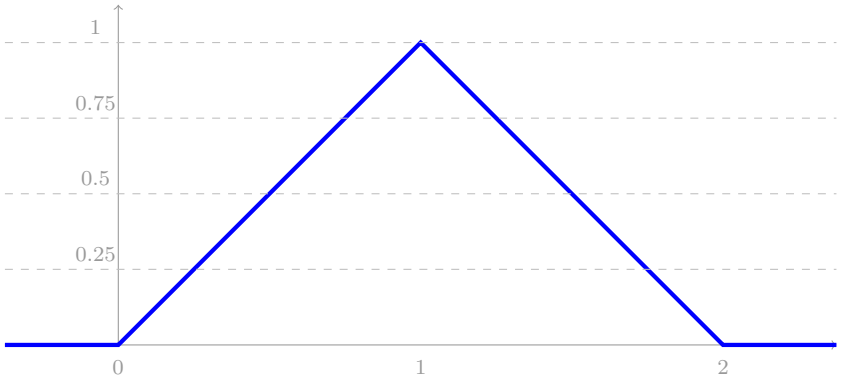


Figure 6.4: The convolution of two uniform densities gives a triangular density, Example 6.7.

6.8 Proposition. *If X and Y are independent normal distributed random variables $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then $X+Y$ is again a normal distributed random variable with mean $\mu_1+\mu_2$ and variance $\sigma_1^2+\sigma_2^2$, $X+Y \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.*

Proof. The proof by integration is somehow tedious and left out here. It can be found in Ross[1]. □

[1] Ross, *A First Course in Probability*, Proposition 5.3.2.

6.4 Expectations, Variance and Covariance

We want to look now on more general properties of jointly distributed random variables, even if they are not independent. Therefore we have first to specify how to calculate the expected value of a function of two (or more) random variables.

6.9 Proposition. *Let X, Y be two continuous random variables with joint density $f_{X,Y}$ or two discrete random variables with joint probability mass function $p_{X,Y}$ and let $h : \mathbb{R} \to \mathbb{R}$ be a real valued function. Then*

$$\mathbb{E}[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) \cdot f_{X,Y}(x,y)\, dx\, dy,$$

or, respectively,

$$\mathbb{E}[h(X,Y)] = \sum_j \sum_k h(x_j, y_k) \cdot p_{X,Y}(x_j, y_k).$$

Proof. The proofs work exactly as those in Propositions 5.8 and 4.9, just having a double integral/sum instead the simple one... □

First we want to inspect the behavior of sums of two random variables. Here we have a very general result.

6.10 Proposition. *If X and Y are jointly distributed random variables, then $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.*

Proof. We prove the statement here for continuous random variables with joint density $f_{X,Y}$, the proof for discrete random variables goes analogously.

$$\begin{aligned}
\mathbb{E}[X+Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y) \cdot f_{X,Y}(x,y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y)\, dx\, dy + \int_{-\infty}^{\infty} y f_{X,Y}(x,y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy\, dx + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} x f_X(x)\, dx + \int_{-\infty}^{\infty} y f_Y(y)\, dy = \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

□

Looking next on products, we will see that a nice result holds only for independent random variables.

6.11 Proposition. *If X and Y are independent random variables, and $g, h : \mathbb{R} \to \mathbb{R}$ are real valued functions, then*

$$\mathbb{E}[g(X) \cdot h(Y)] = \mathbb{E}[g(X)] \cdot \mathbb{E}[h(Y)].$$

Proof. We give the proof for continuous random variables with density $f_{X,Y}$. By independence we have

$$\begin{aligned}
\mathbb{E}[g(X) \cdot h(Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x) \cdot h(y) \cdot f_{X,Y}(x,y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x) \cdot h(y) \cdot f_X(x) \cdot f_Y(y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx \int_{-\infty}^{\infty} h(y) \cdot f_Y(y)\, dy \\
&= \mathbb{E}[g(X)] \cdot \mathbb{E}[h(Y)].
\end{aligned}$$

The discrete case works by the same method. □

6.12 Corollary. *If X and Y are independent random variables, then*

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

Proof. Just setting $g(x) = x$, $h(y) = y$ in Proposition 6.11. □

It is easy to see that this will not hold in general as for the random variables $\bar{X}$ and $\bar{Y}$ from Example 6.4,

$$p_{\bar{X},\bar{Y}}(0,0) = 0, \quad p_{\bar{X},\bar{Y}}(1,0) = \frac{1}{2}, \quad p_{\bar{X},\bar{Y}}(0,1) = \frac{1}{2}, \quad p_{\bar{X},\bar{Y}}(0,0) = 0.$$

we have

$$\begin{aligned}
\mathbb{E}[\bar{X}\bar{Y}] &= 0 \cdot 0 \cdot 0 + 0 \cdot 1 \cdot \frac{1}{2} + 1 \cdot 0 \cdot \frac{1}{2} + 1 \cdot 1 \cdot 0 = 0 \\
\mathbb{E}[\bar{X}] \cdot \mathbb{E}[\bar{Y}] &= \left(0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2}\right) \cdot \left(0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2}\right) = \frac{1}{4}.
\end{aligned}$$

We can thus use the product of two random variables to measure how much they are not independent, i.e. measuring their dependence structure. Practically one uses therefore centered versions of this random variables (i.e. translated so that the expectation is zero).

6.13 Definition. Given two jointly distributed random variables X, Y, we define their *covariance* as

$$\mathbb{C}\mathrm{ov}[X,Y] = \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])].$$

We note first that in the case $X = Y$, this reduces to the ordinary variance, i.e., $\mathbb{C}\mathrm{ov}[X, X] = \mathbb{V}\mathrm{ar}[X]$. As with the variance, there are easier ways to calculate the covariance as just plugging into the definition.

6.14 Proposition. *For to jointly distributed random variables X, Y we have*

$$\mathbb{C}\mathrm{ov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

Proof. By the linearity of the expectation (Corollaries 4.10 and 5.9) we have, writing $\mathbb{E}[X] = \mu_X$ and $\mathbb{E}[Y] = \mu_Y$ to make clear that we are just dealing with real numbers,

$$\begin{aligned}\mathbb{C}\mathrm{ov}[X, Y] &= \mathbb{E}\Big[(X - \mu_X) \cdot (Y - \mu_Y)\Big] = \mathbb{E}[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mu_Y - \mathbb{E}[Y]\mu_X + \mu_X\mu_Y = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y].\end{aligned}$$

□

The covariance helps us also to formulate the behavior of the variance of two random variables. Random variables with zero covariance are called *uncorrelated*.

6.15 Proposition. *Let X, Y be two jointly distributed random variables and a, $b \in \mathbb{R}$ two real numbers, then it holds for the variance that*

$$\mathbb{V}\mathrm{ar}[aX + bY] = a^2\,\mathbb{V}\mathrm{ar}[X] + b^2\,\mathbb{V}\mathrm{ar}[Y] + 2ab\,\mathbb{C}\mathrm{ov}[X, Y].$$

Proof. From the above Proposition 6.14 and the quadratic behavior of the covariance (Propositions 4.15 and 5.12) it follows that

$$\begin{aligned}\mathbb{V}\mathrm{ar}[aX + bY] &= \mathbb{E}[(aX + bY)^2] - (\mathbb{E}[aX + bY])^2 \\ &= \mathbb{E}[a^2X^2 + 2abXY + b^2Y^2] \\ &\quad - a^2\,\mathbb{E}[X]^2 - 2abE[X] \cdot \mathbb{E}[Y] - b^2\,\mathbb{E}[Y]^2 \\ &= a^2\Big(\mathbb{E}[X^2] - \mathbb{E}[X]^2\Big) + b^2\Big(\mathbb{E}[Y^2] - \mathbb{E}[Y]^2\Big) \\ &\quad + 2ab(\mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]) \\ &= a^2\,\mathbb{V}\mathrm{ar}[X] + b^2\,\mathbb{V}\mathrm{ar}[Y] + 2ab\,\mathbb{C}\mathrm{ov}[X, Y].\end{aligned}$$

□

In particular we have the following corollary for the case of independent random variables.

6.16 Corollary. *For two independent random variables X, Y, we have*

$$\mathbb{V}\mathrm{ar}[X+Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y].$$

Proof. Immediate by setting $a = b = 1$ and remembering that independent random variables have zero covariance. □

The covariance between two random variables can take every real number as value. The dependence structure is often easier to understand if one looks on a normalized version of the covariance taking only values between -1 and 1, the correlation.

6.17 Definition. Let X and Y be two jointly distributed random variables. There *correlation* $\rho[X,Y]$ is given by

$$\rho[X,Y] := \frac{\mathbb{C}\mathrm{ov}[X,Y]}{\sqrt{\mathbb{V}\mathrm{ar}[X]\cdot\mathbb{V}\mathrm{ar}[Y]}}.$$

Let's show first that the correlation takes indeed only the values claimed.

6.18 Proposition. *For two jointly distributed random variables X and Y their correlation $\rho[X,Y]$ takes only values in the interval $[-1,1]$. In particular if X and Y are independent we have $\rho[X,Y] = 0$ and for the cases $Y = X$ and $Y = -X$ we have $\rho[X,X] = 1$, $\rho[X,-X] = -1$.*

Proof. To point out the role of the constants, let us denote the standard deviation of X and Y by $\sigma_X = \sqrt{\mathbb{V}\mathrm{ar}[X]}$ and $\sigma_Y = \sqrt{\mathbb{V}\mathrm{ar}[Y]}$. As by Proposition 6.15

$$\mathbb{V}\mathrm{ar}\Big[\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\Big] = \frac{1}{\sigma_X^2}\,\mathbb{V}\mathrm{ar}[X] + \frac{1}{\sigma_Y^2}\,\mathbb{V}\mathrm{ar}[Y] + 2\frac{\mathbb{C}\mathrm{ov}[X,Y]}{\sigma_X\sigma_Y} = 2 + 2\rho[X,Y]$$

and this is always bigger or equal to 0 as the variance is nonnegative. Thus it follows that $\rho[X,Y] \geq -1$. Analogously we have for the upper bound that

$$\mathbb{V}\mathrm{ar}\Big[\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\Big] = \frac{1}{\sigma_X^2}\,\mathbb{V}\mathrm{ar}[X] + \frac{1}{\sigma_Y^2}\,\mathbb{V}\mathrm{ar}[Y] + 2\frac{\mathbb{C}\mathrm{ov}[X,-Y]}{\sigma_X\sigma_Y} = 2 - 2\rho[X,Y]$$

is nonnegative and thus $\rho[X,Y] \leq 1$. Note that here the second step follows from

$$\begin{aligned}\mathbb{C}\mathrm{ov}[X,-Y] &= \mathbb{E}[X\cdot(-Y)] - \mathbb{E}[X]\cdot\mathbb{E}[-Y] \\ &= -\big(\mathbb{E}[XY] - \mathbb{E}[X]\cdot\mathbb{E}[Y]\big) = -\,\mathbb{C}\mathrm{ov}[X,Y].\end{aligned}$$

Moreover, if X and Y are independent, their covariance is zero, hence their correlation is zero. And since $\mathbb{C}\mathrm{ov}[X,X] = \mathbb{V}\mathrm{ar}[X]$, it follows by the above argument $\mathbb{C}\mathrm{ov}[X,-X] = -\,\mathbb{V}\mathrm{ar}[X]$, and the special cases also follow. □

Random variables that have correlation zero are called *uncorrelated*. Finally we want to point out that while independent random variables have covariance zero, the opposite has not to be true.

6.19 Example. Consider the jointly distributed random variables X, Y with joint probability mass function

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

Then the marginals are

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

and

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

Thus X and Y are uncorrelated as

$$\begin{aligned}\mathbb{C}\mathrm{ov}[X,Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\cdot\mathbb{E}[Y] = (-1)\cdot 1\cdot\frac{1}{4} + 0\cdot 0\cdot\frac{1}{2} + 1\cdot 1\cdot\frac{1}{4} \\ &\quad - \Big(-1\cdot\frac{1}{4} + 0\cdot\frac{1}{2} + 1\cdot\frac{1}{4}\Big)\cdot\Big(1\cdot\frac{1}{2} + 0\cdot\frac{1}{2}\Big) \\ &= 0 - 0\cdot\frac{1}{2} = 0.\end{aligned}$$

On the other hand side we have

$$p_{X,Y}(-1,1) = \frac{1}{4} > \frac{1}{8} = \frac{1}{4}\cdot\frac{1}{2} = p_X(-1)\cdot p_Y(1),$$

thus the random variables are not independent. □

6.20 Example (Continuiton of Example 6.5)**.** We take up the analysis of Example 6.5 and compute for it the covariance and correlation. We have

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^1 x\cdot 3x^2\,dx = \Big[\frac{3}{4}x^4\Big]_0^1 = \frac{3}{4}, \\ \mathbb{E}[Y] &= \int_{-\infty}^{\infty} y f_Y(y)\,dy = \int_0^1 y\cdot\frac{3}{2}(1-y^2)\,dy = \Big[\frac{3}{4}y^2 - \frac{3}{8}y^4\Big]_0^1 = \frac{3}{8}, \\ \mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^1 x^2\cdot 3x^2\,dx = \Big[\frac{3}{5}x^5\Big]_0^1 = \frac{3}{5}, \\ \mathbb{E}[Y^2] &= \int_{-\infty}^{\infty} y^2 f_Y(y)\,dy = \int_0^1 y^2\cdot\frac{3}{2}(1-y^2)\,dy = \Big[\frac{1}{2}y^3 - \frac{3}{10}y^5\Big]_0^1 = \frac{1}{5}, \\ \mathbb{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\,dx\,dy = \int_0^1\int_0^x xy\cdot 3x\,dy\,dx \\ &= \int_0^1 \Big[\frac{3}{2}x^2y^2\Big]_0^x dx = \int_0^1 \frac{3}{2}x^4\,dx = \Big[\frac{3}{10}x^5\Big]_0^1 = \frac{3}{10},\end{aligned}$$

and thus

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X] &= \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2 = \frac{3}{5} - \Big(\frac{3}{4}\Big)^2 = \frac{3}{80}, \\
\mathbb{V}\mathrm{ar}[Y] &= \mathbb{E}\big[Y^2\big] - \mathbb{E}[Y]^2 = \frac{1}{5} - \Big(\frac{3}{8}\Big)^2 = \frac{19}{320}, \\
\mathbb{C}\mathrm{ov}[X,Y] &= \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] = \frac{3}{10} - \frac{3}{4} \cdot \frac{3}{8} = \frac{3}{160} \approx 0.02, \\
\rho[X,Y] &= \frac{\mathbb{C}\mathrm{ov}[X,Y]}{\sqrt{\mathbb{V}\mathrm{ar}[X] \cdot \mathbb{V}\mathrm{ar}[Y]}} = \frac{\frac{3}{160}}{\sqrt{\frac{3}{80} \cdot \frac{19}{320}}} = \frac{3}{\sqrt{57}} \approx 0.40.
\end{aligned}$$

Chapter 7

The Classical Limit Theorems

In the last chapter, we have mainly looked at the behavior of the sum of two random variables, now we are interested in the limiting behavior of sums of random variables, i.e.,

$$X_1 + X_2 + \cdots X_n \xrightarrow{n\to\infty} ?$$

7.1 Two Important Inequalities

We start by proving two important inequalities that will be needed.

7.1 Proposition (Markov's inequality)**.** *Let X be a nonnegative random variable and $\alpha > 0$. Then*

$$P[X \geq \alpha] \leq \frac{\mathbb{E}[X]}{\alpha}.$$

Proof. For fixed $\alpha > 0$, we define the random variable

$$I_\alpha := \begin{cases} 1 & \text{if } X \geq \alpha; \\ 0 & \text{else.} \end{cases}$$

We note that we have always

$$I_\alpha \leq \frac{X}{\alpha},$$

since for $X > \alpha$ the left hand term is 1 and the right hand one greater then 1, and for $X \leq \alpha$ the left hand is 0 and the right hand nonnegative. Thus by taking expectations we get

$$\mathbb{E}[I_\alpha] \leq \mathbb{E}\Big[\frac{X}{\alpha}\Big] = \frac{1}{\alpha}\,\mathbb{E}[X].$$

We note however that by the definition of I_α we have in the case of continuously distributed X

$$\mathbb{E}[I_\alpha] = \int_\alpha^\infty 1 \cdot f_X(x)\,dx = \mathbb{P}[X \geq \alpha],$$

or equally for discrete X,

$$\mathbb{E}[I_\alpha] = \sum_{x_j \geq \alpha} 1 \cdot p_X(x_j) = \mathbb{P}[X \geq \alpha].$$

Thus we conclude the proof by uniting the two statements above,

$$\mathbb{P}[X \geq \alpha] = \mathbb{E}[I_\alpha] \leq \frac{\mathbb{E}[X]}{\alpha}.$$

□

7.2 Proposition (Chebyshev's inequality)**.** *Let X be a random variable with finite mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{V}\mathrm{ar}[X]$. Then for every real $\kappa > 0$ it holds that*

$$\mathbb{P}[|X - \mu| \geq \kappa] \leq \frac{\sigma^2}{\kappa^2}.$$

Proof. We note first that $(X - \mu)^2$ is a nonnegative random variable. Thus we can apply Markov's inequality (Proposition 7.1) to it, yielding

$$\mathbb{P}[(X - \mu)^2 \geq k^2] \leq \frac{\mathbb{E}[(X - \mu)^2]}{\sigma^2}$$

by Markov's inequality. Note however since $\kappa > 0$ we can take the square root inside the probability to conclude

$$\mathbb{P}[|X - \mu| \geq \kappa] = \mathbb{P}[|X - \mu|^2 \geq \kappa^2] \leq \frac{\mathbb{E}[(X - \mu)^2]}{\kappa^2} = \frac{\sigma^2}{\kappa^2}.$$

□

7.2 The (Weak) Law of Large Numbers

Now we turn to one of the classical results in probability theory, namely that the weighted sum of independent and identically distributed random variables converges to the mean as one sends the number of summands to infinity. There are several ways to specify the convergence. We keep in in the simplest form that the probability that the sum differs from the mean by more than any given value goes to zero while increasing the number of summands.

One often says just *iid* for independent and identically distributed (and *LLN* for law of large numbers)

7.3 Theorem (The weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of* independent *and* identically distributed *random variables with finite mean $\mu = \mathbb{E}[X_1]$ and variance $\sigma^2 = \mathbb{V}\mathrm{ar}[X_1]$. Then for every $\varepsilon > 0$*

$$\mathbb{P}\Big[\Big|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\Big| \geq \varepsilon\Big] \xrightarrow{n\to\infty} 0.$$

Proof. We note first that by Proposition 6.10 and Corollaries 4.10 and 5.9

$$\mathbb{E}\Big[\frac{X_1 + X_2 + \ldots + X_n}{n}\Big] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_k] = \frac{1}{n}\cdot(n\cdot\mu) = \mu,$$

and by Proposition 6.15

$$\begin{aligned}\mathbb{V}\mathrm{ar}\Big[\frac{X_1 + X_2 + \ldots + X_n}{n}\Big] &= \frac{1}{n^2}\,\mathbb{V}\mathrm{ar}\Big[\sum_{k=1}^{n} X_k\Big] = \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{V}\mathrm{ar}\Big[X_k\Big] \\ &= \frac{1}{n^2}\cdot(n\cdot\sigma^2) = \frac{\sigma^2}{n}.\end{aligned}$$

Thus we have by Chebyshev's inequality, Proposition 7.2,

$$\mathbb{P}\Big[\Big|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\Big| \geq \varepsilon\Big] \leq \frac{\mathbb{V}\mathrm{ar}\big[\frac{X_1+X_2+\ldots+X_n}{n}\big]}{\varepsilon^2} = \frac{\frac{\sigma^2}{n}}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2},$$

and the term on the right hand tends to zero for $n \to \infty$. □

7.3 The Central Limit Theorem

The second classical convergence result is the central limit theorem, which gives us some information that the (appropriately normalized) distance from the mean

$$\frac{\sqrt{n}}{\sigma}\Big(\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\Big)$$

tends to be normally distributed as $n \to \infty$.

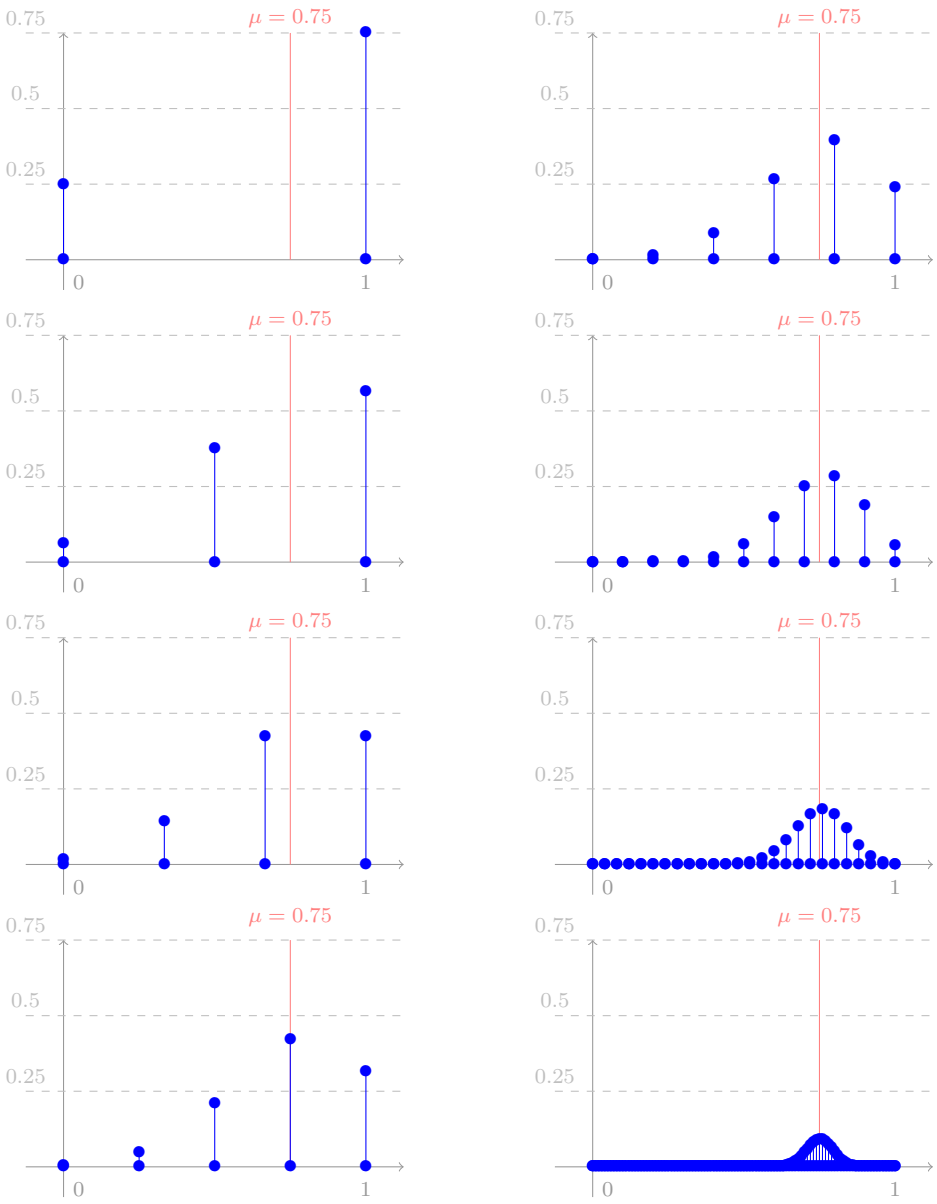


Figure 7.1: *Law of Large Numbers* for $n = 1$, 2, 3, 4, 5, 10, 25 and 100 independent Bernoulli distributed random variables with success probability $p = 0.75$. The weighted sum of the Bernoulli random variables converges to $\mu = 0.75$. The graph of the probability mass function becomes more and more similar to the density of a normal distribution.

7.4 Theorem (The Central Limit Theorem)**.** *Let X_1, X_2, ... be a sequence of* independent *and* identically distributed *random variables with finite mean $\mu = \mathbb{E}[X_1]$ and variance $\sigma^2 = \mathbb{V}\mathrm{ar}[X_1]$. Then for every $z \in \mathbb{R}$*

$$\mathbb{P}\left[\frac{\sqrt{n}}{\sigma}\Big(\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\Big) \leq z\right] \xrightarrow{n\to\infty} \Phi(z),$$

where Φ is the cumulative distribution function of a standard normal random variable.

Proof. The proof is quite involved and will not be given here. Under some additional conditions one can give a shorter proof with the help of generating functions, cf. Ross[1]. □

7.4 The Poisson Limit Theorem

The central limit theorem is a general theorem that holds for all random variables, as long as the conditions are satisfied – in particular it is also applicable for the Bernoulli distribution. As the sum of independent Bernoulli random variables is a Binomial distribution , it is also a limit theorem for $Y_n \sim \mathcal{B}(n, p)$ binomial distributed random variables (in this case it is also known as *de Moivre - Laplace theorem.*

For binomial distributed random variables, there exists also a different limit theorem which does not describe the case of constant p, but of probabilities p_n such that $n \cdot p_n$ converges.

7.5 Theorem. *Let $X_1, \ldots, X_n$ be a sequence of $\mathcal{B}(n, p_n)$ distributed random variables such that $n \cdot p_n \to \lambda > 0$ for $n \to \infty$. Then it holds that*

$$\lim_{n\to\infty} \mathbb{P}[X_n = k] = e^{-\lambda}\frac{\lambda^k}{k!}.$$

As $\mathbb{P}[Z = k] = e^{-\lambda}\frac{\lambda^k}{k!}$ is the probability mass function of a Poisson random variable $Poi(\lambda)$, we see that the limiting distribution of the binomial distribution is a Poisson distribution .

[1] Ross, *A First Course in Probability*, Section 8.3, Section 7.7.

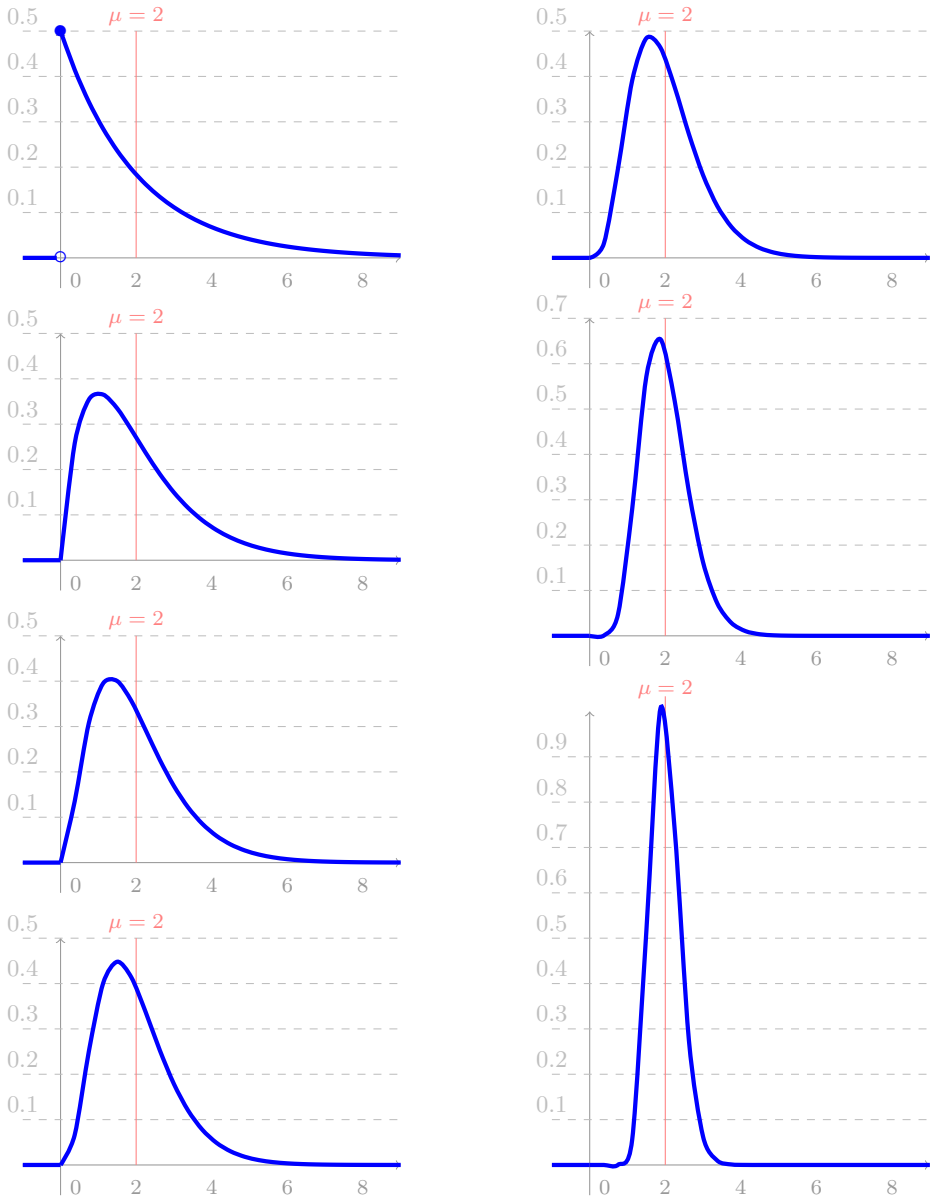


Figure 7.2: *Law of Large Numbers* for $n = 1, 2, 3, 4, 5, 10$ and 25 exponentially distributed random variables with parameter $\lambda = 0.5$. The weighted sum of the exponential random variables converges to $\mu = 2$.

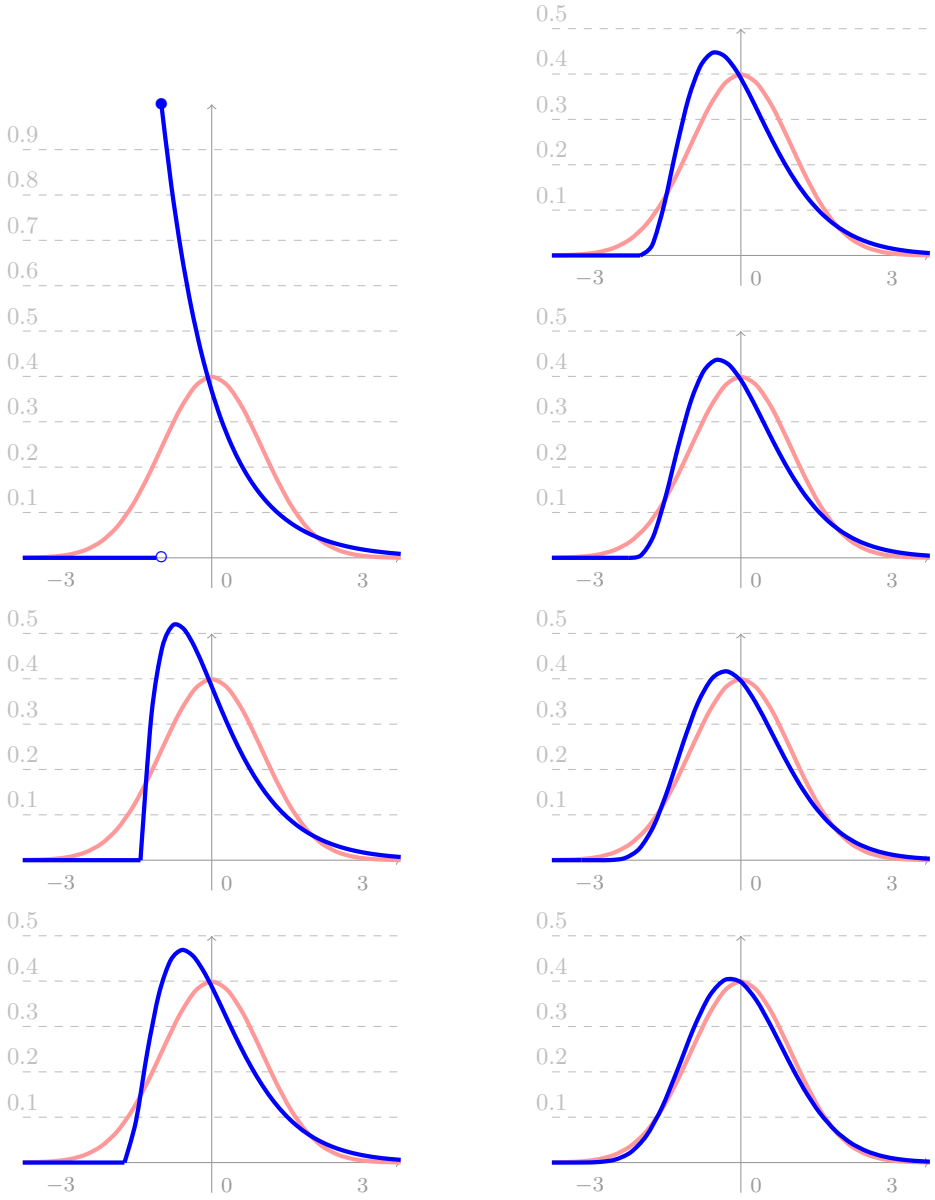


Figure 7.3: *Central Limit Theorem* for independent exponentially distributed random variables X_j with parameter $\lambda = 0.5$. We plot the density of the sum of $\frac{\sqrt{n}}{\sigma}(\sum_{k=1}^{n} X_k - \mu)$ for $n = 1$, 2, 3, 4, 5, 10 and 25 (in blue) in comparison with the density φ of a standard normal distribution (in light red).

Proof. We start by reformulating the probabilities by setting $\lambda_n = p_n \cdot n$. Then we have

$$\begin{aligned}
&\mathbb{P}\big[X_n = k\big] \\
&= \binom{n}{k} p_n^k (1-p_n)^{n-k} = \binom{n}{k} \Big(\frac{\lambda_n}{n}\Big)^k \cdot \Big(1 - \frac{\lambda_n}{n}\Big)^{n-k} \\
&= \frac{n \cdot (n-1) \cdot \ldots \cdot n-k+1}{k!} \Big(\frac{\lambda_n}{n}\Big)^k \cdot \Big(1 - \frac{\lambda_n}{n}\Big)^{n} \cdot \Big(1 - \frac{\lambda_n}{n}\Big)^{-k} \\
&= \frac{n \cdot (n-1) \cdot \ldots \cdot n-k+1}{n^k} \frac{\lambda_n^k}{k!} \cdot \Big(1 - \frac{\lambda_n}{n}\Big)^{n} \cdot \Big(1 - \frac{\lambda_n}{n}\Big)^{-k} \\
&= 1 \cdot \underbrace{\Big(1 - \frac{1}{n}\Big)}_{\to 1} \cdot \ldots \cdot \underbrace{\Big(1 - \frac{k-1}{n}\Big)}_{\to 1} \cdot \underbrace{\frac{\lambda_n^k}{k!}}_{\to \frac{\lambda^k}{k!}} \cdot \underbrace{\underbrace{\Big(\Big(1 - \frac{\lambda_n}{n}\Big)^{-\frac{n}{\lambda_n}}\Big)^{-\lambda_n}}_{\to e} \cdot \underbrace{\Big(1 - \frac{\lambda_n}{n}\Big)^{-k}}_{\to 1}}_{\to e^{-\lambda}} \\
&\xrightarrow{n \to \infty} \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}$$

as $\lambda_n \to \lambda$ for $n \to \infty$ as well as $(1+\frac{1}{x})^x \to e$ for $x \to \infty$ (with $x = -\frac{n}{\lambda_n}$). $\square$

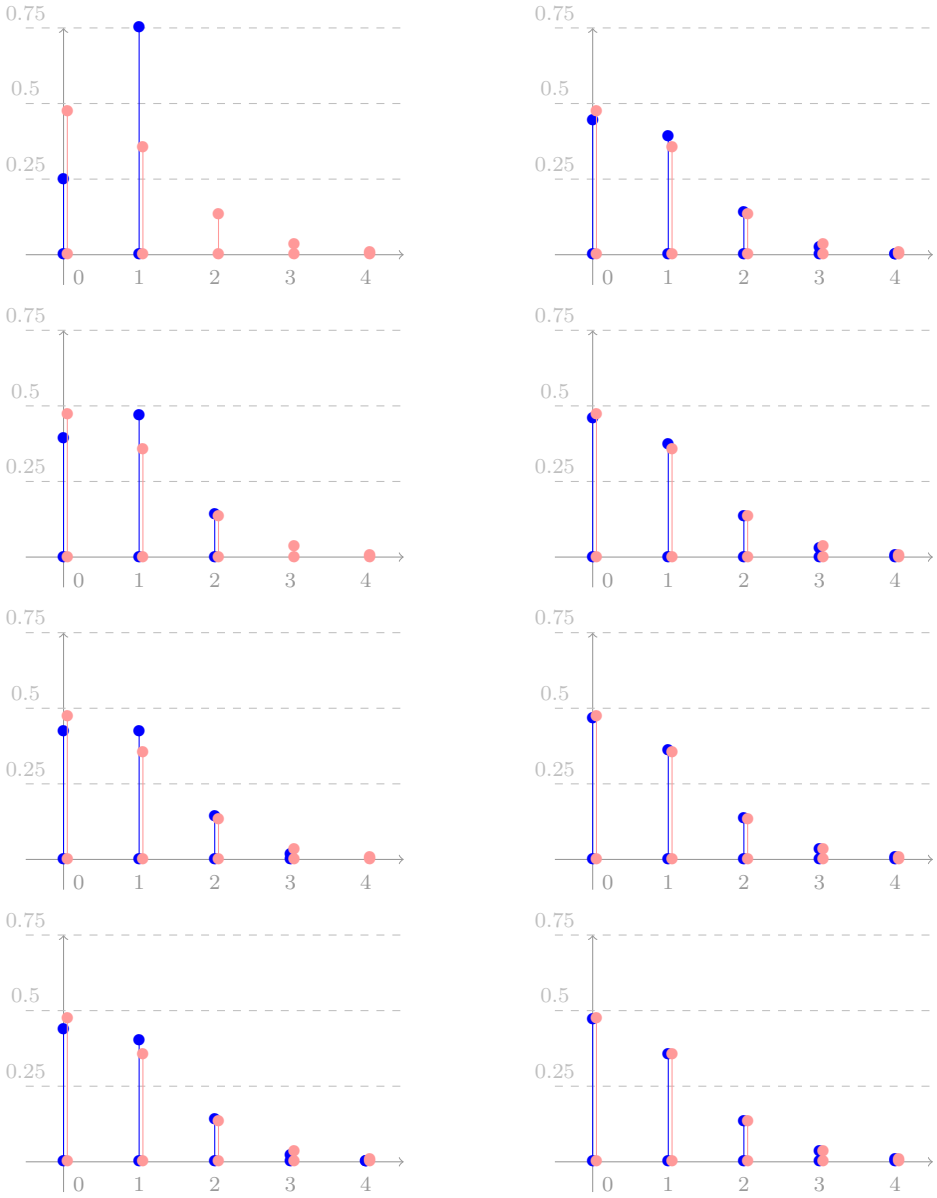


Figure 7.4: *Poisson Limit Theorem* for $n = 1, 2, 3, 4, 5, 10, 25$ and 100 Binomial distributed random variables with success probability $p_n = \frac{0.75}{n}$. The Binomial random variables converge to a Poisson distribution with parameter $\lambda = 0.75$ (in light red). Compare this to the central limit theorem, Figure 7.1.

Index