Contents Reader "Probability Theory for Engineers"

6. Continuous random variables

7. Two or more continuous variables

8. Waiting times

Formula sheet Probability Theory for BIT and TCS in module 4

Distribution	$\boldsymbol{E(X)}$	$\boldsymbol{var(X)}$
Geometric	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hypergeometric	$n \cdot \frac{R}{N}$	$n \cdot \frac{R}{N} \cdot \frac{N-R}{N} \cdot \frac{N-n}{N-1}$
Poisson $P(X=x) = \frac{e^{-\mu}\mu^x}{x!}, x = 0, 1, 2, \ldots$	μ	μ
Uniform on (a, b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Erlang $f_X(x) = \frac{\lambda(\lambda x)^{n-1}e^{-\lambda x}}{(n-1)!}, x \geq 0$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$

$$var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) + \sum_{i \neq j} \sum cov(X_i, X_j)$$

Formula sheet Probability Theory for BA-IEM in module 1

Distribution	$\boldsymbol{E(X)}$	$\boldsymbol{var(X)}$
Geometric	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hypergeometric	$n \cdot \frac{R}{N}$	$n \cdot \frac{R}{N} \cdot \frac{N-R}{N} \cdot \frac{N-n}{N-1}$
Poisson $P(X=x) = \frac{\mu^x e^{-\mu}}{x!}, x = 0, 1, 2, \ldots$	μ	μ
Exponential	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Uniform on (a, b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

==========

Formula sheet Probability Theory for BA-IEM in module 3

$E(X) = \sum_x xP(X=x)$ and $var(X) = E(X-\mu)^2 = E(X^2) - (EX)^2$

$E(aX+b) = aE(X) + b$ and $var(aX+b) = a^2 var(X)$

$$var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) + \sum_{i \neq j} \sum cov(X_i, X_j)$$

Distribution	**Probability / density function**	$\boldsymbol{\mu = E(X)}$	$\boldsymbol{\sigma^2 = var(X)}$
Binomial $B(n,p)$	$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0,\ 1,\ 2,\ \ldots, n$	np	$np(1-p)$
Geometric (p)	$P(X=x) = (1-p)^{x-1}p$, $x = 1,\ 2,\ 3,\ \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson (μ)	$P(X=x) = \frac{\mu^x e^{-\mu}}{x!}$, $x = 0,1,2,\ldots$	μ	μ
Uniform $U(a,b)$	$f(x) = \frac{1}{b-a}$, $a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential $Exp(\lambda)$	$f(x) = \lambda e^{-\lambda x}$, $x \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Chapter 1 Experiment, sample space and probability

1.1 Experiment and sample space

Probability theory deals with (mathematical) models for describing **experiments** in which **chance** plays a role. One could think of rolling a dice or measuring the life time of a certain light bulb. In these cases it is clear what steps should be made to attain an outcome of the experiment. However, the outcome itself is not fixed in advance, it is known only after the experiment has been executed. Such experiments are called stochastic.

An experiment is **probabilistic** or **stochastic** if the experiment does not necessarily lead to the same outcome when it is repeated under equal conditions.
In this reader if we say "**experiment**" we mean a stochastic experiment.

Each experiment always has a result: the **outcome** of the experiment. When rolling a dice the outcome is the face up number; when observing the life time of a light bulb the outcome is the observed life time, a positive real number; launching a satellite could have either the outcome 'success' or 'failure'.

Although the outcome of a stochastic experiment is not known in advance, we can establish all possible outcomes. These **possible** outcomes are inseparably related to the experiment and are, in contrast with the outcome itself, also fixed before the experiment is conducted. The set of all possible outcomes is called the sample space of the experiment. This set is usually indicated by S.

Definition 1.1.1 The **sample space** S of an experiment is the set of all possible outcomes.

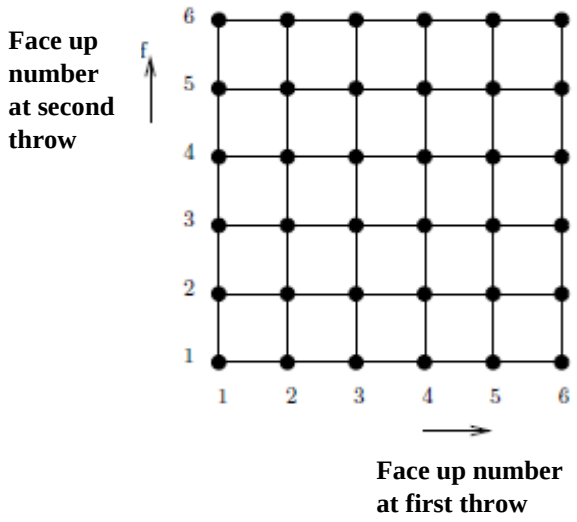
Example 1.1.2 When we toss a coin and record the face up side, there are two possible outcomes H (heads) and T (tails). So we choose as sample space the set $S = \{H, T\}$. ■

Example 1.1.3 Ten chips are selected from a batch of 1000. These 10 chips are all tested and either approved or disapproved. The outcome of this experiment could be the number of approved chips and thus the sample space is $S = \{0, 1, \ldots, 10\}$. ■

Example 1.1.4 We roll a dice twice. The outcome of the experiment is a pair of numbers, in which the first number is the face up number of the first roll and the second number is the face

up number of the second roll. The sample space can be defined as

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



or

$S = \{(i,j)|i = 1,2,\ldots,6 \text{ and } j = 1,2,\ldots,6\}$

and clearly contains 6× 6 = 36 outcomes. Graphically one could see the sample space as shown in the graph. ■

Example 1.1.5 We flip a coin until it lands tails up for the first time. The required number of tosses is the outcome of the experiment and the sample space is $S = \{1,2,3.\ \ldots\}$.
This sample space is not finite as in the previous examples, but it is **countable**, meaning that we can use the natural numbers 1, 2, 3, … to number all of the outcomes. ■

Example 1.1.6 The life time of a light bulb can be seen as the outcome of an experiment. Since the light bulb can break down at any moment, we can take $S = [0,\infty)$ as sample space. This sample space, an interval of real numbers, is not countable. ■

Experiments give rise to certain events. For instance, for the 10 chips in example 1.1.3, we could be interested in the event that more than half of the chips are disapproved. The event "more than half disapproved" occurs when the outcome is 0, 1, 2, 3 or 4.
Therefore we identify this event as the subset $A = \{0,1,2,3,4\}$ of the sample space S.
We will say that "A occurs", if the outcome of the experiment is an element of A.

Definition 1.1.7 An **event** is a subset of the sample space S.

Just like subsets, events are usually denoted by capitals $A, B, C, \ldots$.
In addition the empty set $\emptyset$ and S are events as well. The empty set $\emptyset$ is called the **impossible event**, since it contains none of the outcomes and thus this event never occurs.
The set S is called the **certain event** since every outcome is in S and thus this event will always occur. An event which consists of a single outcome s is called an **elementary event**: $\{s\}$. Both s and $\{s\}$ are sometimes referred to as a "**sample point".**

Example 1.1.8 $S = \{(i,j)|i = 1,2,\ldots,6 \text{ and } j = 1,2,\ldots,6\}$ is the sample space belonging to the experiment where we roll a dice twice (example 1.1.4).
If A is the event where both rolls result in the same face up number, we have

$$A = \{(1,1),(2,2),\ldots,(6,6)\} = \{(i,i)|i = 1,2,\ldots,6\}.$$

The event B in which the total face up number is 5 can be given (use the graph on top of preceding page):

$$B = \{(1, 4), (2, 3), (3, 2), (4, 1)\} = \{(i, 5 - i) | i = 1, 2, 3, 4\}.$$ ■

Many concepts of set theory have a specific interpretation in probability theory, due to the fact that all events are sets. We assume these concepts to be known to the reader, but we will repeat some of them here.

If A and B are events, then:

- $\overline{A}$ is the **complement of A** or **the complementary event** of A (i.e. the event which occurs if A does not occur). Alternative notations of $\overline{A}$ are A^C, $\boldsymbol{S - A}$ and $S \backslash A$
  Pronunciation of $\overline{A}$: "not A" or "A does not occur".
- $\boldsymbol{A \cup B}$, the **union** of A and B, is the event which occurs when **at least one of the events** A and B occurs. Say: "A or B or both (occur)".
- $\boldsymbol{A \cap B}$, or AB, the **intersection** of A and B, is the event which occurs when **both $\boldsymbol{A}$ and $\boldsymbol{B}$** occur. Say "Both A and B (occur)".

Note 1.1.9: The union $\cup$ and the intersection $\cap$ play for subsets (events) the same role as the logical operators $\wedge$ ("and") and $\vee$ ("or") for the elements of sets:
$A \cup B = \{s \in S | s \in A \vee s \in B\}$ and $A \cap B = \{s \in S | s \in A \wedge s \in B\}$.
If there is an expression with both $\cup$ and $\cap$, then as a rule the intersection should be performed first. So $A \cup BC$ means $A \cup (B \cap C)$.

Furthermore: $A \subset B$ (A is a **subset** of B): A implies B, i.e., if A occurs, then so does B.
(In this course we do not distinguish $\subseteq$ and $\subset$: if $A \subset B$, then possibly $A = B$)

Definition 1.1.10 A and B are **mutually exclusive** (or **disjoint**) events if $\boldsymbol{A \cap B = \emptyset}$**.**

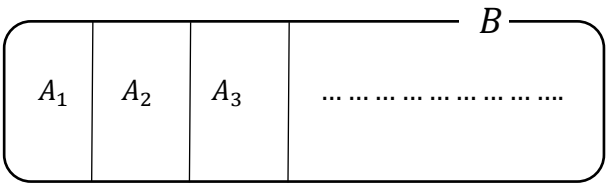
i.e., A and B cannot occur at the same time.

This definition can be extended to a sequence of events, which exists of a countable number of events A_i. Countable means either the sequence is finite $(A_1, A_2, \ldots, A_n)$ or the sequence is countable infinite $(A_1, A_2, \ldots)$. In both cases the sequence is denoted as $\{A_i\}$.

Definition 1.1.11 The events $A_1, A_2, \ldots, A_n$ or $A_1, A_2, \ldots$ are called **mutually exclusive** (or disjoint) if $A_i A_j = \emptyset$ for every possible combination (i, j) for which $i \neq j$.

If $\{A_i\}$ is a sequence of events, then we will denote $\bigcap_i A_i$ for both a finite number of n events $(\bigcap_{i=1}^{n} A_i)$ and an infinite number of events $(\bigcap_{i=1}^{\infty} A_i)$:
$\bigcap_i A_i$ is the event that occurs if each of the events A_i occurs and
$\bigcup_i A_i$ occurs if at least one of the events A_i occurs.

Definition 1.1.12 The sequence of events $\{A_i\}$ is a **partition** of the event B if the events A_i are mutually exclusive and $B = \bigcup_i A_i$

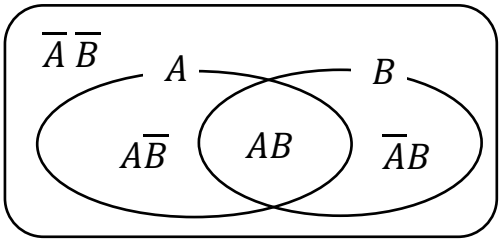


Example 1.1.13
For the purpose of a communication system we can use 32 digital "code words" 00000, 10000, 01000, …, 11111 to code the 26 letters a to z and 6 punctuation marks (. , : ; ? !).
If we choose the code word which is transmitted by the communication system at a random moment in time, then this is a stochastic experiment with sample space

$$S = \{e_1e_2e_3e_4e_5 | e_i = 1 \text{ or } e_i = 0, \text{ for } i = 1, 2, \ldots, 5\}.$$

We can now define A_i for $i = 1, 2, 3, 4, 5$ as the event that the code word has a one (1) on the i^{th} position and A_0 as the event that the code word has no ones, so $A_0 = \{00000\}$
A_1 occurs when that code word starts with 1 and $\overline{A}_1$ consists of all the code words starting with a zero.
A_1 and $\overline{A}_1$ constitute a partition of *S*.
It is clear that $S = \bigcup_{i=0}^{5} A_i$ is true (each element has either at least one 1 (if a 1 is in position i, then it is contained in A_i) or no 1's (then A_0 occurs).
But $\{A_i\}$ is not a partition, since, e.g., the code word 11000 appears in both A_1 and A_2 ($A_1A_2 \neq \emptyset$).
However, if we define B_i as the event that the randomly chosen code word contains (exactly) i ones, then $\{B_0, B_1, \ldots, B_5\}$ is a partition of S. ■

Property 1.1.14 (Properties of combinations of events)



a. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

b. $A \cup B = A \cup (\overline{A}B)$ and
$B = (AB) \cup (\overline{A}B)$.

c. (**De Morgan`s laws**)
$\overline{A \cup B} = \overline{A} \cap \overline{B}$ and
$\overline{A \cap B} = \overline{A} \cup \overline{B}$ $\left(\text{in general: } \overline{\bigcup_i A_i} = \bigcap_i \overline{A}_i \text{ and } \overline{\bigcap_i A_i} = \bigcup_i \overline{A}_i\right)$

Outline of the formal proof of these properties:
The correctness of properties 1.1.14.a and b can be shown by reasoning, using the Venn diagram above: e.g., verify the first part of 1.1.14.b, by considering $A \cup B$, which can be split up in two (mutually exclusive) events A and $\overline{A}B$. The first part of 1.1.14.c follows from the fact that $A \cup B$ occurs, if at least one of A and B occurs. Then the complement $\overline{A \cup B}$ occurs if both A and B do not occur, so if $\overline{A}$ and $\overline{B}$ occur: $\overline{A \cup B} = \overline{A} \cap \overline{B}$.
Similarly, one can verify the second equality.

1.2 Symmetric probability spaces

In the first section we described in each example a stochastic experiment and a corresponding sample space S, and we have seen that events are subsets of S. Now we want to discuss the probability that a certain event occurs, that is, the probability of an event. We want to define a function P which assigns a real number $P(A)$ to every event A of S, and will call $\boldsymbol{P(A)}$ **the probability of event** $\boldsymbol{A}$. We will call the pair $(\boldsymbol{S}, \boldsymbol{P})$ **a probability space**.
Before we give a general mathematical definition of the concept "probability", we will describe a few simple situations.

Example 1.2.1 We roll a dice (once). The face up number is the outcome of the experiment. The sample space is $S = \{1, 2, 3, 4, 5, 6\}$.
When asked what the probability is that the face up number is even, we are inclined to answer $\frac{3}{6}$, since three out of six options are even. Implicitly we assume that the dice is **fair**, i.e., every outcome is equally likely and will each occur with probability $\frac{1}{6}$.
The event $\{2, 4, 6\}$ is one of the $2^6 = 64$ possible events or subsets of S (the number of possible events, 2^6, can be found by reasoning as follows: each of the 6 outcomes **is, or is not**, an element of an event, so there are $2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 64$ different events in total). The probability of each of these events can thus be calculated by dividing the number of outcomes in the event by 6, the total number of outcomes. ■

Let us consider, in a more general setting, an experiment with finite sample space S. The total number of outcomes is called $N(S)$ and the number of outcomes of an event A is denoted by $N(A)$. We will assume the possible outcomes to be "equally likely", meaning that they occur at the same probability rate. Our definition of probability should, in such a case, satisfy the condition that the probability of a particular outcome, or elementary event, is equal to $\frac{1}{N(S)}$. This requirement is fulfilled by the probability **definition by Laplace** (1749 - 1827), as follows:

Definition 1.2.2 When the sample space S of an experiment $N(S)$ contains equally likely outcomes and the event A consists of $N(A)$ outcomes, then the **probability of an event** A, denoted by $P(A)$, equals:

$$\boldsymbol{P(A) = \frac{N(A)}{N(S)}}$$

If the event $A = \{s\}$ is an elementary event (or: a sample point), i.e., it contains only one outcome s, this definition implies that the probability of A is equal to $P(A) = \frac{N(A)}{N(S)} = \frac{1}{N(S)}$, similar to our requirement. For simplicity we denote $P(s)$ instead of $P(\{s\})$.

Definition 1.2.3 If S is a finite sample space of an experiment and the probabilities $P(A)$ of events A are defined according to Laplace's definition, the pair (S, P) is called a **symmetric probability space**.

We will now discuss some more examples of symmetric probability spaces.

Example 1.2.4 We roll a fair dice twice. The sample space is:

$$S = \{(i,j) | i = 1,2,\ldots,6 \text{ and } j = 1,2,\ldots,6\}.$$

We assume every outcome (i,j) to be equally likely.
The probability of a specific outcome, e.g. sample point $(4,2)$, is $\frac{1}{36}$
What is the probability that two rolls of the dice result in a total of 8 face up?
The event "Total face up number is 8" is the subset $A = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$ and contains 5 outcomes. So, the requested probability is $P(A) = \frac{N(A)}{N(S)} = \frac{5}{36}$. ■

Example 1.2.5 We randomly draw a card from a deck of 52 cards. Applying Laplace's definition we find P("diamonds") $= \frac{13}{52} = \frac{1}{4}$ and P("ace of diamonds") $= \frac{1}{52}$. ■

In the previous two experiments we **chose an element arbitrarily, or at random, from a finite sample space**: whenever this is the case, the definition of Laplace applies and can be used to compute probabilities.
In the following example we will show that this is not always the case, when we make a random choice from a few items, which are not all distinguishable.

Example 1.2.6
Consider a box with N marbles: R are red (indistinguishable) and $B = N - R$ are blue.
We choose at random one marble from the box. There are two possible outcomes: r, meaning that the marble is red, and b, the marble is blue. So the sample space is $S = \{r, b\}$.
Note that "choosing at random" concerns the marbles, and not the colors (the elements of S)
Therefore using Laplace's definition for this sample space S leads to the probabilities $P(r) = P(b) = \frac{1}{2}$, which, in general, is evidently incorrect (e.g. if 99 are red and 1 is blue).
However, if we would number (in our mind) the red marbles with the numbers 1 to R and the blue ones with $(R+1)$ to $(R+B) = N$, then the numbers 1, 2, ...,N are the outcomes, so now the sample space is $S' = \{1,2,\ldots,N\}$.
Choosing a marble at random means that every marble is chosen with equal probability $\frac{1}{N}$: we have a symmetric probability space.
If "RED" is the event that the chosen marble is red, then "RED" $= \{1,\ldots,R\}$ and similarly "$BLUE$" $= \{R+1,\ldots,N\}$. Thus we find:

$$P(r) = P(\text{"}RED\text{"}) = \frac{N(\text{"}RED\text{"})}{N(S`)} = \frac{R}{N} \text{ and } P(b) = P(\text{"}BLUE\text{"}) = \frac{N(\text{"}BLUE\text{"})}{N(S`)} = \frac{N-R}{N}$$ ■

In example 1.2.6 we notice that only if $R = B = \frac{1}{2}N$ the original sample space $S = \{r, b\}$ is symmetric. In all other cases a refinement of the sample space was necessary to be able to apply the probability definition by Laplace. The original outcome r is refined to the outcomes $1, 2, 3 \ldots, R$ by numbering the red marbles and similarly b to $R+1, R+2, \ldots, N$.
This example shows that one should not slothfully conclude that a probability space for an experiment is symmetric. In case of the code words from the communication system in example 1.1.13, the code words will generally not all appear equally often, since texts contain more e's than x's. One should therefore check carefully if all outcomes are equally likely. If this is indeed the case, then one determines the probability of an event A by counting the number of outcomes that A is containing. We will return to this approach in chapter 2 (on combinatorial probabilities), since counting can be more difficult than it seems at first sight.

Although we introduced the probability definition by Laplace after formulating one requirement, namely that $P(s) = \frac{1}{N(S)}$ for every outcome s, it is easy to see that this definition fulfills a number of requirements that a probability should intuitively satisfy.
For example:

Property 1.2.7 (Properties of a symmetric probability space)

- **a.** $P(A) \geq 0$ for every event A,
- **b.** $P(S) = 1$,
- **c.** if $A \subset B$, then $P(A) \leq P(B)$,
- **d.** $P(\overline{A}) = 1 - P(A)$,
- **e.** If $A_1, A_2, \ldots, A_n$ are mutually exclusive events , then $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$

Proof:
These properties follow quite directly from Laplace's definition, e.g. e. follows from it, if we use for the mutually exclusive events $A_1, A_2, \ldots, A_n$ that:

$$N\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} N(A_i)$$

From this it follows: : $P\left(\bigcup_{i=1}^{n} A_i\right) = \frac{N(\bigcup_{i=1}^{n} A_i)}{N(S)} = \frac{\sum_{i=1}^{n} N(A_i)}{N(S)} = \sum_{i=1}^{n} P(A_i)$ ■

Laplace's probability definition has two major limitations.
First: the definition assumes a finite sample space, whilst we also want to include experiments which have a sample space with infinitely many elements (examples 1.1.5 and 1.1.6).
Second: even if the sample space of an experiment has a finite number of elements, the presumption that every elementary event has the same probability certainly does not always reflect reality (as in example 1.1.13). We therefore search for a more general definition for the probability concept.

1.3 Relative frequency and the empirical law of large numbers

Example 1.3.1 We want to check whether a coin is "fair" or not, i.e., we want to find out if the probability of tails is indeed $\frac{1}{2}$. One method to "determine" the probability of tails (T) is tossing a coin very often, where both the number of tosses and the number of tails is counted. The number of "Tails" divided by the total number of tosses provides an **estimate** of the probability of tails, for example $\frac{21}{38}$ when tail was counted 21 times in 38 tosses.
When we toss the coin more often, the estimation probably becomes more accurate. ■

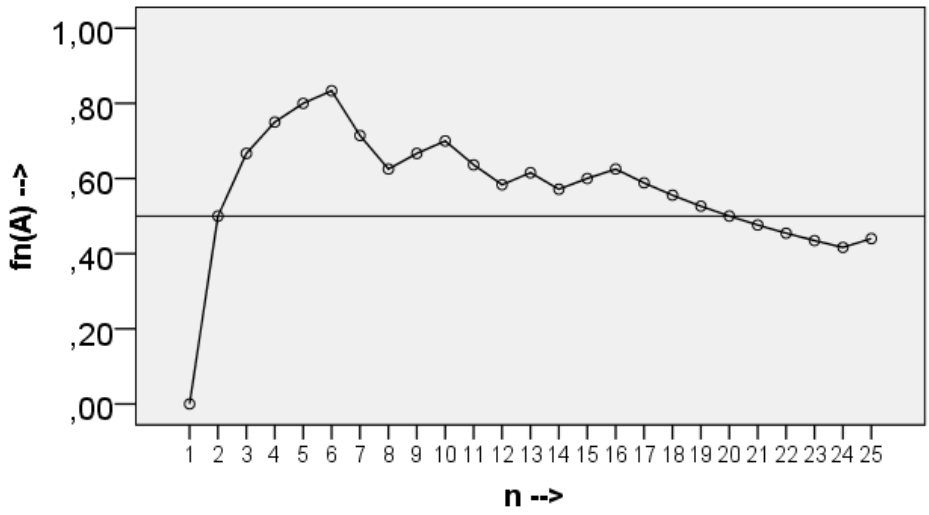
When this type of experiment is often repeated we use the concept of (relative) frequency of an event.

Definition 1.3.2
Assume that we have an experiment with sample space S which we can repeat arbitrarily often. If the event A occurred $n(A)$ times in total with n repetitions, then we define

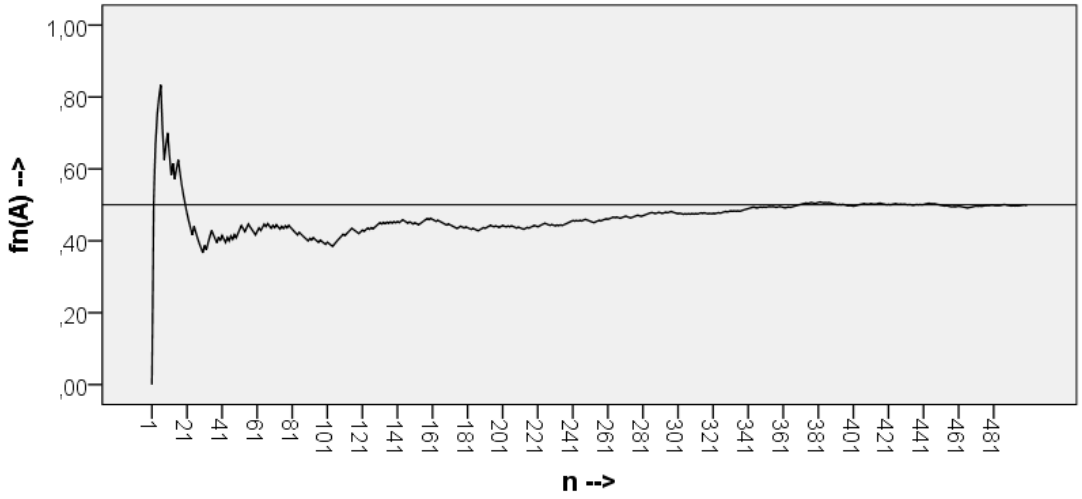
$$\boldsymbol{f_n(A) = \frac{n(A)}{n}}$$

as the **relative frequency** (or: frequency quotient**) of A** in n repetitions.

Experimentally it appears that $f_n(A)$ for increasing n 'converges' to a constant, the probability of A. This phenomenon is called the **empirical law of large numbers**. However, there is no convergence in the usual (mathematical) sense, because the outcomes with consecutive repetitions are not predictable with complete certainty: outliers are always possible, but become less probable with increasing n. We explain this by showing the graph of the relative frequency as a function of n for tossing the coin, as in example 1.3.1.



And continuing the series of tosses after $n = 25$ trials:



Intuitively we are inclined to call " $\lim_{n\to\infty}$ " $f_n(A)$ the probability of the event A.
The problem, however, is that repeating the experiment infinitely is impossible in practice. Moreover, this limit is not mathematically defined, which makes it impossible to use the relative frequency as a mathematical definition of probability.
But, since it is in line with our intuitive understanding of probabilities, we can see the properties of the frequency quotient as a guideline by developing mathematical probability theory.

The **frequency interpretation of probabilities** is based on a large number of repetitions of the experiment. These repetitions do not actually need to be executed.
When a doctor tells a patient that the success probability of surgery is 95%, he does not mean that the patient needs to undertake the surgery many times to be able to show that surgery is successful 95 out of 100 times. The necessary surgery is (hopefully) a once in a life time experience.
But we can imagine a similar thought experiment: draw a random marble from a box with 95 white and 5 black marbles. A black marble stands for failed surgery, a white marble for successful surgery. Unfortunately, only after surgery one can say which marble is drawn.

It is easy to verify that the five properties, given in 1.2.7 for a probability P, also hold for the relative frequency f_n.
It is therefore desirable that properties 1.2.7a-e are true for our future definition of probability. Our definition will be given by a few axioms which a probability has to satisfy, whereby probability is defined as a function P, that assigns a number $P(A)$ to each event A in the sample space.
The system of axioms must, of course, be such that every part of the property 1.2.7 and every other, intuitively desirable property is an axiom itself, or follows from the axioms.
A minimal set of axioms was given by Kolmogorov.

1.4 Kolmogorov`s Axioms

Definition 1.4.1 Consider an experiment with a random non-empty sample space S. A function P which assigns a real number $P(A)$ to every event $A \subset S$, is called a **probability** or **probability measure** on S if:

1. $\boldsymbol{P(A) \geq 0}$ for every event A,
2. $\boldsymbol{P(S) = 1}$ and
3. for every countable sequence of mutually exclusive events $A_1, A_2, \ldots, A_n$ or $A_1, A_2, \ldots$:

$$\boldsymbol{P\left(\bigcup_i A_i\right) = \sum_i P(A_i)}$$

Properties 1., 2. and 3. are known as the **Kolmogorov`s axioms**. By formulating these axioms in his book "*Grundbegriffe der Wahrscheinlichkeitsrechnung*" (1933) A.N. Kolmogorov (1903-1987) provided the basis for modern probability theory.

Note 1.4.2 In theory we can confine axiom 3. to a countable sequence of mutually exclusive events $A_1, A_2, \ldots$. Thus the property is also applicable for a finite sequence $A_1, A_2, \ldots, A_n$. However, we cannot confine axiom 3. to finite sequences only.
An explanation for this would require a (measure-theoretical) approach which is too theoretical for the applied character of this course and is not necessary for the applications of probability that we pursue. ■

Definition 1.4.3 When S is a sample space and P is the probability on S then we call the pair (S, P) a **probability space**.

In properties 1.2.7.a and 1.2.7.b we saw that for a symmetric probability space Laplace's definition satisfies the first two axioms. Property 1.2.7 e concerns a finite number of events A_i which exclude each other, whilst this property also holds for an infinite number of different events, according to axiom 3. of Kolmogorov. Note that a finite number of outcomes is a condition for applying Laplace's definition: in that case the number of events is finite as well.

From the axioms of Kolmogorov we can derive that property 1.2.7.c and 1.2.7.d are true as well, for every probability measure. But first we will show that the probability of the impossible event is indeed 0.

Property 1.4.4 $\boldsymbol{P(\emptyset) = 0}$.

Proof: according to axiom 3 we obtain for the mutually exclusive events A_1 and A_2:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

If we choose $A_1 = S$ and $A_2 = \emptyset$, then:
$A_1 \cap A_2 = S \cap \emptyset = \emptyset$ (so A_1 and A_2 are mutually exclusive) and $A_1 \cup A_2 = S \cup \emptyset = S$.
So $1 = P(S) = P(S \cup \emptyset) = P(S) + P(\emptyset) = 1 + P(\emptyset)$.
Or: $P(\emptyset) = 0$. ■

Property 1.4.5 (**Complement rule**) $\boldsymbol{P(\overline{A}) = 1 - P(A)}$, for every event A.

Proof: $A \cup \overline{A} = S$ and A and $\overline{A}$ are mutually exclusive, so that, according to axioms 2 and 3:
$1 = P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A})$, so $P(\overline{A}) = 1 - P(A)$ ■

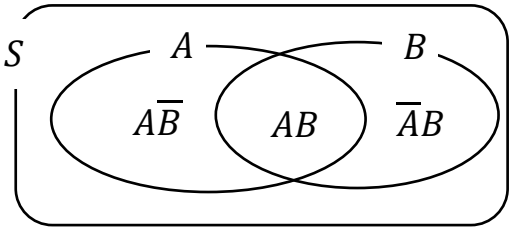
Property 1.4.6 For two events A and B with $\boldsymbol{A \subset B}$ we have: $\boldsymbol{P(A) \leq P(B)}$.

Proof: see exercise 7.

Property 1.4.7 For two events A and B (which are not necessarily mutually exclusive):

$$\boldsymbol{P(A \cup B) = P(A) + P(B) - P(A \cap B)}.$$

Proof: according to property 1.1.14.b:
$A \cup B = A \cup \overline{A}B$, where $A \cap \overline{A}B = \emptyset$ and
$B = AB \cup \overline{A}B$, where $AB \cap \overline{A}B = \emptyset$.
Using axiom 3. We find:



$$P(A \cup B) = P(A) + P(\overline{A}B)$$

and $$P(B) = P(AB) + P(\overline{A}B)$$

So $P(A \cup B) - P(B) = P(A) - P(AB)$ ■

The rule in property 1.4.7 is referred to as "**the general addition rule**".
For mutually exclusive we have the specific "**addition rule for mutually exclusive events**":
If A and B are mutually exclusive, then: $\boldsymbol{P(A \cup B) = P(A) + P(B)}$.

1.5 Exercises

1. Consider three events A, B and C. Express the following events in terms of A, B and C, using complements, unions and intersections:

a. A and B, but not C (occur)
b. All three (occur).
c. At least one of the three.
d. At least two.
e. None.
f. Exactly one of the three.
g. Not more than two.

2. We have a collection of 1200 bolts and consider the following subsets.
A = "the set of bolts with a length of 10 cm"
B = "the set of bolts with a weight of 1 ounce"
C = "the set of bolts with a diameter of 20 mm."
Furthermore, it is known that:

- 400 bolts have a length of 10 cm and a weight of 1 ounce,
- 400 bolts have a length of 10 cm and a diameter of 20 mm,
- 400 bolts have a weight of 1 ounce and a diameter of 20 mm and
- 300 bolts have a weight of 1 ounce and a diameter of 20 mm and a length of 10 cm.

Compute the probability that an arbitrary bolt (chosen at random from the population of 1200 bolts) occurs in at least two of the 3 events A, B and C.

3. We toss two coins once. One could distinguish 3 outcomes of this experiment: two Tails, two Heads and a Head and a Tail. D'Alembert (1717-1783) stated that this sample space is symmetric. Check, experimentally and by reasoning, whether you agree.

4. In his novel *Bomber* Len Deighton reported that a World War II pilot had a 2% probability to be shot down during each flight.
Therefore, he concluded he has a 40% probability to be shot down in 20 flights.
Argue whether this is correct.

5. Use a Venn diagram of the events A, B and C to express $P(A \cup B \cup C)$ in $P(A)$, $P(B)$, $P(C)$, $P(AB)$, $P(BC)$, $P(AC)$ and $P(ABC)$.
You should find a similar rule as for $P(A \cup B)$.

6. Choosing one out of many possibilities, completely at random, is not always as easy as it seems. For example, if somebody is choosing one out 4 answers (a, b, c or d) on *multiple-choice* questions at random, he usually unconsciously chooses one option more often than another. To ensure the choice is at random, an approach is to simulate the choice, e.g. using a vase with numbered balls – and choosing one at random.
Another way is using one or more dice.
How many dice does one at least need to simulate a random choice **in one roll** if we have to answer a multiple choice item with

a. 2 **b**. 4 or **c.** 5 possible answers?

7. Prove, using Kolmogorov's axioms, that from $A \subset B$ (A is a subset of B) it follows that $P(A) \leq P(B)$ (property 1.4.6). First draw A and B in a Venn diagram.

8. Given is that $P(A) = \frac{1}{2}$, $P(AB) = \frac{1}{3}$ and $P(A \cup B) = \frac{8}{9}$.
Compute $P(B)$ and $P(\overline{A} \cap \overline{B})$.

Hints for solving exercises of chapter 1.

1. Sketch a Venn diagram of the 3 events A, B and C, such that every couple of two events or all three events can occur simultaneously. From the area's you can reason how to use $A, B, C, \overline{A}, \overline{B}, \overline{C}$ and e.g. $AB\overline{C}$.
2. See 1.
3. Instead of 2 coins at once we could also flip one coin twice: which outcomes can you distinguish now? Are they equally likely?
4. What is the probability **not** to be shot down in one flight?
   What is the probability to be shot down in 2 flights? And in 20 flights.
5. See 1.
6. –
7. Consider the Venn diagram, in which A is included in B: How can you define the part of B which is not in A? So: $B = A \cup$ .....
8. Use a Venn diagram for A and B: in which (disjoint) parts can you split up B?
   And $A \cup B$? Use the Venn diagram to relate the known probabilities to the unknown.

Chapter 2 Combinatorial Probability

2.1 Theory and examples

Many problems in probability theory can be solved by using the probability definition by Laplace, if the assumption of "equally likely" outcomes is justified.
With this definition one determines the probability of event A by computing the proportion of the number of elements of *A (the favorable number)* and the total number of possible outcomes:

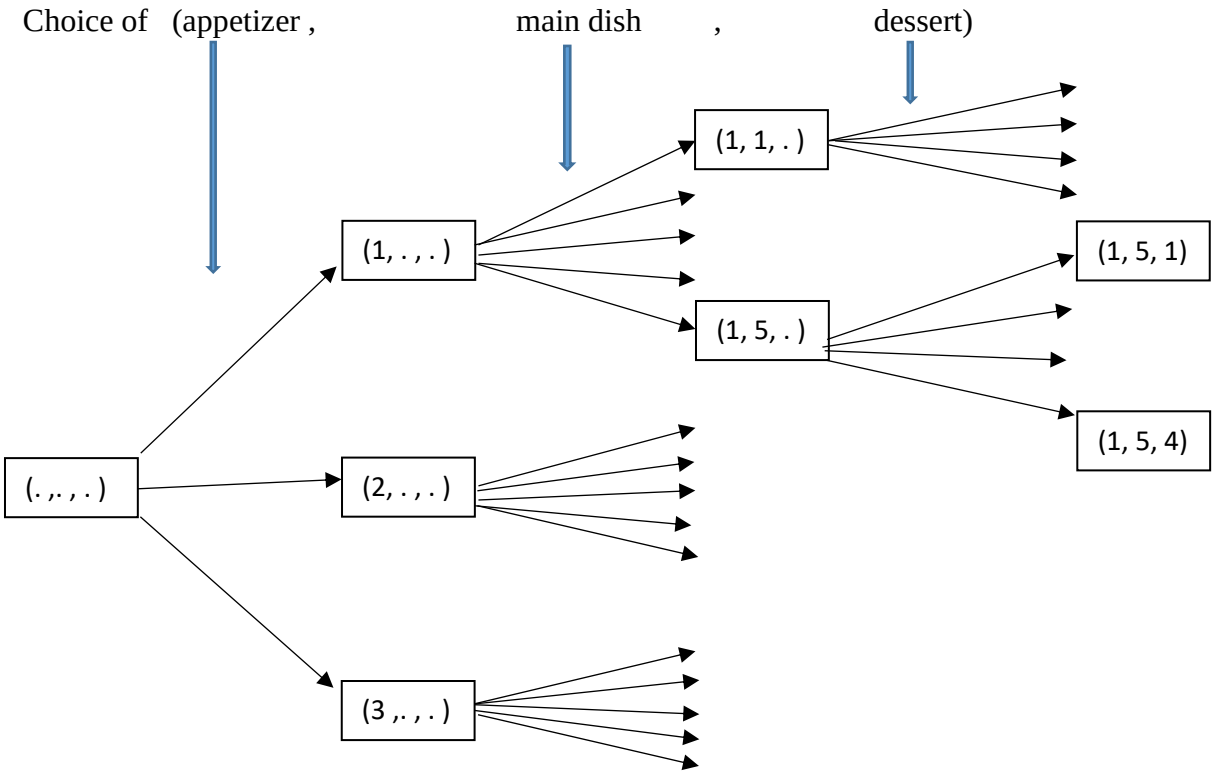
$$P(A) = \frac{N(A)}{N(S)} = \frac{\text{"favorable number"}}{\text{"total number"}}$$

As a consequence we can find probabilities by using the theory of counting: the "art of counting" is called **Combinatorics**. We will show some important results in combinatorics, mostly by applying it to several applications.

Example 2.1.1 A menu contains 3 appetizers, 5 main courses and 4 desserts.
A man chooses a three course menu randomly. What is the probability that he has chosen exactly the same menu as his wife did?
From the description we know that the set of possible menus from which the man chooses is a symmetric probability space. There are 3×5×4 possible menus. The probability that the man chooses the exact same menu is $\frac{1}{3\times5\times4} \approx 1.7\%$
(*note: "≈" means "is approximately equal to", usually we will give the final answers of probabilities in 3 decimals or as a percentage in one decimal*)
Only few people will have trouble understanding this calculation, but you might wonder **why** we have to multiply the numbers of appetizers, main courses and desserts (3×5×4), or: why don`t we, e.g., add the numbers. The reasoning is as follows:
for every appetizer one can choose 5 main courses, which results in 3 times 5 different combinations of an appetizer and main course. For each of these combinations one can choose 4 different desserts, and thus: (3×5)×4 different menus.
Every outcome of a menu consists of an ordered threesome (appetizer, main course, dessert). So when we number each of the courses we find the sample space of this experiment:

$$S = \{(i,j,k) | i = 1,2,3 \text{ and } j = 1,\ldots,5 \text{ and } k = 1,\ldots,4\}.$$

One can visualize the calculation of the number of menus in a "decision tree". The first branching gives the choice of appetizer, the second gives the main course and the third gives the dessert. The number of end-points is the total number of menus, as can be seen in the following diagram.



3 possibilities 3×5 couples 3×5×4 menus ■

In example 2.1.1 the experiment "choosing a random menu" was split up in three partial experiments for choosing an appetizer, a main course and a dessert, respectively. Every partial experiment has a fixed number of possible outcomes, independent of the results of the previous partial experiments. More general:

Property 2.1.2 (**The product rule**)
If an experiment consists of performing k partial experiments and the i^{th} partial experiment has n_i possible outcomes ($i = 1, \ldots k$), no matter what the results of the partial experiments are, then $\boldsymbol{n_1 \times n_2 \times \ldots \times n_k}$ outcomes of the total experiment are possible.

This rule is easily proven by using induction on k, the number of partial experiments.

Example 2.1.3 A soccer match in the Champions League has to be decided by taking 5 penalties. Coach T. Rainer has chosen his penalty-specialists in advance and chooses the order by letting his assistant-trainer draw tickets with the names out of his cap. There are 5 options when drawing the first ticket for shooting the first penalty, 4 options for the second penalty, etc. There are thus 5 partial experiments with resp. 5, 4, 3, 2 and 1 outcomes, which leads to 5×4×3×2×1 = 5! possible orders according to the product rule,
as is illustrated by an example of one specific order:
(5! is pronounced as "5 **factorial**".)

position	*1*	*2*	*3*	*4*	*5*
outcome	3	2	5	4	1

The probability for one order to occur in this symmetric probability space is thus

$$\frac{1}{5!} \approx 0.83\%.$$ ■

With most of the (simple) calculators you can use the *x!*-button *(shift-x^{-1})* to calculate 5!
Generalizing this example we find the number of orders of k different elements:

position	*1*	*2*	……	*k*
outcome			……..	

Property 2.1.4 (**The permutation rule**)
The number of orders or **permutations** in which k different things can be arranged is $\boldsymbol{k!}$

The next example shows that one should avoid "blindly" multiplying numbers.

Example 2.1.5 We want to determine the probability that an arbitrary 3-digit number has the digit 2 as lowest digit. Numbers less than 100 can be interpreted as 3 digit numbers as well: e.g. 28 = 028. In this way there are 1000 of these numbers: 0, 1, 2, …, 999.
If we choose one of these numbers **at random**, we have 1000 equally likely outcomes.
We can compute the probability of $A =$ "the lowest digit in the 3-digits number is 2." by counting the number of elements of A and dividing by 1000.
The event A occurs if all digits are at least 2: every digit is 2, 3, 4, 5, 6, 7, 8 or 9.
But $\boldsymbol{N(A) \neq 8 \times 8 \times 8}$, since the multiplication rule does not apply: if the first two digits are both 2, there are 8 possibilities to choose the third, but if the first two digits are both 4 the third digit has to be 2. Evidently, the number of possibilities for drawing the three digits are dependent!
Another false approach to determine $N(A)$: first choose the position of the 2 that should be in the number: if A_i is the event that the arbitrary 3 digit number has a 2 in position i (for $i = 1$, 2, 3) and on the other positions digits at least 2.Then the following is true:

$$N(A_1) = 1 \times 8 \times 8 = N(A_2) = N(A_3) \quad \text{and} \quad A = \bigcup_{i=1}^{3} A_i$$

$$\text{But:} \quad N(A) \neq \sum_{i=1}^{3} N(A_i) = 3 \times 8 \times 8$$

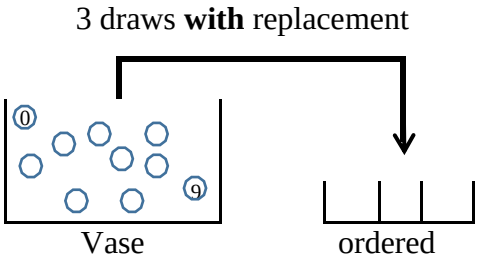
We cannot compute $N(A)$ in this way, because A_1, A_2 and A_3 are not mutually exclusive, because, e.g., $272 \in A_1 \cap A_3$. (272 is "counted double" in the total number).
A correct solution for $P(A)$ should be given as an answer to exercise 2. ■

Counting by considering sub-events is only useful, if those sub-events are a **partition** of the original event. In that case we can use the following property (which we also used in the proof of property 1.2.7).

Property 2.1.6 When $A_1, A_2, \ldots, A_k$ are **mutually exclusive** events, then:

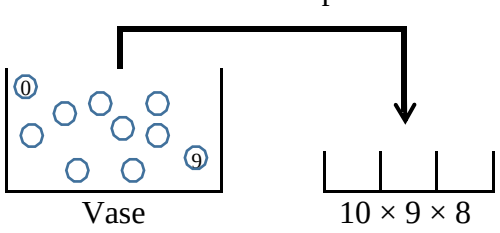
$$\boldsymbol{N\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} N(A_i)}$$

Example 2.1.7 What is the probability that a random 3-digit number contains the number 2 once? As we saw in example 2.1.1, this is a symmetric probability space with $N(S) = 1000$ possible outcomes. We can generate a random 3-digit number by drawing balls randomly from a vase with 10 balls (numbered 0 to 9).
The first ball drawn determines the value of the first digit. We return the drawn ball to the vase and, after shaking the vase properly, we repeat the experiment a second, and a third time, to determine the second and third digit. Each of these partial experiments has 10 possible outcomes, which leads to $N(S) = 10 \times 10 \times 10 = 1000$ possible 3-digit numbers.



When A is the event that a 3-digit number has exactly one 2, we can determine $N(A)$ by first choosing the position of the 2 (position 1, 2 or 3), and then choosing both other digits to be unequal to 2. According to the product rule this can be done in $N(A) = 3 \times 9 \times 9$ ways.
So $P(A) = \frac{N(A)}{N(S)} = 0.243$. ■

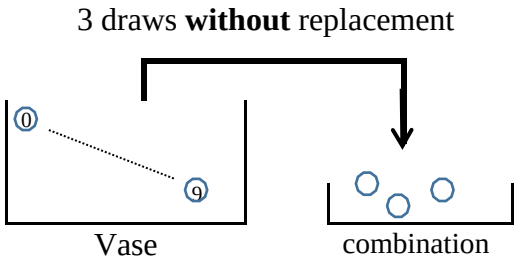
Example 2.1.8 A factory makes combination locks with 3 different digits. Each code occurs equally often. A bike-thief needs 3 seconds to check whether a code is correct or not.
What is the probability of the event A that the thief opens the lock within 5 minutes?
We can generate a random combination lock by drawing 3 balls from a vase with ten balls numbered 0 up to 9, for resp. the first, second and third number of the code. To prevent the repetition of numbers the drawn balls are not put back in the vase, so: after every draw there is one ball less in the vase ("draws without replacement").
See the illustration.



So, in the first draw there are 10 options, in the second draw there are 9 options, and in the third draw there are 8 options, regardless of the outcomes of the previous draws.
According to the product rule we get $N(S) = 10 \times 9 \times 8$ different locks, which all occur with probability $\frac{1}{10\times9\times8}$. $N(S) = 10 \times 9 \times 8$ can be written as $\frac{10!}{7!}$ or $\frac{10!}{(10-3)!}$ and is called the number of **permutations (**or **variations)** of 3 out of 10.
Most of the simple calculators have a nPr-button: type 10 nPr 3 to calculate $\frac{10!}{(10-3)!}$.
Conclusion: A thief can try
20 (different) codes in one minute, so the probability $P(A)$ that he can open a combination lock in five minutes is $\frac{N(A)}{N(S)} = \frac{5\times20}{10\times9\times8} \approx 0.139$ ■

When simulating probability experiments we often use the so called **vase model**.
In the examples above we drew a ball from a vase with balls numbered 0 to 9 three times, once **with replacement** (example 2.1.7) and once **without replacement** (example 2.1.8).
In both cases we considered the **ordered** three-digit numbers as outcomes.
Each different order gives a different number or code.
The two cases in which we consider **unordered** three-digit numbers are treated in the following two examples.

Example 2.1.9 We draw three balls from an urn with 10 balls numbered 0 up to 9, randomly and without replacement.
Determine the probability of the event A that the sum of the drawn numbers is higher than 5.
The order of drawing balls is not important, the sample space S consists of combinations or subsets of 3 elements from the set $\{0, 1, \ldots, 9\}$.
See the illustration with the vase model.



When we consider one subset, e.g. $\{2, 5, 6\}$, we can order the three numbers in 3! orders or make 3! different combination locks as in the previous example. This is true for all subsets, so there are 3! times as many combination locks as there are subsets.
The number of subsets is thus the number of combination locks divided by 3!:

$$N(S) = \frac{10 \times 9 \times 8}{3!} = \frac{10!}{3!\,7!} = \binom{10}{3}$$

Most of the simple calculators have a nCr-button: type 10 nCr 3 to calculate $\binom{10}{3}$, pronounced as "**10 choose 3**".
Since the combination locks, so the ordered 3-digit numbers whilst drawing without replacement, are a symmetric probability space, it is also the case for corresponding unordered outcomes (the subsets).
Subsequently, the probability that the sum is greater than 5 can be determined as follows, using the complement rule for the event A:

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{N(\overline{A})}{N(S)} = 1 - \frac{4}{120} \approx 0.967,$$

because $\overline{A}$ consists of the subsets $\{0, 1, 2\}$, $\{0, 1, 3\}$, $\{0, 1, 4\}$ and $\{0, 2, 3\}$. ■

The number $\binom{10}{3}$ determines the number of subsets (or combinations) of 3 out of 10, but is also the number of (ordered!) sequences with 3 ones and 7 zeros, or the number of terms a^3b^7 which appears while working out the binomium $(a + b)^{10}$.
Due to the last observation $\binom{10}{3}$ is also called the **binomial coefficient**.

In example 2.1.9 we could have drawn three balls from the vase at the same time. This is equivalent to the experiment of drawing one ball, three times, without replacement and without taking into account the order in which they are drawn.

Example 2.1.10 We draw three balls, randomly and with replacement, from a vase with ten balls numbered 0 to 9.
Determine the probability of event A that exactly one draw results in a 2.
The order of drawing is not important, so we choose as sample space S the set of all **unordered** threesomes.
When we draw a 2 twice and an 8 once we will denote the outcome as 2 2 8.
So, 2 2 8 = 2 8 2, etc. (The order can be discarded).

$$A = \{2\,i\,j \mid i, j = 0, 1, 3, \ldots, 9\}.$$

But the sample space S of all unordered threesomes out of 10 (with possibly equal digits) is not symmetric, and as a consequence of this observation, we cannot apply Laplace's formula: $P(A) \neq \frac{N(A)}{N(S)}$, in general.
For example 2 2 8 will occur more often than 2 2 2: one can reason this inequality of the probabilities after refinement of S to a new sample space $S`$, consisting of all **ordered** threesomes: now $S`$ does constitute a symmetric probability space, as in example 2.1.7.
$2\ 2\ 8 \in S$ is refined to the event $\{(2,2,8),(2,8,2),(8,2,2)\} \subset S`$, and $2\ 2\ 2 \in S$ to one sample point $(2,2,2)$ in $S`$. So $P(2\ 2\ 8) = 3 \cdot P(2\ 2\ 2)$.
If we define $A`$ as the event that an ordered outcome contains one 2, we find:

$$P(A) = P(A`) = \frac{N(A`)}{N(S`)} = \frac{3 \times 9 \times 9}{10 \times 10 \times 10} = 0.243. \qquad \blacksquare$$

In examples 2.1.7-10 we have seen 4 different ways of randomly choosing three out of a set of 10 different elements. Only in the last case probability spaces are not symmetric. In the other three cases we can directly determine probabilities of events with the probability definition by Laplace. An overview of the methods and numbers is given in the table below.

		Method of drawing	
		Without replacement	With replacement
Kind of outcome	ordered	**Permutations** of 3 out of 10: $S =$ {numbers with 3 different digits} $\boldsymbol{N(S) = 10 \times 9 \times 8 = \frac{10!}{7!}}$	$S =$ { numbers with, possibly the same, 3 digits} $\boldsymbol{N(S) = 10^3}$
	unordered	**Combinations** of 3 out of 10: $S =$ {subsets with 3 different digits} $\boldsymbol{N(S) = \binom{10}{3}}$	$S =$ {Unordered threesomes with repetitions} $\left(N(S) = \binom{12}{3}\right)$ (S, P) is **not symmetric**: refine the sample space to the corresponding symmetric probability space of ordered threesomes.

In the last case we have not determined the number of elements of the sample space *S*. This is not important for determining probabilities, as we have seen in example 2.1.10. We can generalize these four cases, when drawing k times randomly out of n different elements.

Property 2.1.11 We randomly draw *k* times from a set of n different elements, then in the following three cases the probability space is **symmetric**:

- **a.** Draw with replacement, ordered outcomes: $N(S) = \boldsymbol{n^k}$.
- **b.** Draw without replacement, ordered outcomes: (**variations/permutations of $\boldsymbol{k}$ out of $\boldsymbol{n}$**): $N(S) = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{\boldsymbol{n!}}{\boldsymbol{(n-k)!}}$
- **c.** Draw without replacement, unordered outcomes: (**combinations of $\boldsymbol{k}$ out of $\boldsymbol{n}$**).

$$N(S) = \binom{\boldsymbol{n}}{\boldsymbol{k}}$$

And in the last case **non-symmetric**:

- **d.** If the draws are with replacement and unordered outcomes, the probability space is non-symmetric. We can redefine a symmetric probability space by considering the corresponding ordered outcomes (transferring to case a.).

Outline of the proof: the proof of the numbers in a, b and c is analogue to the derivation given in examples 2.1.7 to 2.1.9, for the case where $k = 3$ and $n = 10$. In the same way we can determine the number of combinations of k out of n (case c) from the number $\frac{n!}{(n-k)!}$ of permutations (case b) of k out of n: every combination of k out of n can be written in $k!$ orders, permutations with of the same k elements.
So there are $\frac{n!}{(n-k)!}/k! = \binom{n}{k}$ combinations *k* out of *n*. Moreover, it follows that when the permutations form a symmetric probability space, this is also true for the combinations. ■

Note 2.1.12 We return to example 2.1.10 where we considered the unordered threesomes, possibly with repetition (such as 2 2 8, meaning that 2 draws resulted in a 2 and one in 8). Though in this case (d. in property 2.1.11) the number of outcomes cannot be used to compute probabilities, we will give the number of these outcomes, $\binom{12}{3}$, and the reasoning to find it: every combination of 3 out of 10 with repetition can be represented by a sequence of 9 ones and 3 zeros. For example, 2 2 8 is represented by 110011111101: no 0 to the left of the first 1 and no 0 between the first two 1`s mean "no 0 and 1 in the three draws". Two 0`s between the second and third 1 mean "two 2's in the three draws". The last 0 is between the eighth and the ninth 1, meaning an "8 in the three draws".
For each result of the draws there is one unique order of 9 ones and 3 zero`s, and conversely. So the total number of outcomes is equal to the total number of orders of 9 ones and 3 zeros: $\binom{12}{3}$. In general: the total number of combinations with repetition of k out of n is $\binom{n+k-1}{k}$ ■

Simple calculators can compute the value of the number of permutations (usually with the nPr-button) and combinations (nCr-button), but for small numbers we can easily compute them "by hand": $\binom{10}{3} = \frac{10!}{3!7!} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 5 \cdot 3 \cdot 8 = 120$, just by simplifying.
There is only one subset of 10 out of 10 (S it self): $\binom{10}{10} = \frac{10!}{10!0!} = 1$, if $0! = 1$.
That is why we have to define $\mathbf{0! = 1}$.
A few examples will follow in which we apply property 2.1.11 (cases a., b. and c.).

Example 2.1.13 We toss 6 (fair) coins and want to determine the probability that we get 2 T (tails) and 4 H (heads). The outcome apparently is a sextet, where the order is not important. The sample space S consists of seven (unordered) outcomes, which we denote as

$$6 \times H, 5 \times H, \ldots, 0 \times H$$

But these seven outcomes are not equally likely.
We refine the sample space by numbering the coins. The result is a symmetric sample space of $N = 2^6$ ordered sextets:

$$HHHHHH,\ HHHHHT,\ HHHHTH,\quad \ldots\ldots\ldots\ldots\quad ,\ TTTTTT$$

For each of the 6 positions an "H" or a "T" can be chosen so in total 2^6 different sextets.
The probability of A = "$2 \times H$" is determined by the numbers of sequences with 2 H's and 4 T's and the total number of ordered sextets. These computations can be illustrated as follows:

position 1 2 3 4 5 6 outcome: ☐☐☐☐☐☐ On every position a H or T could be chosen: 2 possibilities for every position, giving a total of 2×2×2×2×2×2 possibilities.	Example with 2 H`s (and consequently 4 $T$`s): *position* 1 2 3 4 5 6 outcome: *T* *H* *T* *T* *T* *H* The total number of these sequences equals the number of combinations of 2 positions for the H`s chosen out of 6, or de 4 positions out of 6 for the $T$`s: $\binom{6}{2} = \binom{6}{4}$

$$\text{So } P(A) = \frac{N(A)}{N(S)} = \frac{\binom{6}{2}}{2^6} = \frac{15}{64} \approx 0.234$$ ■

Example 2.1.14 What is the probability that 2 friends are selected in the same group if they are included in a total group of 15 persons, which is randomly distributed in two groups of 7 (group 1) and 8 persons (group 2) respectively?
In this case "randomly distributed" means that, e.g., from a box with 15 numbered marbles (each person has a unique number) 7 are drawn, at random and without replacement: these persons form group 1; the remaining 8 form group 2. The set of all possible distributions is thus a symmetric sample space (S, P) with $N(S) = \binom{15}{7}$.
One can also argue this number in the following way: we can place 15 people in $15!$ orders. For each order we assign the first 7 to group 1 and the last 8 to group 2. The order of the first 7 and last 8 people do not change the distribution in sub groups. So, every distribution in 2 groups correspond with $7! \times 8!$ orders of 15 people.
Consequently, the total number of distributions in 2 groups of 7 and 8 persons must be:

$$\frac{15!}{7!\,8!} = \binom{15}{7}$$

Furthermore we note that choosing the second group of 8 first would lead to the answer $\binom{15}{8}$, but:$\binom{15}{8} = \frac{15!}{8!7!} = \binom{15}{7}$.
The occurrence of event A, that 2 friends join the same group, can be split into two (mutually exclusive) events A_1 and A_2, where A_i is the event that two friends are in group i.
$N(A_1)$ is determined, by appointing 5 of the remaining 13 persons to group 1 and 8 to group 2, assuming that two friends are in group 1: $N(A_1) = \binom{13}{5}$. Similarly: $N(A_2) = \binom{13}{6}$.
Applying property 2 gives us the probability (which is a little smaller than 50%):

$$P(A) = \frac{N(A)}{N(S)} = \frac{N(A_1) + N(A_2)}{N(S)} = \frac{\binom{13}{5} + \binom{13}{6}}{\binom{15}{7}} \approx 46.7\%.$$ ■

In school classes of approximately 30 students it often turns out that students have the same day of birth. Is this coincidence? Or, is the probability of this event higher than, e.g., 50%? When determining such a probability we have to make assumptions with respect to the birthdays: a (probability) **model of reality**. In this case it might be reasonable to assume that every day of the year occurs at the same rate as a birthday of a randomly chosen person. The days of the years form a symmetric probability space.
If we have all the birthday information of the whole population we can check out the correctness of the assumption, but this information is not available. Thus we thus need to realize that our calculation of the probability is correct, provided that the **model** we made of the reality is correct.

Example 2.1.15 Consider a random group of n persons.
What is the probability that two or more of them have the same birthday?
For simplicity we assume nobody is born on the 29-th of February and every birthday occurs at the same rate ($\frac{1}{365}$).

Choosing a group of n persons can be seen as drawing, in a specific order and with replacement, n birthdays out of 365 days, where every outcome, so a series of n birthdays, is equally likely. We thus have a symmetric probability space with 365^n outcomes.
When A is the event that two or more of the drawn days are the same, then $\overline{A}$ is the event that all drawn days are different. Event $\overline{A}$ consists of all outcomes without repetition, so all permutations of n out of 365 days.
The number of permutations is $\frac{365!}{(365-n)!}$ so that the requested probability is:

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{365!/(365 - n)!}{365^n}$$

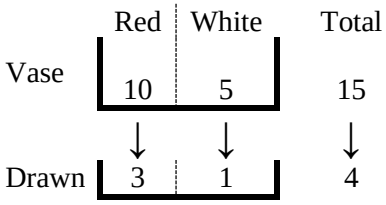
Numerical computation shows that for n = 23 persons $P(A) > \frac{1}{2}$:

n	1	2	3	4	……	22	23	24
$P(A) \approx$	0	0.0027	0.0082	0.0164		0.4757	0.5073	0.5383

■

2.2 Combinatorics and random variables

Example 2.2.1 From a vase, filled with 10 red and 5 white balls, 4 balls are drawn, randomly and without replacement.
Determine the probability that we draw exactly 3 red (and thus 1 white) balls.
We can schematically display this as follows:



To get a symmetric sample space we number the balls 1 to 15 (in our mind, to make them distinct), where the red balls get the numbers 1 to 10. The sample space then consists of $\binom{15}{4}$ unordered outcomes, combinations of 4 out of 15. The event A = "2 red and 2 white" or A= "2 red" for short, occurs when there is an (unordered) outcome with 2 numbers between 1 and 10 and 2 numbers between 11 and 15. We can choose the 3 red balls in $\binom{10}{3}$ ways, and the white ball in 5 = $\binom{5}{1}$ ways.
The requested probability is:

$$P(A) = \frac{\binom{10}{3} \cdot \binom{5}{1}}{\binom{15}{4}} \approx 44.0\%$$

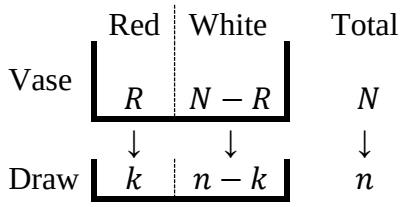
■

The expression with binomial coefficients is an example of the hypergeometric formula.

Property 2.2.2 (**hypergeometric formula**)

If we draw n times, at random and without replacement, from a set of N balls, consisting of R red and $N - R$ white balls, the probability of event A_k, that we draw k red (and $n - k$ white) balls, is given by:

$$\boldsymbol{P(A_k) = \frac{\binom{R}{k} \cdot \binom{N-R}{n-k}}{\binom{N}{n}}}$$



The number of drawn red balls is between 0 and R, so $0 \leq k \leq R$.
Similarly the condition for the white balls is: $0 \leq n - k \leq N - R$.

Using this formula we can compute probabilities of a specific number of k red balls (or, equivalently, of $n - k$ white balls). Such a numerical variable is called a **random variable** $\boldsymbol{X}$, a quantitative variable in a stochastic experiment. In example 2.2.1 X is "the number of red balls in 4 draws from the vase without replacement". The event A_3 = "3 red balls in 4 draws" can be written as "$X = 3$". Or as "$Y = 1$", where Y is the number of white balls in the 4 draws. $X = 3$, $Y = 1$ and A_3 are indeed equivalent events, so

$$P(X = 3) = P(Y = 1) = P(A) = \frac{\binom{10}{3} \cdot \binom{5}{1}}{\binom{15}{4}}.$$

If we determine for X all possible probabilities $P(X = k)$, using the hypergeometric formula (for $k = 0, 1, 2, 3$ and 4), this list of probabilities is called the **hypergeometric distribution of** X. In this case:

$$P(X = k) = \frac{\binom{10}{k} \cdot \binom{5}{4-k}}{\binom{15}{4}}, \qquad \text{for } k = 0, 1, 2, 3, 4$$

$N = 15$, $R = 5$ and $n = 4$ are the **parameters** of the hypergeometric distribution.
In chapter 4 we will give a general definition of random variables and list some families of often used distributions, like the hypergeometric distribution.

2.3 Exercises

1. Some counting problems.
- **a.** Compute the number of all possible orders for the digits 1 to 7.
- **b.** If we use the digits $0, 1, \ldots, 9$ to compose a number with 7 (different) digits, how many of these numbers can we compose?
- **c.** How many combinations of 7 digits can we compose, using the digits $0, 1, \ldots, 9$?
- **d.** A deck of cards consists of 16 face cards (including the aces) and 36 number cards. In the card game Bridge a player receives 13 of the 52 cards at random. What is the probability he has at least 2 face cards?
- **e.** In how many ways can we compose 4 groups of 6, 7, 8 and 9 persons, given a total number of 30 persons?

2. What is the probability that a random number of 3 (possibly identical) digits has a 2 as the lowest digit (see example 2.1.5)?

3. Four balls are drawn at random and without replacement from a box containing 3 red and 7 white balls. Compute the probability that the fourth draw results in a red ball.

4. We draw a card from a deck of 52 cards five times, randomly and without replacement.

- **a.** Determine the probability that all drawn cards are face cards (ace, king, queen or jack: 16 in total).
- **b.** Determine the probability that the first card drawn is an ace and the last card is a king.
- **c.** Determine the probability that of the 5 drawn cards we have exactly one ace and exactly one king.

5. Someone claims to be a connoisseur of wine. He is subjected to a test to prove his claim: he is presented the names of 6 well known (red) wines and 6 glasses of wine, one of each kind. After tasting he has to say which glass contains which wine. He has to name every type once. His expertise is acknowledged if he names at least 4 of the wines correctly. What is the probability that somebody is considered to be a connoisseur, if, in reality, he does not know anything about wines?

6. A manufacturer of rubber rings guarantees that no more than 10% of the rings are bad (substandard). These rings are sold in packets of 100. One of the buyers has the habit of randomly picking 10 rings from the packet and testing those, whether they are of good quality.
When one ring is bad in the test, the buyer refuses the packet from which the ring came.

- **a.** What is the probability that a packet is refused whilst it just satisfies the guarantee bound (10%)?
- **b.** What is the probability that a packet, of which 20% is bad, is nevertheless accepted?

7. A lottery consists of 100 tickets. There are 4 main prizes and 10 consolation prizes. Drawing takes place without replacement. If one buys 5 tickets, what is the probability that he gets:

- **a.** (exactly) one main prize and (exactly) one consolation prize,
- **b.** exactly one prize,
- **c.** no prize,
- **d.** at least one prize?

8. Buying tickets in exercise 7 can be seen as drawing from a box with 3 types of items: 4 main prizes, 10 consolation prizes, and 86 non-prizes.
We can generalize this situations as follows: a box contains N items. Of these items there are N_1 of kind 1, N_2 of kind 2, … , N_k of kind k, such that $N_1 + N_2 + \cdots + N_k = N$.
We draw, at random and without replacement, n items from the box.
What is the probability that we draw n_1 items of kind 1, n_2 items of kind 2, . . . , n_k items of kind k ($n_1 + \cdots + n_k = n$)?

Some extra combinatorial exercises:

9. A deck of 26 cards contains 6 cards of spades. Peter and Paul each get 13 of these cards at random. Determine the probability that:

a. one of both gets exactly 4 spades.
b. both get 3 spades.

10. (The **multinomial coefficient**)

a. In how many ways can we distribute 13 persons over 2 groups of 6 and 7 persons?
b. In how many ways can we distribute 30 persons over 4 groups of 6, 7, 8 and 9 persons?
c. In how many ways can we distribute n persons over k groups of resp. $n_1, n_2, \ldots, n_k$ persons? (This number is called the multinomial coefficient, a generalization of the binomial coefficient.)

11. 5 zeros and 6 ones are randomly placed in a sequence. A maximal uninterrupted subsequence of symbols is called a run. For example: the sequence **0 1 000 11 0 111** counts 6 runs. The length of a run is the number of symbols that the run contains.

a. Determine the probability that a sequence starts with a run of length 3.
b. Determine the probability that a sequence contains 5 runs.

12. Consider a vase with ten balls, numbered 1 to 10. We randomly draw 4 balls, without replacement. What is the probability that the balls have increasing numbers?

Some hints for solving exercises of chapter 2.

1. Check first whether the order, of e.g. the draws, is important (permutations) or not (combinations). Write down the correct formula and use your calculator (button nPr or nCr) to compute the answer. At d.: compare to the solution of example 2.2.1
At e.: use reasoning to find the solution: first compute the number of possible choices of group 1, then compute the number of possible compositions of group 2 using the remaining persons, etc.

3. Do **not** use "conditioning", that is distinction of the results of the first 3 draws (e.g. assuming 1 red in the first 3 draws, computing the probability of a red ball in the 4th draw): this approach is possible but tiresome. Instead, consider the total number of results of 4 balls and count the number with a red ball on the 4th position among them.

4. For each part check whether to use permutations or combinations: **if both methods can be used, choose combinations**, preferably!

5. Imagine the problem vividly, e.g. in a diagram: 6 glasses and 6 nametags: what is the total number of orders in which you can tag those 6 names to the glasses (arbitrarily) and how many of these ordered lists of names lead to the event "6 correct"? 5 correct? 4 correct?

6. How many bad rings should there be in one full packet if the packet "satisfies exactly the guarantee condition". Are we drawing here with or without replacement?

7. In this case we do not have 2 but 3 types of things in a box: with or without replacement?

11. **a**. Write down an example of such a sequence, starting with a run of exactly 3.
How many of these sequences can you "construct" if we have to use 5 zero's and 6 ones?

12. Consider one combination of 4 digits, e.g. 2, 8, 5, 6.
In how many orders can they be positioned? And how many of these orders are increasing?
So, how many orders are there with increasing numbers, considering all combinations of 4 out of 10 digits?

Chapter 3 Conditional probability and independence

3.1 Conditional probability

About the outcome of an experiment often some information is known beforehand.
When an arbitrary person on the street in Amsterdam is asked whether he votes the Dutch liberal party VVD, then the probability that he votes VVD will be different from the answer of a random person in Wassenaar (the "millionaires village" of Holland).
Consider the experiment of randomly choosing a Dutchman, provided that he is an inhabitant of Amsterdam. Under this condition we can ask for the probability that he will vote VVD. We cannot determine this "probability" with the tools we applied so far. We need a new definition. In the example we ask for the **conditional probability** that a random Dutchman votes VVD, **given** that he lives in Amsterdam.
Intuitively it seems obvious to equate this probability with the probability that a random **person, living in Amsterdam,** votes VVD.
With the event A we denote the Dutch people who vote VVD and with B we denote people living in Amsterdam. Then this conditional probability is equal to $\frac{N(AB)}{N(B)}$, where $N(B)$ denotes the number of people living in Amsterdam and $N(AB)$ the number of these people in Amsterdam who vote VVD. We denote this conditional probability as $P(A|B)$.
If $N(S)$ is the number of all Dutchmen, we get for this symmetric sample space:

$$P(A|B) = \frac{N(AB)}{N(B)} = \frac{N(AB)/N(S)}{N(B)/N(S)} = \frac{P(AB)}{P(B)}$$

One should see the difference between the probabilities $P(A), P(A|B)$ and $P(AB)$ clearly.
In this example the probabilities can be interpreted as follows:

- $P(A)$: the proportion of VVD voters among all Dutch voters.
- $P(A|B)$: the proportion of VVD voters among voters in Amsterdam.
- $P(AB)$: the proportion of VVD voters in Amsterdam among all Dutch voters.

The equation above shows that we can determine the conditional probability $P(A|B)$ with the unconditional probabilities $P(AB)$ and $P(B)$. From now on this intuitive result for the symmetric probability space of all Dutch voters is used as definition for conditional probability in any probability space.

Definition 3.1.1 When A and B are events and $P(B) > 0$, then we define

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

as **the (conditional) probability of *A* under condition of *B*** **(**or: **the (conditional) probability of *A* given *B*)**.

Moreover, from the definition it follows that **for fixed B with $P(B) > 0$** the conditional probability $P(\cdot|B)$ is a probability as well, i.e. it fulfills Kolmogorov's axioms.
The requirement $P(B) > 0$ is not a heavy restriction, as an event with probability 0 cannot have occurred! Kolmogorov`s axiom (2), e.g., is also true for conditional probabilities, since:

$$P(S|B) = \frac{P(SB)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

So $\big(S, P(\cdot|B)\big)$ is a probability space as well. As a result this conditional probability also has the properties following from the axiom. It follows that: $P\big(\overline{A}\big|B\big) = 1 - P(A|B)$, etc.

Example 3.1.2 A company owns two factories which fabricate the same products.
During a certain period factory 1 produces 1000 products of which 100 are defective and factory 2 makes 4000 products of which 200 are defective.
From the total production one product is chosen randomly and it appears to be defective.
What is the probability that the product is manufactured in factory 1?
If A_1 is the event "produced by factory 1", A_2 "produced by factory 2" and D the subset of defective products, then $P(A_1|D)$ is the requested probability. The total production is 5000 products, the number of defective products by factory 1 is 100 and the total number of defective products is 300, so intuitively it is clear that one of three defective products are produced in factory 1. Applying the definition we find:

$$P(A_1D) = \frac{100}{5000}, P(D) = \frac{300}{5000} \text{ and } P(A_1|D) = \frac{P(A_1D)}{P(D)} = \frac{1}{3}$$ ■

Example 3.1.3 A box contains two coins, one of which is fair, i.e., Heads and Tails are tossed with equal probability, whilst the other coin is not fair: it has Tails on both sides.
We randomly choose one of the coins and toss it. The result is "Tails": then, if we turn around the chosen coin, what is the probability that the other side will show Tails as well?
A is the event that the coin lands Tails up and B is the event that the coin has Tails on both sides, then $P(B|A)$ is the requested conditional probability. We have $P(A) = \frac{3}{4}$, since three of the four sides of the two coins are Tails and all 4 sides land up with equal probabilities $\left(\frac{1}{4}\right)$. Moreover, we have $P(BA) = P(B) = \frac{1}{2}$, since $B \subset A$ (if both sides are Tails, A occurs) and both coins have equal probability to be chosen. So $(B|A) = \frac{P(BA)}{P(A)} = \frac{1/2}{3/4} = \frac{2}{3}$. ■

From the definition of conditional probability we immediately get:

$\boldsymbol{P(AB) = P(A) \cdot P(B|A)}$, which is known as the (general) **product rule**.

Note that this is the product rule for events which is different from the product rule for counting (property 2.1.2). If we substitute $A = A_1A_2$ and $B = A_3$ in the product rule above, we find for the intersection of 3 events:

$$\boldsymbol{P(A_1A_2A_3) = P(A_1A_2) \cdot P(A_3|A_1A_2) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1A_2)}$$

This property can be generalized to n events and is used intuitively in many practical

situations, as the following example illustrates.

Example 3.1.4 We draw three marbles, randomly and without replacement, from a vase with 10 marbles numbered 1 to 10. We are interested, e.g., in the probability that we draw the marbles with numbers 1, 2 and 3 in this order.
The probability space is symmetric and $S = \{(i,j,k) | i,j,k = 1,2,\ldots,10 \text{ and } i,j,k \text{ different}\}$ and consists of $N(S) = \frac{10!}{(10-3)!}$ permutations. So $P\big((1,2,3)\big) = \frac{1}{10 \cdot 9 \cdot 8}$.
This answer can also be derived (intuitively) by reasoning that the probability of drawing marble 1 the first time is $\frac{1}{10}$, the probability that marble 2 is drawn from the remaining 9 marbles then is $\frac{1}{9}$ and finally the probability that marble 3 is drawn is $\frac{1}{8}$.
That's why many people will directly state: the probability of drawing (1, 2, 3) is $\frac{1}{10} \cdot \frac{1}{9} \cdot \frac{1}{8}$.
Why is this multiplication correct?
We define A_i as the event that the i-th draw gives marble i with $i = 1, 2, 3$. Then we have:

$$P(A_1) = \frac{1}{10} \;\; P(A_2|A_1) = \frac{1}{9} \;\; \text{en } P(A_3|A_1A_2) = \frac{1}{8}$$

According to the rule above we find:

$$P(A_1A_2A_3) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1A_2) = \frac{1}{10} \cdot \frac{1}{9} \cdot \frac{1}{8}$$

We note that $A_2 = \{(i,2,k) | i,k = 1,3,4,\ldots,10 \text{ en } i \neq k\}$.
So $P(A_2) = \frac{N(A_2)}{N(S)} = \frac{9 \cdot 8}{10 \cdot 9 \cdot 8} = \frac{1}{10} = P(A_1)$. Clearly $P(A_2|A_1) \neq P(A_2)$
Similarly, $P(A_3) = \frac{1}{10}$, and we see that $P(A_1A_2A_3) \neq P(A_1) \cdot P(A_2) \cdot P(A_3)$. ■

Generalizing the previous, we find:

Property 3.1.5 (general product rule)

For n events $A_1, A_2, \ldots, A_n$ with $n \geq 2$ and $P(A_1A_2 \ldots A_{n-1}) > 0$ we have:

$$\boldsymbol{P(A_1A_2 \ldots A_n) = P(A_1) \cdot P(A_2|A_1) \cdot \ldots \cdot P(A_n|A_1A_2 \ldots A_{n-1})}.$$

The proof can be given using the definition of conditional probability fort $n = 2$ events and induction on n, as shown in the extension to $n = 3$ events.

3.2 Law of total probability and Bayes` rule

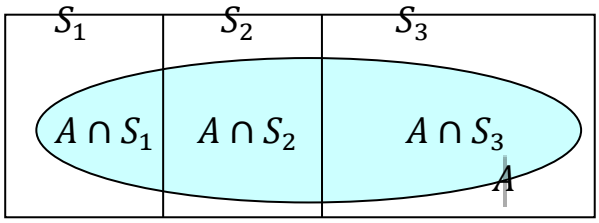
Example 3.2.1 A company has three factories (1, 2 and 3) which all produce the same smartphones. These factories contribute resp. 15%, 35% and 50% to the total production. The probability that a smartphone, produced by these factories, is defective is 0.01, 0.05 and 0.02, resp. If we see buying a smartphone as a random draw of a smartphone from the total production, we wonder how we can find the answer to the following questions:

a. What is the probability that the smartphone is defective?
b. If the smartphone is defective, what is the probability that it was produced in factory 1?

Question a. is easily answered intuitively: the probability of getting a defective smartphone is the average probability of a defective smartphone, using as weighing factors the given proportions of the total production:

$$0.01 \cdot 0.15 + 0.05 \cdot 0.35 + 0.02 \cdot 0.50 = 2.9\%.$$

Question b. is a lot harder to answer intuitively, so let us describe the situation in a probability model. Another goal of this course is to "prove properly", what seems to be intuitively correct, or at least make it convincingly likely. Therefore we define S as the set of all produced smartphones, S_i is the event that the smartphone is produced in factory i ($i = 1, 2, 3$) and A the event that it is defective (see the Venn diagram below).



The proportions of the production of the three factories are the given probabilities:

$$P(S_1) = 0.15, P(S_2) = 0.35 \text{ en } P(S_3) = 0.50.$$

The given probabilities of a defective smartphone are conditional:

$$P(A|S_1) = 0.01, P(A|S_2) = 0.05 \text{ en } P(A|S_3) = 0.02.$$

According to the product rule is $P(AS_i) = P(A|S_i) \cdot P(S_i)$, so answering question a.:

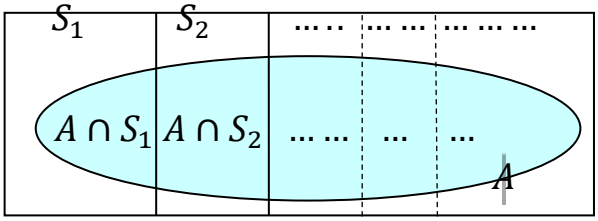
$$\begin{aligned} P(A) &= P(AS_1) + P(AS_2) + P(AS_3) \\ &= P(A|S_1) \cdot P(S_1) + P(A|S_2) \cdot P(S_2) + P(A|S_3) \cdot P(S_3) \\ &= 0.01 \cdot 0.15 + 0.05 \cdot 0.35 + 0.02 \cdot 0.5 = 2.9\%. \end{aligned}$$

Now we can see that question b. refers to a conditional probability: a smartphone being made in factory 1, given the fact that it is defective: the probability $P(S_1|A)$ can be found by simply applying the definition of conditional probability and using what we already know:

$$P(S_1|A) = \frac{P(S_1 A)}{P(A)} = \frac{P(A|S_1) \cdot P(S_1)}{P(A)} = \frac{0.01 \cdot 0.15}{0.029} \approx 5.2\%$$ ■

In this example S_1, S_2 and S_3 are a partition of S. The probabilities of the parts S_i and the conditional probabilities of defective smartphone for each factory are known. In a. we computed the "overall" (total) probability of defective smartphones. We will now generalize this property, which can be applied for any partition $\{S_i\}$ of S: $\{S_i\}$ is a finite sequence of

(mutually exclusive) parts S_1 to S_n, like in example 3.2.1, or a numerable infinite sequence of parts $S_1, S_2, \ldots$.



Property 3.2.2 (The law of total probability)

If $\{S_i\}$ is a partition of S such that $P(S_i) > 0$ for all i, then for each event A we have:

$$\boldsymbol{P(A) = P(A|S_1) \cdot P(S_1) + P(A|S_2) \cdot P(S_2) + \cdots = \sum_i P(A|S_i) \cdot P(S_i)}$$

The computation of the b-part of example 3.2.1 can be generalized as well:

Property 3.2.3 (Bayes` rule)

If $\{S_i\}$ is a partition of S with $P(S_i) > 0$ for each i, then for each event A with $P(A) > 0$ we have:

$$\boldsymbol{P(S_k|A) = \frac{P(AS_k)}{P(A)} = \frac{P(A|S_k)P(S_k)}{\sum_i P(A|S_i) \cdot P(S_i)}}$$

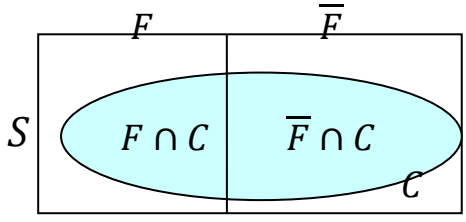
The proper application of Bayes` rule (and the law of total probability) is illustrated in the following (former) exam exercise:

Example 3.2.4 According to a cyclist 10% of all professional cyclists use forbidden stimulants. The use is checked with a test and if a cyclist is caught (positive), he will be suspended. From tests we know that cyclists who use these stimulants are caught in 85% of the cases. However, the test is also positive for 5% of the non-users.
Determine on these grounds the probability that a cyclist, who is tested positive, is falsely suspended.

Solution:

The probability model: we define the sample space S as all professional cyclists, $\boldsymbol{F}$ as the event that a professional uses **f**orbidden stimulants and $\boldsymbol{C}$ the event that a professional is **c**aught (so tested positive).

Given is thus: $P(F) = 0.1$, $P(C|F) = 0.85$ and $P(C|\overline{F}) = 0.05$.



Note that F and $\overline{F}$ are a partition of S, as the Venn diagram illustrates

Then the requested conditional probability of a non-user, given a positive test, is according to Bayes' formula:

$$P(\overline{F}|C) = \frac{P(\overline{F} \cap C)}{P(C)} = \frac{P(C|\overline{F}) \cdot P(\overline{F})}{P(C|F) \cdot P(F) + P(C|\overline{F}) \cdot P(\overline{F})} = \frac{0.05 \cdot (1 - 0.1)}{0.85 \cdot 0.1 + 0.05 \cdot 0.9} \approx 34.6\% \quad \blacksquare$$

3.3 Independence of events and random variables

When pointing at a random European, then knowing whether a woman or a man is chosen, does not influence the probability that the person comes from Sweden. It would be different if we know that the chosen person is blond, since there are relatively more blond people in Sweden than there are blond Europeans. If the occurrence of *B* does not affect the probability on the occurrence of another event *A*, then we say *A* is independent of *B*. Then we should have:

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

Is it also true that the occurrence of event A does not influence the probability of the occurrence of B?
Yes, since from $P(A|B) = P(A)$ we get:

$$\frac{P(AB)}{P(B)} = P(A), \quad \text{provided that } P(B) > 0$$

Or:

$$P(AB) = P(A) \cdot P(B)$$

If $P(A) > 0$ is true as well, we conclude from this equality (divide by $P(A)$) that

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

$P(A|B)$ or $P(B|A)$ are not defined if $P(A) = 0$ or $P(B) = 0$, but the equality $P(AB) = P(A) \cdot P(B)$ is always defined. That is why we use the following definition:

Definition 3.3.1 The events *A* and *B* are **independent** when $\boldsymbol{P(AB) = P(A) \cdot P(B)}$

When two events *A* and *B* are mutually exclusive, so if $P(AB) = 0$, then *A* and *B* can only be independent if $P(A) = 0$ or $P(B) = 0$.

Example 3.3.2 From a deck of 52 cards we randomly draw one card. H is the event that the drawn card is of hearts, and J is the event that the drawn card is a Jack. Then we get:

$$P(H) = \frac{13}{52} = \frac{1}{4} \quad \text{and } P(J) = \frac{4}{52} = \frac{1}{13}$$

Since HJ is the event that we draw a Jack of hearts we get:

$$P(HJ) = \frac{1}{52} \quad \text{and} \quad P(H)P(J) = \frac{1}{4} \cdot \frac{1}{13} = \frac{1}{52}.$$

The events H and J are thus independent. ■

In the previous example we have proven the independence of the events H and J. Often we do not know the probabilities and decide upon independence of two events in a different way.

Example 3.3.3 We roll a dice twice. A is the event that we roll 5 the first time and B the event that we roll a 3 or higher the second time. Assuming an unbiased dice we have $P(A) = \frac{1}{6}$ and $P(B) = \frac{2}{3}$.
In general: $P(AB) = P(A) \cdot P(B|A)$. However, the conditional probability $P(B|A)$ can be determined by assuming that the result of the first roll does not influence the result of the second roll, i.e., we assume A and B to be independent. Then we have $P(AB) = P(A) \cdot P(B)$.
The probability of rolling a 5 the first time and a 3 or higher the second time, is thus:

$$P(AB) \overset{\text{ind.}}{=} P(A)P(B) = \frac{1}{6} \cdot \frac{2}{3} = \frac{1}{9}$$ ■

Note that in the example above two assumptions have been made considering the probability model: **the dice is unbiased and the two rolls are independent**. These assumptions are equivalent to the assumption that we have a symmetric probability space.
Since A_i is the event that we roll i face up the first time and B_j the event that we roll j face up $(i, j = 1, 2, \dots, 6)$, then based on the unbiasedness of the dice we have

$$P(A_i) = P(B_j) = \frac{1}{6}$$

And, due to independence, we have: $P(A_i B_j) \overset{\text{ind.}}{=} P(A_i)P(B_j) = \frac{1}{36}$
We could have determined the probability of the event in example 3.3.3 by using the probability definition of Laplace. However, the described method of calculation, directly using of independence, is intuitive and easier.
From now on we will assume independence of such experiments (repeatedly tossing a coin, repeatedly drawing a marble from a vase **with** replacement, etc.) without explicitly stating it.
When we say that two **experiments** are **independent**, we mean that every pair of events A and B, where A only relates to the first experiment and B only to the second, can be assumed independent.

Example 3.3.4 A device consists of two components. A_1 is the event that the first component works and A_2 the event that the second component works. The device only works if both components work and we have good reasons to assume that the working of one component does not influence the working of the other component, i.e., A_1 and A_2 are independent. Under this assumption, the probability that a device works is:

$$P(A_1A_2) \overset{\text{ind.}}{=} P(A_1)P(A_2).$$ ■

We want to extend the definition of independence to the case where we have more than two events. It seems logical that, if we have 3 independent experiments and for each experiment a an event A_i $(i = 1,2,3)$ is defined, then $P(A_1A_2A_3) = P(A_1)P(A_2)P(A_3)$ should be true. But for any pair of experiments we have independence as well, so $P(A_1A_3) = P(A_1)P(A_3)$.
Is any threesome of events A_1, A_2 and A_3 independent if $P(A_1A_2A_3) = P(A_1)P(A_2)P(A_3)$?
Or, if any pair of A_1, A_2 and A_3 is independent, then $P(A_1A_2A_3) = P(A_1)P(A_2)P(A_3)$?
The answer to these intriguing questions is negative!
We will call the events $\{A_i\}$ **pairwise independent** when each pair in this sequence of events is independent. Pairwise independence does not rule out that there is a certain dependence between events, as the following example illustrates.

Example 3.3.5 We toss a fair coin twice. The sample space is $S = \{HH, HT, TH, TT\}$ and each of these outcomes occur with equal probability (due to the assumptions of

independence).
Say A is "first toss is heads", B is "second toss is heads" and C is "both tosses give the same result", so: $A = \{HH, HT\}$, $B = \{HH, TH\}$ and $C = \{HH, TT\}$
A, B and C are pairwise independent, since

$$P(AB) = \frac{1}{4} = P(A)P(B)$$
$$P(AC) = \frac{1}{4} = P(A)P(C)$$
$$P(BC) = \frac{1}{4} = P(B)P(C).$$

But $P(C|AB) = \frac{P(ABC)}{P(AB)} = 1 \neq P(C)$,
so AB provides "information" about (the occurrence of) C.
To exclude every form of dependence between events A, B and C, more requirements are needed for pairwise independence. It should for example also be true that:

$$P(A|BC) = P(A) \quad \text{or} \quad P(ABC) = P(A) \cdot P(B) \cdot P(C).$$ ■

This last requirement on itself is not sufficient to guarantee pairwise independence, as the following example shows:

Example 3.3.6 We roll a fair dice twice. Say, A is the event that the first roll results in 1, 2 or 3 and B the event that the first roll results in 3, 4 or 5.
C is the event that the sum of the two throws is 9, so C consists of four outcomes: $(3,6), (4,5), (5,4)$ and $(6,3)$.
$P(A) = P(B) = \frac{1}{2}$ and $P(C) = \frac{4}{36}$
The intersection of A, B and C consists of one outcome: $(3,6)$.
So

$$P(ABC) = \frac{1}{36} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{9} = P(A) \cdot P(B) \cdot P(C)$$

But:

$$P(AB) = \frac{1}{6} \neq \frac{1}{2} \cdot \frac{1}{2} = P(A) \cdot P(B)$$
$$P(BC) = \frac{3}{36} \neq \frac{1}{2} \cdot \frac{1}{9} = P(B) \cdot P(C)$$
$$P(AC) = \frac{1}{36} \neq \frac{1}{2} \cdot \frac{1}{9} = P(A) \cdot P(C)$$

A, B and C are not pairwise independent. ■

A, B and C can only be called independent (or **mutually independent**) when for each pair and for the threesome the probability of the intersection can be written as the product of the separate probabilities. (Mutual) independence implies pairwise independence, but not reverse. For a finite or countable infinite number of events A_1, A_2, A_3, ... we should thus give this requirement for every couple, threesome, quadruplet, etc.

Definition 3.3.7 The events $A_1, A_2, A_3, \ldots$ are independent if for each subsequence $A_{i1}, A_{i2}, \ldots, A_{ik}$ with $k \geq 2$, it is true that

$$\boldsymbol{P(A_{i1}A_{i2} \ldots A_{ik}) \overset{\text{ind.}}{=} P(A_{i1}) \cdot P(A_{i2}) \cdot \ldots \cdot P(A_{ik})}$$

If two events A and B are independent, then so are A and $\overline{B}$, and $\overline{A}$ and $\overline{B}$. (see exercise 8).

Similar properties one can also give for more than two events, e.g.:
if A, B, C and D are independent, then AB and $C \cup D$ are independent as well, since

$$\begin{aligned} P(AB(C \cup D)) &= P(ABC \cup ABD) \\ &= P(ABC) + P(ABD) - P(ABCD) \\ &\overset{\text{ind.}}{=} P(A)P(B)P(C) + P(A)P(B)P(D) - P(A)P(B)P(C)P(D) \\ &= P(A)P(B)[P(C) + P(D) - P(CD)] \\ &\overset{\text{ind.}}{=} P(AB)P(C \cup D). \end{aligned}$$

Similarly, one can prove that, e.g., $\overline{A}$ and BCD are independent, etc.

Similar to the independence of two experiments, if we can reasonably assume independence of n experiments, the each sequence of n corresponding events are independent.

Example 3.3.8 We roll a fair dice ten times and record the number of times we roll 6.
We want to determine the probability of the event B_k, that out of 10 rolls k result in a 6 head up ($k = 0, 1, \ldots, 10$). This experiment consists of 10 sub-experiments, each with two possible outcomes: 6 and not-6, denoted by A and $\overline{A}$. As the result of one roll does not influence the result of other throws, these sub-experiments may be assumed independent.
Defining A_i as the event that the i^{th} roll results in a 6 with $1 \leq i \leq 10$, then, due to independence, we have for example:

$$P(A_1 A_2 \overline{A}_3) \overset{\text{ind.}}{=} P(A_1) \cdot P(A_2) \cdot P(\overline{A}_3) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{5}{6}$$

B_3 occurs when for example the first three rolls result in a 6 and the other 7 rolls result in a not-6. Then we have:

$$P(A_1 A_2 A_3 \overline{A}_4 \ldots \overline{A}_{10}) \overset{\text{ind.}}{=} P(A_1) \cdot P(A_2) \cdot P(A_3) \cdot P(\overline{A}_4) \cdot \ldots \cdot P(\overline{A}_{10}) = \left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^7$$

Each order with three sixes and seven not-sixes occurs with this probability and there are $\binom{10}{3}$ of those orders, so we get:

$$P(B_3) = \binom{10}{3}\left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^7 \approx 15.5\%.$$

And: $$P(B_k) = \binom{10}{k}\left(\frac{1}{6}\right)^k \cdot \left(\frac{5}{6}\right)^{10-k}$$

As before in the combinatorial probability chapter 2, we can define random variables to describe the numerical variable: Define $X =$ "The number of six's in 10 rolls"
Since $B_k = \{X = k\}$ we can give the **probability distribution** of X as follows:

$$P(X = k) = \binom{10}{k}\left(\frac{1}{6}\right)^k \cdot \left(\frac{5}{6}\right)^{10-k}, \text{where } k = 0, 1, \ldots, 10$$

X is said to be **binomially distributed with parameters $n = 10$** (the number of trials) **and $p = \frac{1}{6}$**, the success probability. ■

The last probability formula is an example of the binomial formula, which can be applied in similar situations to compute probabilities. Other examples are: answering a multiple choice test randomly; determining the number of drawn red balls, when drawing balls randomly and with replacement from a box with 7 red and 13 white balls; the number of successful shots on a basket; the number of defective products in a quality control sample.
In all these examples the repeated trials have basically two outcomes, success and failure, and the outcome of one of the trials does not influence the outcome of other trials: the probability of an outcome is the same for each experiment and is not influenced by the result of other experiments. These type of experiments are called Bernoulli-experiments or Bernoulli trials.

Definition 3.3.9 A series of experiments is called **Bernoulli experiments** or **trials** if
1) each experiment has two possible outcomes, often denoted with 'Success' and 'Failure',
2) the experiments are independent and
3) the probability of success is the same for each experiment.

The success probability is usually denoted by $\boldsymbol{p}$ and the probability of failure with $\boldsymbol{1-p}$.
We can now generalize example 3.3.8 as follows (the proof is similar to the derivation in example 3.3.8).

Property 3.3.10 (The **binomial formula**)

If X is the number of successes in n Bernoulli experiments with success probability p, then:

$$\boldsymbol{P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}},\ \text{where } k=0,1,2,\dots,n$$

Remember:
"$p^k(1-p)^{n-k}$ is the probability that the first k trials are successful and the last $n-k$ are failures and $\binom{n}{k}$ the number of possible orders of k successes and $n-k$ failures."

Another question we can pose whilst rolling a dice is: what is the probability that we succeed in rolling the 6 in the tenth roll?
Similar questions can arise in the same kind of situations:

- If we observe passing cars, what is probability that we have to wait until the tenth car to observe the first Audi?
- while conducting quality control on products, what is the probability that the 16^{th} product is the first to be substandard?

More generally: what is the probability of the event $\{Y=k\}$, that is the probability that the first success occurs in the k^{th} experiment, if we execute a series of Bernoulli experiments?

The random variable Y can be defined as the **number of trials** until we achieve our first success. Furthermore we define A_i as the occurrence of success in the i^{th} experiment, then

$$\{Y=4\}=\overline{A}_1\cap\overline{A}_2\cap\overline{A}_3\cap A_4 \quad = \quad \overline{A}_1\overline{A}_2\overline{A}_3A_4$$

And in general: $\{Y=k\}=\overline{A}_1\overline{A}_2\dots\overline{A}_{k-1}A_k$
and due to the independence of the experiments we have (for $k=1,2,3,\dots$):

$$P(Y=k) = P\left(\overline{A}_1\overline{A}_2 \dots \overline{A}_{k-1}A_k\right) \overset{\text{ind.}}{=} P\left(\overline{A}_1\right)P\left(\overline{A}_2\right) \dots P\left(\overline{A}_{k-1}\right)P(A_k)$$
$$= (1-p)(1-p) \dots (1-p)p = (1-p)^{k-1}p$$

In this way we have derived:

Property 3.3.11 (**the geometric formula**)

If we conduct Bernoulli trials with success probability p until a success occurs and X is the number of required trials, then

$$\boldsymbol{P(X=k) = (1-p)^{k-1}p}, \text{ where } k = 1, 2, 3, \dots$$

X is said to have a **geometric distribution with parameter $\boldsymbol{p}$** (the success probability for each trial).

3.4 Exercises

1. **a.** Compute $P(A|B)$ if we know that $P(A \cup B) = 0.8$, $P(A \cap B) = 0.3$ and $P(A) = 0.35$.
b. Compute $P(ABC)$ if we know that $P(B|AC) = \frac{1}{2}$, $P(C) = \frac{4}{5}$ and $P(A|C) = \frac{3}{4}$.

2. At the end of a production line a final test is conducted on the products. We know that 98% of the total production is approved. Whether a product really meets all requirements is something we will experience later. From statistics we know that 97% of the approved products really meet the requirements. For 5% of the disapproved products this is the case as well. Define for a randomly chosen product the events A and B as follows.
$A =$ "the product is approved" and $B =$ "the product meets all requirements".

a. Express the given probabilities (0.98, 0.97, 0.05) in A and B and compute $P(B)$.
b. Compute the probability that a product is disapproved if it does not meet the requirements.
c. Are the events A and B independent? Motivate your answer.

3. A car driver who causes an accident has to undergo a blood test. From past experience we know that, when someone is "under influence" (of alcohol or drugs) the probability of a positive blood test is 75%. When the driver is not under influence the probability of a positive test is 2%. Let us assume that 5% of the drivers who cause an accident is under influence.
What is the probability that someone who causes an accident is under influence when the blood test is positive? (Answer the question by 1. defining proper events, 2. expressing the given probabilities in these events and 3. using rules of probability to compute the requested probability).

4. In the West-African country Gambia there are 3 mobile networks available: *Africell*, *Gamcel* and *Comium*. The market share of *Africell* is twice the market share of *Gamcel*. The proportions of (non-prepaid) subscriptions are for *Africell* 10%, for *Gamcel* 20% and for *Comium* 30%. According to the government 15% of all mobile phone users in Gambia has a non-prepaid subscription.
Compute the market share of *Comium* (use the same approach as in exercise 3)

5. A cupboard has three drawers. The first drawer contains two golden coins, the second drawer contains two silver coins and the third drawer contains one silver and one golden coin. A drawer is chosen at random and one coin is drawn randomly from that drawer: this coin turns out to be made of gold.
What is the probability that the other coin in the drawer is also made of gold? (Again: define events etc. to prove what you might give as an intuitively correct computation)

6. A vase contains 5 red and 7 white marbles. We roll a fair dice and draw randomly without replacement as many marbles as the face up number of the dice (the number of draws depends on the dice roll).

a. Determine the conditional probability of 3 red marbles when we have rolled 5 with the dice.
b. Determine the probability of (exactly) 3 red marbles.

7. A student participates in a multiple choice test with two possible answers per question. When he does not know the correct answer, he guesses by tossing a fair coin.
To 60% of the questions he knows the answer (assume that this answer is really correct in that case).
What is the probability that he knew the answer to a question that he answered correctly?

8. Prove, using the definition:
a. If A and B are independent, then A and $\overline{B}$ are independent, and $\overline{A}$ and $\overline{B}$ as well.
b. If A, B and C are independent, then so are A and BC.

9. "Mutually exclusive events are independent" Is this statement correct? Motivate your answer.

10. Compute the probability that we need more than 6 rolls of a dice to obtain the first 6 as a result of the roll.

11. For a game of chance one has to predict the result of 12 football matches (1, 2 or 3 for victory, loss and draw for the home team, respectively). If somebody would give a completely random prediction of all twelve matches, what is the probability that he will have at least 10 correct predictions?

Extra exercise, illustrating the use of conditional probability in case of detecting rare illnesses:

12. The ELISA-test was introduced in the eighties to check whether blood donors are HIV-infected (AIDS): the test detects antibodies if they are present in the blood.
Research showed that if antibodies are present in the blood ELISA is positive at a rate of 0.997 and negative at rate 0.003. If the person is not infected, ELISA is negative with probability 0.985 and positive with 0.015 ("false positives")
(since ELISA is designed to avoid contaminated blood entering the blood banks, the relatively large probability of a false positive (1.5%) is accepted against the small probability of not discovering antibodies (0.3%)).
Assume we have a population where 1% is HIV-infected.

a. Compute for a randomly chosen person in the population the probability of a positive ELISA test result.
b. Compute the probability that a person is really HIV-infected if he receives the message that his test result is positive.
(*This exercise illustrates that when a population on diseases is under consideration (such as AIDS, a type of cancer or illegal drugs) one should oversee the consequences in advance: if the phenomenon occurs at a (very) low rate, the probability of false positives could be unacceptably high, even if the percentages of correct test results for all groups is high*).

Some hints for solving exercises of chapter 3.

1. **a.** Use a Venn diagram to find quickly what rules of probability you could apply.
   b. For conditional probabilities a Venn diagram is not very helpful; use the definition of conditional probability and the product rule instead.
2. Sketch the Venn diagram such that the partition you use consist of parts of which you know the (unconditional) probability. The law of total probability can be derived from the diagram by computing the intersections using the product rule. Bayes' rule follows directly from the definition of conditional probability.
3. Use recognizable names for the events, e.g. $U =$ "under influence" and express first the given and requested probabilities in these events
4. Consider this exercise to be a puzzle: what is the relation between the 3 market shares? And could you apply the law of total probability on the event S = "Mobile phone subscription" to derive the market shares.
5. Distinguish the drawers and the material of the first chosen coin (the second is gold as well means….).
6. **a.** Use the hypergeometric formula (drawing without replacement)
   b. Which case can you distinguish?
7. Use the same approach as in exercises 2 and 3.
8. First give the definition and what you want to prove: establish a relation between the two by considering the Venn diagram.
9. First give the definition of both concepts!
10. Reason how you can give the answer in one simple formula. (using an addition is possible too but much more work).
11. Can you assume independence of the trials here? And if so, should you use the geometrical or the binomial formula?

Chapter 4 Discrete random variables

4.1 Random variable

In previous chapters we discussed experiments with corresponding probability spaces. The outcome of an experiment is sometimes a real number, like when rolling a dice: $S = \{1, 2, 3, 4, 5, 6\}$. Or observing the life time of a light bulb: $S = [0, \infty)$.
Whether or not you win the lottery, however, is an experiment with outcome either "success or "failure": $S = \{s, f\}$ consists of non-numerical outcomes.
It is also possible that outcomes are composed of multiple numbers, like a communication channel which sends code words consisting of 5 zeros or ones.
Even with these types of experiments we want to assign a number to each possible outcome. After a lottery draw for example the amount of money that has to be paid: 1 million for success and 0 for failure. When sending the code word, for example, the number of ones in the corresponding code word. Then we will assign a real number to each code word.
The function X which assigns a number (= number of ones) to each code word is called a **random variable** (or stochastic variable).
When the experiment is executed, we get an outcome through a probability mechanism, e.g. 01101, to which the random variable X assigns a function-value: for 01101 we find the value 3 and we denote $X(01101) = 3$ or simply $X = 3$.
The number 3 is called the **realization** of X.

> **Definition 4.1.1** If S is the sample space of an experiment, then a real function $X: S \to \mathbb{R}$, which assigns a real number $X(s)$ to each outcome $s \in S$, is a **random variable.**

(In definition 4.1.1 $\mathbb{R}$ is the set of all real numbers).
We use capitals for random variables: X_1, X_2, X_3, Y, Z, ... or (in case of integer numbers) N.
A realization is an observed value: $X = x$ means that the variable X attains the real value x.

Example 4.1.2 For demographic research a random Dutchman is chosen and asked for his age. For this experiment the Dutch are a symmetric probability space with S = {all Dutch people}. We now define X as "the age of the chosen person". The random variable X gets the age $X(s)$ for person s in the population S. The ages vary from 0 to 120 years, so the set of all realizations $X(s)$ is $\{0, 1, 2, \ldots, 120\}$: this is the **range** S_X of the variable X.
To the same, arbitrarily chosen person we can affix his "weight in kg" or his "length in cm", thereby defining additional random variables Y and Z, with realizations, e.g. $Y(s) = 80$ kg and $Z(s) = 185$ cm for person s.
For one experiment we can introduce many random variables, one for each desirable **quantitative aspect** of this experiment. ■

Definition 4.1.3 The **range** S_X of a random variable X, defined on a sample space S is the set of all possible realizations $X(s)$.

So $S_X = \{X(s) | s \in S\}$.
The range of a variable can be

- **finite,** e.g. if X = "the face up number of a rolled dice": $S_X = \{1, 2, 3, 4, 5, 6\}$,
- **countably infinite** ("countable" using the natural numbers 1, 2, 3, …), e.g. if Y = "the number of required of rolls of a dice to achieve a 6": $S_Y = \{1, 2, 3, 4, \dots\}$ or
- **not countably infinite**, e.g. if Z = "the life time (in hours) of an arbitrary processor": $S_Z = [0, \infty)$.

X and Y are examples of discrete random variables, which will be discussed in this chapter.
Z is an example of a continuous (interval) variable, to be discussed in chapter 6.

Definition 4.1.4 A **random variable** X is **discrete** if the range S_X is denumerable.

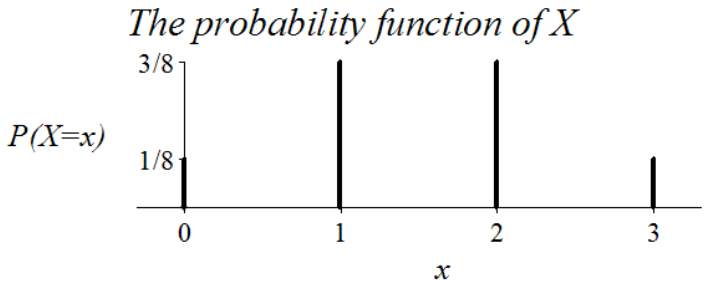
If X is discrete, S_X has the shape $\{x_1, x_2, \dots, x_n\}$ or $\{x_1, x_2, x_3, \dots\}$.

4.2 The probability function of a discrete random variable

Example 4.2.1 We flip a fair coin three times and define X as the number of tails.
For this experiment the $2^3 = 8$ outcomes (each flip either Head or Tail) in
$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$ form a symmetric probability space, so that $P(A) = \frac{N(A)}{8}$ for each event A.
$X =$ "the number of tails in three rolls" can be 0, 1, 2 or 3, so $S_X = \{0, 1, 2, 3\}$.
When we want to determine the probability that we obtain tails once in three rolls, we ask for the probability of the event $\{X = 1\} = \{s \in S | X(s) = 1\}$.
This event occurs in three of the outcomes: $\{X = 1\} = \{HHT, HTH, THH\}$,
so $P(\{X = 1\}) = \frac{3}{8}$.
For a more compact notation we omit the braces: $P(X = 1) = \frac{3}{8}$
Similarly, we can determine the probabilities for $X = 0$, 2 of 3:

$$P(X = 0) = \frac{1}{8}, \quad P(X = 1) = \frac{3}{8}, P(X = 2) = \frac{3}{8} \quad \text{and} \quad P(X = 3) = \frac{1}{8}$$

This is the **probability function** $P(X = x)$ of X, which is usually graphed using a so called bar graph of probabilities, with x in S_X on the X-axis and the probabilities $P(X = x)$ on the Y-axis.



Note that the total length of the "bars" is the total probability 1. This is not an amazing observation: $\{X = 0\}, \{X = 1\}, \{X = 2\}$ and $\{X = 3\}$ is a partition of S, so that:

$$\sum_{x=0}^{3} P(X = x) = P(X \in \{0, 1, 2, 3\}) = P(S) = 1. \qquad \blacksquare$$

Definition 4.2.2 If X is a discrete random variable, then we will call the function that assigns a probability $P(X = x)$ to each $x \in S_X$ the **probability function** of de X.

In example 4.2.1 we noticed that the sum of all probabilities $P(X = x)$ equals 1. In general:

Property 4.2.3 For the probability function of a discrete random variable X we have:

1) $\boldsymbol{P(X = x) \geq 0}$ for $x \in S_X$ and

2) $\boldsymbol{\sum_{x \in S_X} P(X = x) = 1}$

Conversely, any function which satisfies conditions 1) and 2) is a probability function. Probability statements regarding discrete random variables can now be expressed in the corresponding probability function. E.g., in example 4.2.1 we have:

$$P(X > 1) = P(X = 2) + P(X = 3) = \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$$

$$\text{Or in alternative notation: } P\big(X \in (1, \infty)\big) = \frac{1}{2}$$

More general, for each subset B of real numbers, $B \subset \mathbb{R}$ we have:

$$P(X \in B) = \sum_{x \in B} P(X = x),$$

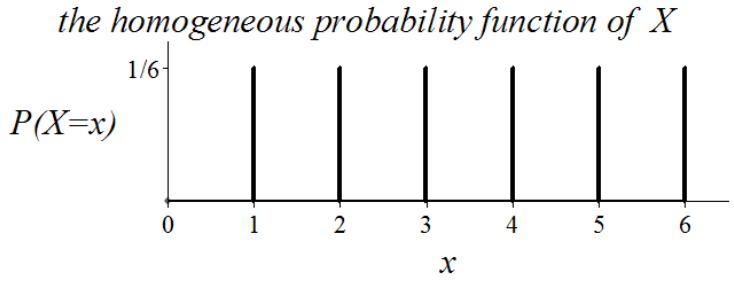
in which we sum over the values x of S_X, so actually over $x \in B \cap S_X$.

This way we again defined a probability function P on S_X (the axioms of Kolmogorov are fulfilled), so that $(\boldsymbol{S_X, P})$ is a probability space. The probabilities $P(X \in B)$ for each $B \subset S_X$ are, all together, called the **(probability) distribution** of the random variable X.
If these probabilities can be determined from the probability function of X, we can also give the probability distribution with the probability function $P(X = x)$ for all $x \in S_X$.

Example 4.2.4 The distribution of X = "the face up number after rolling an (unbiased) dice" is given by:

$$P(X = x) = \frac{1}{6}, \quad \text{for } x \in S_X = \{1, 2, 3, 4, 5, 6\}.$$

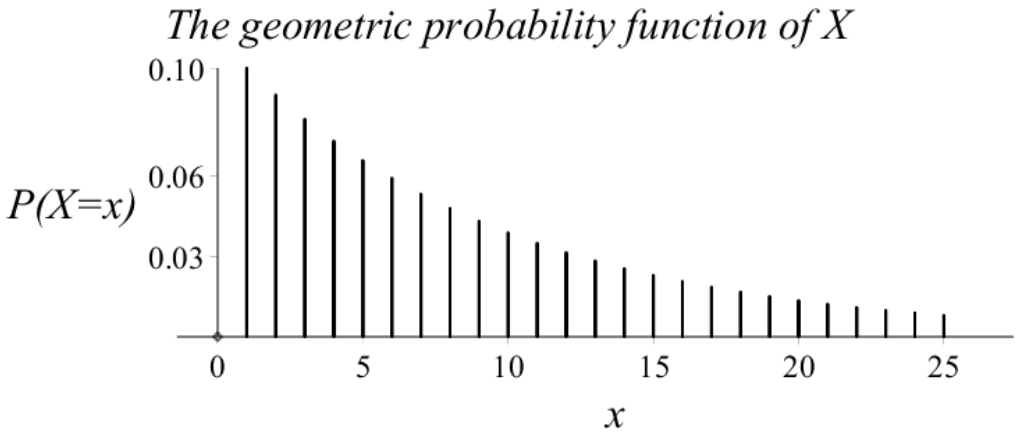
The probability distribution can be shown graphically as a bar graph of probabilities:



Since all probabilities are equal we will say that X has a **homogeneous distribution** on $\{1, 2, 3, 4, 5, 6\}$. ■

Example 4.2.5 A traveling salesman sells cookware sets to, on average, 1 out of 10 clients. One day he decides to visit customers until he has sold one set of cookware. Presume all costumers decide independently to buy with probability $\frac{1}{10}$, then we consider his trials to be Bernoulli experiments. If X is the number of visited customers on that day, then the probability of the event $X = k$, that he sells his set to the k^{th} customer, is given by the geometric formula (Property 3.3.11). X also has a so called geometric distribution with success probability $p = \frac{1}{10}$:

$$P(X = k) = \left(\frac{9}{10}\right)^{k-1}\left(\frac{1}{10}\right), \quad for\ k \in 1, 2, 3, \ldots.$$



For this probability function we can verify property 4.2.3 in this section:

1) $P(X = k) = \left(\frac{9}{10}\right)^{k-1}\left(\frac{1}{10}\right) > 0$ for all $k = 1, 2, 3, \ldots$

2) $\sum_{k \in S_X} P(X = k) = \sum_{k=1}^{\infty}\left(\frac{9}{10}\right)^{k-1}\left(\frac{1}{10}\right) = \left(\frac{1}{10}\right)\sum_{k=1}^{\infty}\left(\frac{9}{10}\right)^{k-1} = \left(\frac{1}{10}\right)\cdot\frac{1}{1-\frac{9}{10}} = 1$

This last summation is a consequence of the formula for a **geometric series:**
$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$, voor $|x| < 1$ (see the appendix Mathematical Techniques in this reader).
What is the probability that he has to visit more than 10 customers to sell a first set?
One could calculate:

$$P(X > 10) = 1 - P(X \leq 10) = 1 - \sum_{i=1}^{10} P(X = i) = \cdots$$

But if we reason that: "The probability that we have to visit more than 10 customers to sell a first set is the same as the probability that we do not sell a single set to the first 10 customers:

$$P(X > 10) = 0.9^{10} \approx 34.9\%.$$ ■

4.3 The expectation of a discrete random variable

Example 4.3.1 When we want to determine what the face up number is whilst rolling a **fair** dice, we can do this by determining the **average** face up number we roll, we can do this by rolling the dice a large number of times and keeping track of the results in a frequency table. Suppose $f_{1000}(x)$ is the relative frequency of the event "x is the face up number in 1000 rolls" and the result of the experiment is:

x	1	2	3	4	5	6	total
$f_{1000}(x)$	$\frac{180}{1000}$	$\frac{163}{1000}$	$\frac{164}{1000}$	$\frac{161}{1000}$	$\frac{162}{1000}$	$\frac{170}{1000}$	$\frac{1000}{1000}$

Then the mean $\overline{x}$ of the face up numbers in 1000 rolls can be given by.

$$\overline{x} = \frac{180 \cdot 1 + 163 \cdot 2 + 164 \cdot 3 + 161 \cdot 4 + 162 \cdot 5 + 170 \cdot 6}{1000}$$

Which can be computed similarly as:

$$\overline{x} = \sum_{x=1}^{6} x \cdot f_{1000}(x) = 1 \cdot \frac{180}{1000} + 2 \cdot \frac{163}{1000} + 3 \cdot \frac{164}{1000} + 4 \cdot \frac{161}{1000} + 5 \cdot \frac{162}{1000} + 6 \cdot \frac{170}{1000} = 3.472$$

The mean face up number $\overline{x}$ can be interpreted as the **weighted mean** of all values 1, 2, 3, 4, 5 and 6, where the relative frequencies can be used as the weighing factors. ■

We define X as the face up number at a roll of a dice and we know the probability function $P(X = x)$ for $x \in S_X = \{1, 2, \ldots, 6\}$, then we can determine, by analogy, the average of the values of x in S_X by weighing these values with $P(X = x)$, since $f_{1000}(x)$ is a an estimate of $P(X = x)$. This average w.r.t. the values of X is called the expected value $E(X)$ of the random variable X:

$$E(X) = \sum_{x=1}^{6} x \cdot P(X = x)$$

For an unbiased dice we have: $E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

Definition 4.3.2 The **expectation** or **expected value** $\boldsymbol{E(X)}$ of a discrete random variable X is given by

$$\boldsymbol{E(X) = \sum_{x \in S_X} xP(X = x)},$$

provided that this summation absolute convergent is (that is: $\sum_{x \in S_X} |x| \cdot P(X = x) < \infty$).

The condition on absolute convergence is mostly fulfilled in practice. In example 4.3.4 we will encounter an example, for which the summation does not converge (absolutely), so that

in that case the **expected value does not exist**.
Instead of the symbol $E(X)$, in literature $\boldsymbol{EX}$**,** $\boldsymbol{\mu}$ or $\boldsymbol{\mu_X}$ is also used, and in physics $< X >$.

Furthermore we will abbreviate the notation "summation over $x \in S_X$" by only mentioning x:

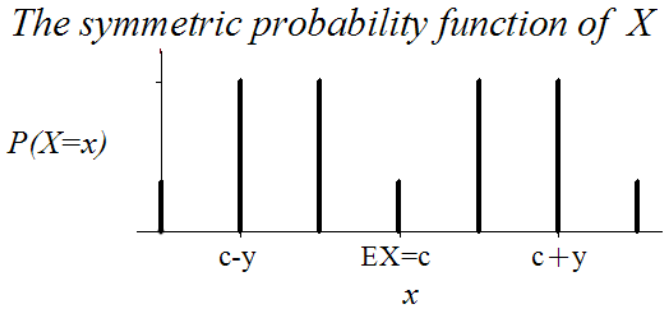
$$E(X) = \sum_x xP(X = x)$$

We have to bear in mind that the expected value of X or the expectation $E(X)$ can be interpreted as the average value of all possible values x of X, with as weighing factors the relevant probabilities $P(X = x)$: the summation of the weighing factors is, of course, 1. Since the variable X with range S_X and probability function $P(X = x)$ form a probability model for a population, $E(X)$ is often referred to as the **population mean**, and therefore denoted with μ, the Greek letter m for *mean*. In statistics we will come across another mean, **the sample mean $\overline{x}$,** the **average value of observations in a sample**, drawn from a (usually large) population. Both are called "mean" in daily life, but we will have to interpret whether the sample or population mean is meant. The distinction is conceptually important: the mean of the population μ is a fixed, but often unknown value, where the sample mean is just an estimate $\overline{x}$ of the real value of μ. Another sample will give you another estimate. Nevertheless we feel that, the larger the sample size, the closer the estimate $\overline{x}$ will be to μ. This intuitive observation is called the **frequency interpretation of the population mean $\boldsymbol{E(X)}$**: if an experiment to determine the value of a random variable is repeated very often (under equal circumstances), then the observed mean value will be close to $E(X) = \mu$.

Note that $E(X)$ is not necessarily "the middle value in the range", but it is a **measure of the center** (middle) of the probability distribution.
A more physical interpretation of $E(X)$ is that of a "**point of balance**": if we would see the X-axis as a weightless bar and the probabilities $P(X = x)$ as weights, hanging on the bar at the points x on the X-axis, then the bar is in balance if supported at the point $E(X)$.

Property 4.3.3 If the probability function is symmetric with respect to $x = c$, then $\boldsymbol{E(X) = c}$

This property is illustrated in the following graph:



In the expression $E(X) = \sum_{x \in S_X} xP(X = x)$ the values $c - y$ and $c + y$ have the same probability and the average of both values is c: the overall mean is, on average, $E(X) = c$.
In example 4.3.1 we have seen that X = "the result of one roll with a dice" has a symmetric probability function on {1,2,3,4,5,6}, where indeed $(X) = 3.5 = \frac{1+6}{2}$.

Example 4.3.4 The number X of visited clients by the traveling salesman in example 4.2.5 was geometrically distributed with success probability $p = \frac{1}{10}$.
The expected number of clients to visit can be computed, using the definition of expectation and some mathematical techniques:

$$EX = \sum_{k \in S_X} kP(X = k) = \sum_{k=1}^{\infty} k \cdot \left(\frac{9}{10}\right)^{k-1} \frac{1}{10} = \frac{1}{10} \sum_{k=1}^{\infty} k \cdot \left(\frac{9}{10}\right)^{k-1} \overset{*}{=} \frac{1}{10} \cdot \left(\frac{1}{1 - \frac{9}{10}}\right)^2 = 10$$

The equality $\overset{*}{=}$ follows from the summation of a geometric series, after differentiation (see the appendix "Mathematical Techniques").
$E(X)$ is sometimes confused with the so called **median**: that is the value M, such that $P(X \leq M) \geq 50\%$ and $P(X \geq M) \geq 50\%$.
After some computation we will find in this example that the median $M = 7$, different from the expectation $E(X) = 10$. (For more examples see exercise 16). ■

Example 4.3.4 Two players A and B toss an fair coin in turns. The player who flips tails first wins. Both bet a euro and A starts. Each time tail is not the outcome, the stakes are doubled. This seems attractive for A: in the first toss he already has a probability of 50% of winning, so $P(\text{"}A \text{ wins"}) > \frac{1}{2}$.
If we define X as the winning of A, then $X = 1$ if A flips tails in the first trial, $X = -2$ if A flips heads and B flips tails in his first trial, $X = 4$ if A flips tails in his second trial, etc.:

$$S_X = \{1, -2, 4, -8, \ldots\} = \{(-2)^n | n = 0, 1, 2, \ldots\}$$

The probability to have the first tail in the k^{th} trial is $\left(\frac{1}{2}\right)^k$, so the distribution of X can be given by:

$$P[X = (-2)^n] = \left(\frac{1}{2}\right)^{n+1}, \text{for } n = 0, 1, 2, \ldots$$

$E(X)$, the expected winning by A, can be computed in the following way:

$$\sum_{n=0}^{\infty} xP(X = x) = \sum_{n=0}^{\infty} (-2)^n \cdot \left(\frac{1}{2}\right)^{n+1} = \frac{1}{2} - \frac{1}{2} + \frac{1}{2} - \frac{1}{2} + \cdots,$$

This is a so called alternating series, that does not converge. It is neither absolute convergent:

$$\sum_{n=0}^{\infty} 2^n \cdot \left(\frac{1}{2}\right)^{n+1} = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots = \infty$$

Conclusion: $E(X)$ does not exist! ■

4.4 Functions of a discrete random variable; variance

Example 4.4.1 At a fair visitors are challenged to bet 4 euros for the following game: if the visitor rolls X as the face up number of an fair dice, he will be paid $(X-3)^2$ Euro.
E.g. if you roll $X=6$, he will pay you $(6-3)^2=9$ Euro and the profit will be $9-4$ Euro;
if you roll $X=3$, he will pay $(3-3)^2=0$ Euro and the profit will be $0-4$ Euro.
You are likely to decide to play the game if the expected profit is positive, so if the expected value of $Y=(X-3)^2-4$ is positive.
Y is a function of X and is a random variable as well: it takes on values 0, −3, −4, −3, 0, 5 if the face up number X equals 1, 2, 3, 4, 5 and 6, respectively: $S_Y=\{-4,-3,0,5\}$ and the distribution of Y can be given as follows: $P(Y=0)=P(X=1)+P(X=5)=\frac{1}{6}+\frac{1}{6}=\frac{1}{3}$

Similarly $P(Y=-3)=\frac{1}{3}$

and $P(Y=-4)=P(Y=5)=\frac{1}{6}$

Computing the expected profit:

$$E(Y)=\sum_{y\in S_Y} yP(Y=y)$$
$$=(-4)\cdot\frac{1}{6}+(-3)\cdot\frac{1}{3}+0\cdot\frac{1}{3}+5\cdot\frac{1}{6}=-\frac{5}{6}\text{ Euro,}$$

a negative expected profit, having the following frequency interpretation: If we would play this game very often, then the average profit will be -5/6 Euro, (or: every game will cost me on average almost 1 Euro)
The expected profit can also be computed by immediately using the distribution of X: we will have to weigh each profit value $(x-3)^2-4$ with the probability that $X=x$, so:

$$E[(X-3)^2-4]=\sum_{x\in S_X}[(x-3)^2-4]\cdot P(X=x)$$
$$=0\cdot\frac{1}{6}+(-3)\cdot\frac{1}{6}+(-4)\cdot\frac{1}{6}+(-3)\cdot\frac{1}{6}+0\cdot\frac{1}{6}+5\cdot\frac{1}{6}=-\frac{5}{6}\text{ Euro}\quad\blacksquare$$

In the previous example $Y=(X-3)^2-4$ is an example of a function $Y=g(X)$ of a discrete random variable X. In the example it was not difficult to derive the distribution of Y by computing the range of Y and determining the probabilities for Y by adding the probabilities of corresponding values of X.
Subsequently $E(Y)$ can be determined using the distribution of Y. We found another, more direct approach by computing $E(Y)=E(g(X))$, using the distribution of X.
This approach is given in the following property (without general proof).

Property 4.4.2 If X is a discrete random variable en g a (real) function, then:

$$\boldsymbol{E\big(g(X)\big)=\sum_{x\in S_X} g(x)P(X=x)},$$

(provided that the summation is absolute convergent).

Note that $Eg(X)$ means $E(g(X))$, e.g.: EX^2 means $E(X^2)$, and does **not** mean $(EX)^2$.

If ***Y* is a linear function of *X***, that is $Y=aX+b$ for any real constants $a,b\in\mathbb{R}$, then according property 4.4.2 we have:

$$E(aX+b) = \sum_{x\in S_X} (ax+b)\cdot P(X=x)$$
$$= \sum_{x\in S_X} ax\cdot P(X=x) + \sum_{x\in S_X} b\cdot P(X=x)$$
$$= a\cdot \sum_{x\in S_X} x\cdot P(X=x) + b\cdot \sum_{x\in S_X} P(X=x)$$
$$= a\cdot E(X) + b\cdot 1$$

We have proven the first part of the following property:

Property 4.4.3 If X is a discrete random variable and g and h are real functions, then for real constants $a, b \in \mathbb{R}$ we have:

$$\textbf{1)}\ \boldsymbol{E(aX+b) = aE(X)+b}$$
$$\textbf{2)}\ \boldsymbol{E[ag(X)+bh(X)] = aEg(X)+bEh(X)}.$$

The proof of 2) is analogous to the proof of 1) as given above.
Apparently in example 4.4.1 we could state that:
$E(Y) = E((X-3)^2-4) = E(X^2-6X+5) = E(X^2) - 6E(X) + 5.$
where $E(X) = 3.5$ (using symmetry) and $E(X^2) = (1^2+2^2+3^2+4^2+5^2+6^2)\cdot\frac{1}{6} = \frac{91}{6}$
Again we find: $E(Y) = \frac{91}{6} - 6\cdot\frac{7}{2} + 5 = -\frac{5}{6}$
In Probability Theory the expected values of the functions $g(X) = X^k$ have many applications. Property 4.4.2. enables us to compute its value:

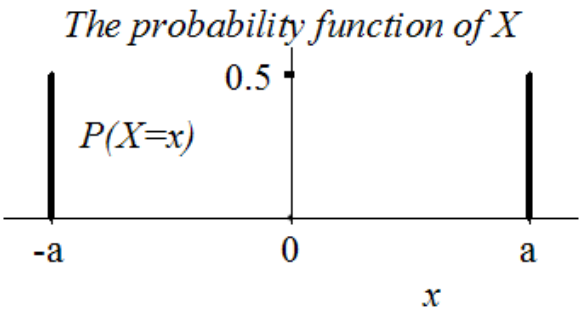
$$\boldsymbol{E(X^k) = \sum_x x^k P(X=x)}$$

Definition 4.4.4 $\boldsymbol{E(X^k)}$ is the $\boldsymbol{k^{th}}$ **moment** of the de random variable X, $k = 1, 2, 3, \ldots$.

Of course the k^{th} moment is only defined if the corresponding summation is absolutely convergent. Since this is mostly fulfilled in practice, we will not check this property every time. The first moment $E(X^1)$ is known as the expected value $E(X)$ of a random variable X. This weighted average can be considered as a **measure for the center** of the distribution of X (the median is a different measure for the center).
However, $E(X)$ does not tell us anything about the **magnitude of the differences** in the values of X. The following example introduces the concept of **measure of variation**.

Example 4.4.5
Consider the distribution given by:
$P(X=a) = P(X=-a) = 1$ with $a > 0$

Using symmetry we can immediately state that $E(X) = 0$, regardless of the value of a. The variation of X depends in this case on the difference $2a$ between the two points of S_X (or their deviations from 0): the variation will increase as a increases.

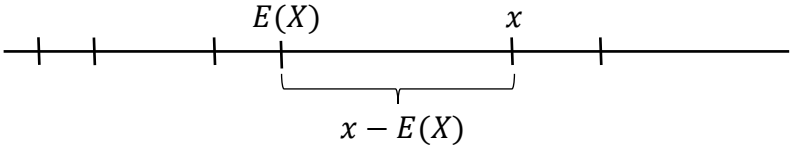


The moments of X can be computed ($k = 1, 2, 3, \ldots$):

$$E(X^k) = a^k \cdot \frac{1}{2} + (-a)^k \cdot \frac{1}{2} = \begin{cases} a^k & \text{for even } k \\ 0 & \text{for odd } k \end{cases}$$

The second moment or "the mean of the squared values" of X is $E(X^2) = a^2$: it increases as a increases, which is true for all even moments. ■

Searching for a measure of variation we could consider the differences $X - E(X)$: "the deviation of X from $E(X)$".
Then $E(X - E(X))$ is the weighted average of the deviations $x - EX$.
If $X = x$, then the deviation $x - E(X)$ has the weighting factor $P(X = x)$:



According to property 4.4.3 we have: $E[X - E(X)] = E(X) - E(X) = 0$.
This result is not surprising: the deviations can be positive or negative and *E*(*X*) is defined such that the "weighted differences" $[x - E(X)] \cdot P(X = x)$ are in total 0.
Being always zero, $E[X - E(X)]$ is not a suitable measure of variation. We could turn to the mean of the absolute deviations: $E\{|X - E(X)|\}$, or as an alternative:

$E\left[\left(X - E(X)\right)^2\right]$, the mean of the squared deviations.

Because of its convenient properties this last alternative is chosen. The brackets [ ] are usually omitted, in the same way as EX and EX^2 are alternative notations for $E(X)$ and $E(X^2)$.

Definition 4.4.6 The **variance** of *X* (notation: $\boldsymbol{var(X)}$ or $\boldsymbol{\sigma_X^2}$) is defined as

$$\boldsymbol{var(X) = E(X - \mu_X)^2}$$

According to property 4.4.2 we can compute $var(X)$ as follows:

$$var(X) = \sum_x (x - \mu_X)^2 \cdot P(X = x)$$

But we can also apply property 4.4.3 to express $var(X)$ in the first and second moment:

$$\begin{aligned} \boldsymbol{var(X)} = E(X - \mu_X)^2 &= E(X^2 - 2\mu_X \cdot X + \mu_X^2) \\ &= E(X^2) - 2\mu_X \cdot E(X) + \mu_X^2 \\ &= \boldsymbol{E(X^2) - \mu_X^2} \end{aligned}$$

Note that μ_X is a (fixed) real number, so $E(\mu_X) = \mu_X$ or $E\big(E(X)\big) = E(X)$.
Similarly: $E(\mu_X^2) = \mu_X^2$

In most cases this formula is preferred instead of the definition, for computational reasons:
$var(X)$ = "the 2nd moment minus the square of the 1st moment"
This and other properties of expectation and variance should be known by heart to be applied if necessary.

Example 4.4.7 If X is the number of sixes in two rolls of a fair dice, then:

$$P(X=2)=\frac{1}{36}, P(X=1)=\frac{10}{36} \text{ and } P(X=0)=\frac{25}{36}$$

The computation of $E(X)=\sum_x xP(X=x)$ and $var(X)=E(X^2)-\mu_X^2$ performed using a neat table (avoiding computational errors).

x	0	1	2	Total
$P(X=x)$	$\frac{25}{36}$	$\frac{10}{36}$	$\frac{1}{36}$	1
$x\cdot P(X=x)$	$0\cdot\frac{25}{36}$	$1\cdot\frac{10}{36}$	$2\cdot\frac{1}{36}$	$\frac{1}{3}=E(X)=\mu_X$
$x^2\cdot P(X=x)$	$0\cdot\frac{25}{36}$	$1\cdot\frac{10}{36}$	$4\cdot\frac{1}{36}$	$\frac{14}{36}=E(X^2)$
$(x-\mu_X)^2\cdot P(X=x)$	$\left(-\frac{1}{3}\right)^2\cdot\frac{25}{36}$	$\left(\frac{2}{3}\right)^2\cdot\frac{10}{36}$	$\left(\frac{5}{3}\right)^2\cdot\frac{1}{36}$	$\frac{10}{36}=var(X)$

And:
$var(X)=E(X^2)-\mu_X^2$
$=\frac{14}{36}-\left(\frac{1}{3}\right)^2=\frac{10}{36}$

We added the last row to compare the "direct" computation (using the definition) to the computational formula $var(X)=E(X^2)-\mu_X^2$.

In this example clearly $\frac{14}{36}=E(X^2)\neq(EX)^2=\left(\frac{1}{3}\right)^2$. In general, the mean of the squares of numbers is greater than the square of the mean, e.g. $17=\frac{3^2+5^2}{2}>\left(\frac{3+5}{2}\right)^2=16$ ■

Because $var(X)$ is defined as an average of squares, $var(X)=E(X^2)-\mu_X^2$ cannot be negative, so $E(X^2)\geq\mu_X^2$.

The equality $var(X)=0$ can occor if all terms in the summation $\sum_x(x-\mu_X)^2\cdot P(X=x)$ equals zero. Consequently, if $P(X=x)>0$, then $x=\mu_X$:

X can only attain one value (with probability 1): $P(X=\mu_X)=1$.

In such a case we will call the distribution of X **degenerate**: there is no more chance, we know for sure that the experiment will lead to one outcome (μ_X).

Since $var(X)$ is an average of squared deviations $(x-\mu_X)^2$, the unit of $var(X)$ is the square of the unit of X: if X is a length in cm is, then $var(X)$ is in cm^2.

Returning to the original unit we will have to root the variance.

But note that $\sqrt{E(X-\mu_X)^2}\neq E|X-\mu_X|$).

Definition 4.4.8 the **standard deviation** of X (notation: σ_X) is the square root of the variance:

$$\boldsymbol{\sigma_X=\sqrt{var(X)}}$$

So, $var(X)=\sigma_X^2$ and $\sigma_X=\sqrt{var(X)}$ are exchangable measures of variation.

In practice the standard deviation will be used more often because it has the same unit as X.

Property 4.4.9 (Properties of variance and standard deviation)

a. $\boldsymbol{var(X)\geq 0}$ and $\boldsymbol{\sigma_X\geq 0}$.

b. $\boldsymbol{var(X)=E(X^2)-\mu_X^2}$ (the computational formula).

c. if $var(X)>0$, so if X is not degenerate, we have $E(X^2)>(EX)^2$.

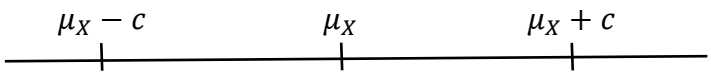
d. $\boldsymbol{var(aX+b)=a^2\cdot var(X)}$ and $\boldsymbol{\sigma_{aX+b}=|a|\cdot\sigma_X}$.

Proof: a., b. and c. are discussed in the text prior to the property.
d. claims that if you apply a **linear transformation** to X, a shift $(+b)$ does not affect the measures of variation, but a multiplication factor does:

$$\begin{aligned} var(aX+b) &= E[aX+b-E(aX+b)]^2 \\ &= E[aX+b-aE(X)-b]^2 \\ &= a^2 \cdot E(X-\mu_X)^2 \\ &= a^2 \cdot var(X) \end{aligned}$$

And: $\sigma_{aX+b} = \sqrt{var(aX+b)} = \sqrt{a^2 \cdot var(X)} = |a| \cdot \sigma_X$ ■

If we consider the measures of center and variation, μ_X and σ_X^2 (or σ_X), of a random variable X, we can state something about the probability of X in a symmetric interval around μ_X.
In the sketch we see such an interval with bounds deviating $c > 0$ from μ_X.



Property 4.4.10 (**Chebyshev`s inequality**)

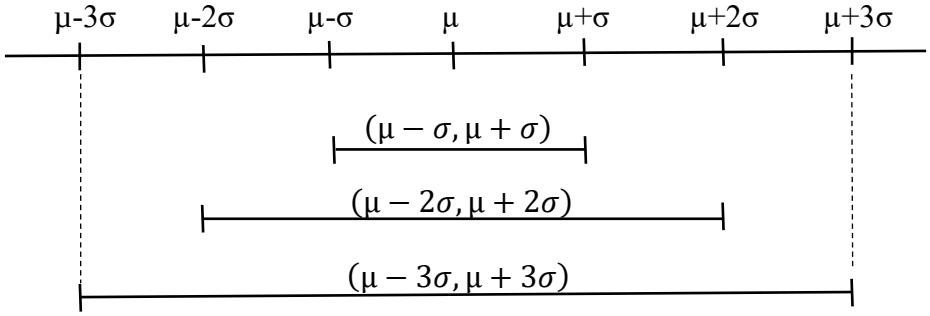
For any real number $c > 0$, we have: $\boldsymbol{P(|X - \mu_X| \geq c) \leq \frac{var(X)}{c^2}}$

We will not prove this theoretical result, but it allows us to give an interpretation with respect to the standard deviation. The inequality is valid for **any** random variable X and gives us an **upper bound of the probability** of values **outside the interval** $(\boldsymbol{\mu_X - c, \mu_X + c})$, so deviating more than c from μ.
The relation with the standard deviation can be made by choosing $c = k \cdot \sigma_X$.
Then the interval is $(\mu_X - k \cdot \sigma_X, \mu_X + k \cdot \sigma_X)$ and the upper bound of probability of observing values outside this interval is $\frac{var(X)}{c^2} = \frac{var(X)}{k^2 \sigma_X^2} = \frac{1}{k^2}$.
If $c < \sigma_X$ the upper bound of the probability is greater than 1 (not very informative), but choosing k greater we find, e.g.:

If $k = 2$, then $P(|X - \mu_X| \geq 2\sigma_X) \leq \frac{1}{2^2} = 25\%$

If $k = 3$, then $P(|X - \mu_X| \geq 3\sigma_X) \leq \frac{1}{3^2} \approx 11\%$

The latter means that there is a probability of at most 11% that X deviates more than 3 standard deviations from the mean μ and the probability to find a value within the interval $(\mu_X - 3 \cdot \sigma_X, \mu_X + 3 \cdot \sigma_X)$ is at least 89%.



Chebyshev`s rule is valid for any distribution, but the so called **Empirical Rule** is only valid for distributions that are (approximately) symmetric and bell (hill) shaped. The normal distribution, to be discussed in chapter 6, is the "standard" of such a distribution.

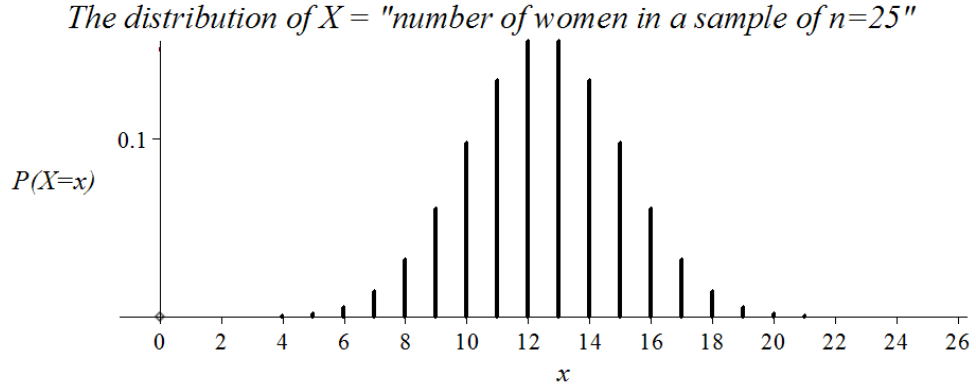
Empirical rule
If the graph of the distribution of X shows a bell shape, then the approximately probabilities for X having a value within the interval

- $(\mu-\sigma,\mu+\sigma)$ is **68%.**
- $(\mu-2\sigma,\mu+2\sigma)$ is **95%.**
- $(\mu-3\sigma,\mu+3\sigma)$ is **99.7%.**

This rule is sometimes referred to as the 68-95-99.7%-rule: in chapter 6 we will show its validity when discussing the normal distribution, on which the rule is based.

Example 4.4.11 In Enschede 50% of the adults are female. If we choose 25 inhabitants of Enschede for a survey, we will do so without replacement (you will not choose one person twice). Since the population is very large the probability of choosing one person twice is negligibly small, so we could as well assume that we draw with replacement (guaranteeing independence). Then for X, the number of women in the sample, probabilities can be given, using the binomial formula with parameters: sample size $n=25$ and probability of success $p=\frac{1}{2}$:

$$P(X=k)=\binom{25}{k}\left(\frac{1}{2}\right)^{25}, \quad \text{where } k=0,1,2,3,\ldots$$



The probability function of X has indeed a bell shaped graph, for which the Empirical rule applies. We need the values of $\mu=E(X)$ and $\sigma^2=var(X)$:
$E(X)=\frac{1}{2}\cdot 25=12.5$ (intuitively: we expect half of the sample to be female).
$var(X)=np(1-p)=6.25$. These formulas will be discussed in the next section.
We will compare Chebyshev`s rule and the Empirical rule to the real values of the probabilities for this distribution in the following table:

Interval	$P(X$ in interval$)$	Probability accord. Empirical rule	Probability accord. Chebyshev`s rule
$(\mu-\sigma,\mu+\sigma)=(10,15)$	**57.6%**	68%	$\geq 0\%$
$(\mu-2\sigma,\mu+2\sigma)=(7.5,17,5)$	**95.7%**	95%	$\geq 75\%$
$(\mu-3\sigma,\mu+3\sigma)=(5,20)$	**99.6%**	99.7%	$\geq 89\%$

The actual probabilities are close to those of the Empirical bounds and the values of Chebyshev`s rule prove to be lower bounds. ■

4.5 The binomial, hypergeometric, geometric and Poisson-distribution

The Binomial distribution

In probability theory and statistics **Bernoulli experiments or (Bernoulli) trials** play an important role: whether or not a specific phenomenon in repeated experiments occurs can be characterized by the occurrence of a success (=1) or a failure (=0) in each of the experiments. We will only call these Bernoulli experiments if they are independent. In a large series of Bernoulli experiments it is a natural choice to estimate the success probability by computing the proportion of successes in n repetitions.

Example 4.5.1 From a population of voters we draw n times with replacement a person and ask him/her whether he/she will vote for a specific party A. If we define X as the number of party A voters, the event $\{X = k\}$ occurs if we have k successes (party A voters) in n Bernoulli experiments (n persons), where the success probability p = "the probability that an arbitrary person votes party A".
Applying property 3.3.10 we find:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{met } k \in S_X = \{0,1,2,\dots,n\}$$

Number of orders of "k successes and $n - k$ failures."

Probability of "first k successes and then $n - k$ failures".

This is indeed a probability distribution, since

1) $P(X = k) \geq 0$ for $k = 0, 1, \dots, n$ (and, of course, $0 \leq p \leq 1$)

2) $\sum_{k \in S_X} P(X = k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1$,

according to Newton`s Binomial Theorem: $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$ (see the appendix)■

Definition 4.5.2 (**the binomial distribution**)
X is **binomially distributed with parameters n and p**, for all $n = 1, 2, \dots$ and $p \in [0, 1]$, if the probability function of X is given by:

$$\boldsymbol{P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}}, \quad \text{where } k = 0,1,2,\dots,n$$

Short notations: X is $B(n, p)$-distributed, or: $\boldsymbol{X \sim B(n, p)}$

One can apply the binomial distribution as a probability model of real life situations, whenever there is a series of n similar experiments for which the conditions of Bernoulli trials hold, i.e.:

- A phenomenon occurs (or does not occur) at a fixed success rate p (or failure rate $1-p$)
- Independence of the trials.

When we have **random draws** from a population the independence is only secured if we draw **with replacement**, which was the case when choosing the voters in example 4.5.1. Another situation, in which the conditions apply, is the repeated execution of experiments with two possible outcomes "under the same conditions", such as the flipping of a coin, 3 times in a row in example 4.2.1; we found a probability function that we can label as a $B\left(3,\frac{1}{2}\right)$-distribution. The number of sixes after two rolls of a dice in example 4.4.7 is apparently binomially distributed with parameters $n=2$ and $p=\frac{1}{6}$.
We found $E(X)=\frac{1}{3}$ and $var(X)=\frac{10}{36}$, which satisfy the general formulas

$$E(X)=n\cdot p \text{ and } var(X)=np(1-p),$$

The expected number of sixes in 60 rolls of a dice is intuitively equal to $\frac{1}{6}\cdot 60=10$, confirming the same general formulas given below:

If X is $B(n,p)$-distributed, then expected value and variance are given by:

$$\boldsymbol{E(X)=np} \text{ and } \boldsymbol{var(X)=np(1-p)}.$$

These formulas can be derived from the definitions of EX and $var(X)$, e.g.:

$$E(X)=\sum_{k\in S_X} k\cdot P(X=k)=\sum_{k=0}^{n} k\cdot\binom{n}{k}p^k(1-p)^{n-k}=\cdots=np$$

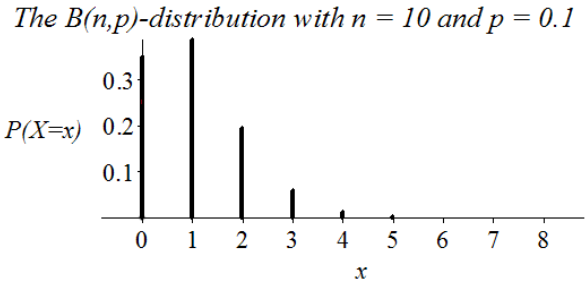
This derivation requires careful analytic work (e.g. application of Newton's Binomium). But in chapter 5 we will develop a more probabilistic and insightful approach to derive the formulas for $E(X)$ and $var(X)$ presented in this chapter.

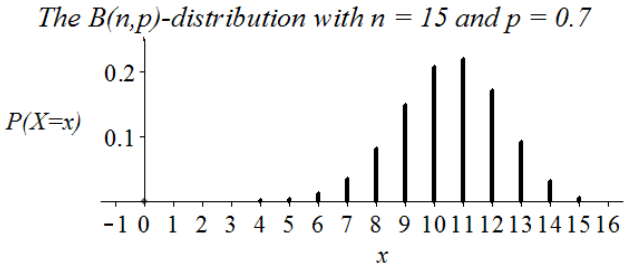
We will highlight some special values of n and p, the parameters of the $B(n,p)$-distribution:

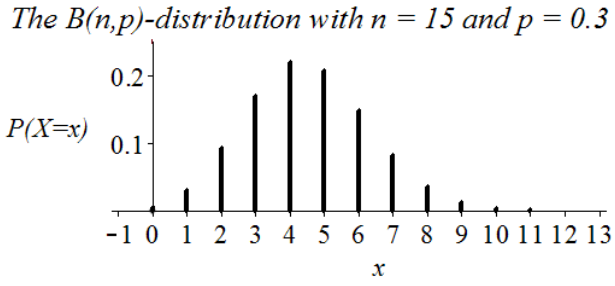
- If $\boldsymbol{p=1}$ ("success guaranteed"), then $P(X=n)=1$ and $E(X)=n$: X has a **degenerate distribution** in n. Similarly, if $p=0$, then $P(X=0)=1$ and $E(X)=0$.

- If $\boldsymbol{n=1}$**,** that is, if only one trial is conducted (one shot on the basket, the quality of one product is assessed, etc.), X is said to have an **alternative distribution with success probability $\boldsymbol{p}$,** which is a $B(1,p)$-distribution. It follows that:
  $P(X=1)=p$ and $P(X=0)=1-p$, so:
  $E(X)=\sum_x xP(X=x)=1\cdot p+0\cdot(1-p)=p$
  And: $E(X^2)=\sum_x x^2P(X=x)=1^2\cdot p+0^2\cdot(1-p)=p$
  We find: $var(X)=E(X^2)-(EX)^2=p(1-p)$, the variance of a $B(1,p)$-distribution.

Below we give some graphs of binomial distributions with different parameters n and p:







To simplify the computation of probabilities for $B(n,p)$- distributed variables X, so called cumulative binomial tables can be found at the end of this reader.
These tables contain probabilities of the shape $P(X \leq c) = \sum_{k=0}^{c} \binom{n}{k} p^k (1-p)^{n-k}$.
For, e.g., $n = 15$ and $p = 0.3$ (see the graph on the previous page) we can find:

- $P(X \leq 5) = 0.7216$
- $P(X = 5) = P(X \leq 5) - P(X \leq 4) = 0.7216 - 0.5155 = 0.2061$
  You might check this result with the binomial formula: $P(X = 5) = \binom{15}{5} 0.3^5 0.7^{10}$
- $P(X > 5) = 1 - P(X \leq 5) = 1 - 0.7216 = 0.2784.$
- Sometimes (as is the case for the tables in this reader) the probability tables are only available for success probabilities $p \leq 0.5$. If $\boldsymbol{p > 0.5}$ we can use this table anyhow, if we compute the probability of the corresponding number of failures.
  If Y is, e.g., $B(15, 0.7)$-distributed (see the graphs above), the probability of a failure is $0.3 < 0.5$: $X = 15 - Y$, the **number of failures**, is $B(15, 0.3)$-distributed, so:
  $P(Y = 10) = P(X = 15 - 10) = 0.2061.$

The Hypergeometric distribution

Example 4.5.3 A hotel manager wants to acquire flatscreens for 5 of his hotel rooms. He considers the opportunity of buying from a bankrupt competitor, who offers 20 used flatscreens of which it is known that 5 have serious defects. If the manager has to buy a flatscreens at random (no quality control allowed), how many flatscreens does he have to buy such that he has a probability of at least 90% that 5 of the flatscreens are working well?
To solve this problem let us first suppose he buys 8 flatscreens: 75% of them, so 6 flatscreens are expected to be working well, but we want to compute the probability that at least 5 work well.
The choice of 8 flatscreens can be seen as 8 draws without replacement from a population with 15 well and 5 not working flatscreens. So the hypergeometric formula applies (2.2.2).
The probability of (exactly) $X = 5$ well working flatscreens is:

$$P(X = 5) = \frac{\binom{15}{5}\binom{5}{3}}{\binom{20}{8}} \approx 23.8\%$$

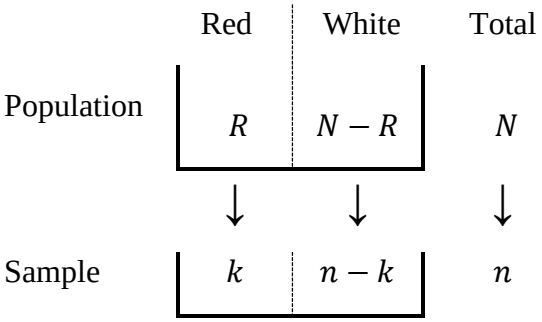
And $P(X \geq 5) = \sum_{k=5}^{8} \frac{\binom{15}{k}\binom{5}{8-k}}{\binom{20}{8}} \approx 94\%$

	Well	Not	Total
Available	15	5	20
	↓	↓	↓
Choice	5	3	8

Conclusion: purchasing 8 of the 20 flatscreens fulfills the condition of a 90% probability of at least 5 working flatscreens. (Check that this is not the case, when buying 7) .
Note that the expected number $E(X)$ of working flatscreens has the shape of np: $8 \cdot \frac{15}{20} = 6$. ■

If the probability function of the random variable X can be given by the hypergeometric formula, X is said to have a hypergeometric distribution. We can apply this distribution whenever we consider a number of **random draws without replacement** from a so called **dichotomous** population: consisting of elements which do or do not have a specific property, such as the red and white balls in property 2.2.2.

$X =$ "number of red balls in the sample"



Definition 4.5.4 (**the hypergeometric distribution**)
X is hypergeometricly distributed (with parameters N, R and n) if

$$\boldsymbol{P(X=k)} = \frac{\binom{\boldsymbol{R}}{\boldsymbol{k}}\binom{\boldsymbol{N-R}}{\boldsymbol{n-k}}}{\binom{\boldsymbol{N}}{\boldsymbol{n}}}, \qquad \text{where } k = 0, 1, 2, \ldots, n$$

For the number X in example 4.5.3 we would find: $P(X=2) = \frac{\binom{15}{2}\binom{5}{6}}{\binom{20}{8}} = ?$
Two working flatscreens imply that he bought 6 not working flatscreens. But only 5 not working flatscreens are available. This explains why the (unknown) binomial coefficient $\binom{5}{6}$, "the number of combinations of 6 chosen from 5" should be 0: the event is impossible.
If we define that $\binom{5}{6} = 0$, the probability $P(X=2)$ is 0 as well.
The expected proportion of working flatscreens in the sample is equal to the proportion $\frac{15}{20}$ of working flatscreens in the population, so the expected number is $E(X) = 8 \cdot \frac{15}{20} = 6$.

In general: $\boldsymbol{E(X) = np}$ and $\boldsymbol{var(X) = np(1-p) \cdot \frac{N-n}{N-1}}$, where $\boldsymbol{p = \frac{R}{N}}$

In chapter 5 we will prove the correctness of these formulas.

Random draws from a dichotomous population lead to the hypergeometric distribution of the number of "successes" if we draw without replacement, but on the other hand, if the draws are with replacement, we can use the binomial distribution: in that case the draws should be independent.
Using the sketch of the vase at 4.5.4, we have dependence or drawing without replacement:

$$P(\text{"1}^{\text{st}}\text{ drawn ball is red"}) = P(\text{"2}^{\text{nd}}\text{ drawn ball is red"}) = \frac{R}{N}$$

$$\text{but: } P\left(2^{\text{nd}}\text{ red} \,\middle|\, 1^{\text{st}}\text{red}\right) = \frac{R-1}{N-1}$$

The probability of the "2^{nd} red" depends on the result of the first draw.
But if the two subpopulations are (very) large we have approximately $\frac{R}{N} \approx \frac{R-1}{N-1}$.
This is, e.g.. the case if the population consists of "the Dutch voters" and considering a large party. For relatively small numbers of draws (e.g. $n = 1000$ persons out of $N = 10$ milion voters), which are much smaller than R and $N - R$, we have approximate independence.
This is formulated in the following (without proof):

Property 4.5.5 For relatively large R and $N - R$ and relatively small n the hypergeometric distribution with parameters N, R and n can be approximated by a $\boldsymbol{B\left(n, \frac{R}{N}\right)}$-distribution**.**

Note that the variances of the hypergeometric and binomial distributions under these conditions are almost equal: $np(1-p) \cdot \frac{N-n}{N-1} \approx np(1-p)$.

A (quite strict) rule of thumb for approximating by the binomial distribution in property 4.5.5 is:

$$\boldsymbol{N > 5n^2}$$

E.g., if the sample size $n = 1000$ and the sampling is without replacement, we can use the binomial approximation when $N > 5 \cdot 1000^2 = 5\,000\,000$

The Geometric distribution

We applied this distribution examples 4.2.5 and 4.3.3 for a traveling salesman who visited clients until he sold a cookware set.
The independent sales trials all have a success probability $\frac{1}{10}$.
Determination of the number of Bernoulli trials until a success occurs, can be encountered in many situations, such as the rolling of a dice to get a 6 face up ($p = \frac{1}{6}$)., checking products until one is substandard or participating in a lottery until you win a prize.

Definition 4.5.6 X has a **geometric distribution with parameter** $\boldsymbol{p} \in (0, 1]$, if

$$\boldsymbol{P(X = k) = (1-p)^{k-1}p}\,, \text{ where } k = 1, 2, \ldots.$$

If $p = 1$ the distribution is degenerate: $P(X = 1) = 1$.
Using the properties of geometric series (see the appendix "Mathematical Techniques") we can prove:

$$\boldsymbol{E(X) = \frac{1}{p}} \quad \text{en} \quad \boldsymbol{var(X) = \frac{1-p}{p^2}}$$

The following formula is convenient whenever we have to compute a summation of geometric probabilities:

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

The reasoning is as follows: the probability that we need **more than k trials** to score a success equals the probability that we are **not successful in the first k trials**.

The Poisson distribution

Example 4.5.7 An academic hospital reported that**,** on average, the demand of IC units for babies is 4 on an arbitrary day. How many units should be available to ensure that the probability that the demand exceeds the number of available IC-units is less than 0.001?
To answer this question we could define X as the number of demanded IC-units on an arbitrary day.
Which distribution can we apply for X?
Subsequently we could use this distribution to compute the minimum number m such that $P(X > m) \leq 0.001$.
If we would assume that:

- on that day there are (e.g.) 1000 just born babies in the region of the hospital.
- every just born baby will require a IC-unit with probability $\frac{4}{1000}$, independent of all other babies.

Then the demand X can be modelled with a $B\left(1000, \frac{4}{1000}\right)$-distribution with the given expected demand $E(X) = np = 4$.
Of course, using this distribution can cause computational problems, e.g. in computing the binomial coefficients $\binom{1000}{k}$ and factors $\left(\frac{4}{1000}\right)^k \left(\frac{996}{1000}\right)^{1000-k}$
Furthermore the number of just born babies is not known exactly and will vary from day to day. Let us assume that there are n just born babies, that independently need IC with probability $p = \frac{4}{n} = \frac{\mu}{n}$. Then we can rewrite the binomial formula:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)\ldots(n-k+1)}{k!} \cdot \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k}$$

If we know that n is large, approaching infinity, we can mathematically show that:

$$\lim_{n\to\infty} P(X = k) = \lim_{n\to\infty} \frac{\mu^k}{k!} \cdot \frac{n(n-1)\ldots(n-k+1)}{n \cdot n \cdot \ldots\ldots \cdot n} \cdot \left(1 - \frac{\mu}{n}\right)^{n-k} = \frac{\mu^k e^{-\mu}}{k!}$$

This result is obtained using the following limits:
$\lim_{n\to\infty} \frac{n}{n} = \cdots = \lim_{n\to\infty} \frac{n-k+1}{n} = 1$ and $\lim_{n\to\infty} \left(1 - \frac{\mu}{n}\right)^{n-k} = \lim_{n\to\infty} \left(1 - \frac{\mu}{n}\right)^{n} = e^{-\mu}$.
The last limit is a consequence of the "standard limit" $\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x$ (see your calculus book).

The limit distribution of X can be used approximately for large n
$P(X = k) = \frac{\mu^k e^{-\mu}}{k!}$, where $k = 0,1,2$ ....

This distribution is called the Poisson distribution and can be used to solve the question at the start of this example: determine m such that $P(X > m) \leq 0.001$. X has a Poisson distribution with a mean demand μ = 4. For this and other values of μ cumulative probabilities $P(X \leq m)$ are given in the Poisson table.
Since $P(X > m) \leq 0.001$ is equivalent to $P(X \leq m) \geq 0.99$ we find $m = 11$ IC-units ■

Definition 4.5.8 X has a **Poisson distribution** with parameter $\mu > 0$ if

$$\boldsymbol{P(X = k) = \frac{\mu^k e^{-\mu}}{k!}}, \qquad \text{for } k = 0,1,2 \ldots.$$

This is a probability function: all probabilities are at least 0 and the sum of all probabilities is 1. To prove this we use the Taylor series of the function e^x about 0:

$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ (see the appendix Mathematical Techniques)

$$\sum_k P(X = k) = \sum_{k=0}^{\infty} \frac{\mu^k e^{-\mu}}{k!} = e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} = e^{-\mu} \cdot e^{\mu} = 1$$

In example 4.5.7 we interpreted μ as the mean (demand). Indeed μ is the expected value:

$$\boldsymbol{E(X) = \sum_{k=0}^{\infty} k \cdot \frac{\mu^k e^{-\mu}}{k!} = e^{-\mu} \sum_{k=1}^{\infty} \frac{\mu^k}{(k-1)!} = \mu e^{-\mu} \cdot \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} = \mu e^{-\mu} \cdot e^{\mu} = \mu}$$

For derivation of the formula of the variance, $var(X) = \mu$, we need some mathematical tricks: We will use: $var(X) = E(X^2) - (EX)^2 = E\big(X(X-1)\big) + E(X) - (EX)^2$
In the last expression is $E(X) = EX = \mu$ and

$$E(X(X-1)) = \sum_{k=0}^{\infty} k(k-1) \cdot \frac{\mu^k e^{-\mu}}{k!} = \mu^2 e^{-\mu} \sum_{k=2}^{\infty} \frac{\mu^{k-2}}{(k-2)!} = \mu^2 e^{-\mu} \cdot e^{\mu} = \mu^2$$

So: $\boldsymbol{var(X) = E\big(X(X-1)\big) + E(X) - (EX)^2 = \mu^2 + \mu - \mu^2 = \mu}$

Poisson probabilities are given in (cumulative) probability tables for $P(X \leq c)$: in example 4.5.7 we applied the table to compute the minimum m of available IC units such that $P(X \leq m) \geq 0.99$.

The property we used in this example can be state in general as:

Property 4.5.9 If $\boldsymbol{X}$ has a $\boldsymbol{B(n,p)}$-distribution with **"large $\boldsymbol{n}$ and small $\boldsymbol{p}$"**, then X has **approximately** a **Poisson** distribution with parameter $\boldsymbol{\mu = np}$.

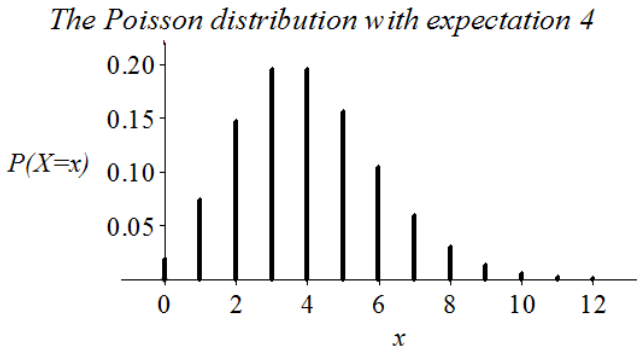
A **rule of thumb** for applying this approximation is:

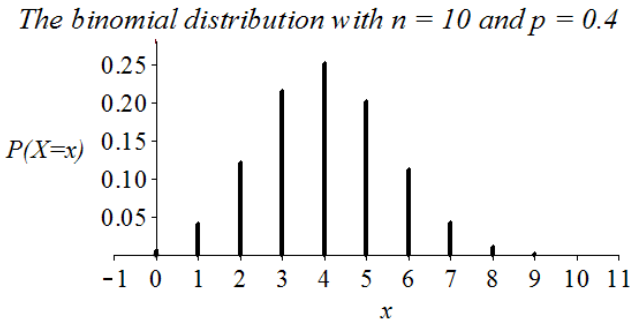
$$\boldsymbol{n > 25} \text{ and } \boldsymbol{np < 10} \text{ or } \boldsymbol{n(1-p) < 10}.$$

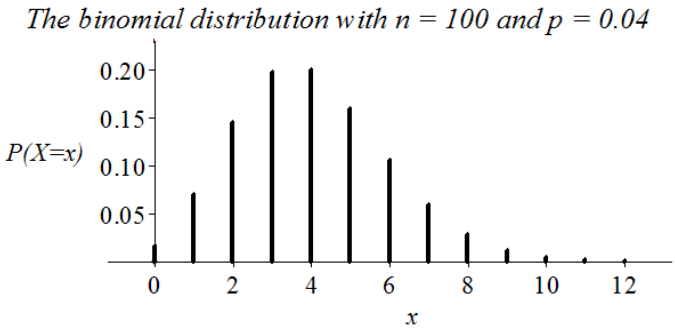
By the way, these approximations are also applicable in case of "large n and large p" (p close to 1)", because we noticed before that if the number of successes X is $B(n,p)$ with p close to 1, then the number of failures, $n-X$, is $B(n,1-p)$, with $1-p$ close to 0.

As an illustration of property 4.5.9 we will compare the probabilities of the Poisson distribution with μ = 4 with the $B\left(10,\frac{2}{5}\right)$- and the $B\left(100,\frac{1}{25}\right)$-distribution.
Note that all three distributions have expectation $E(X)=4$.

Distribution	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X=4)$	$P(X=5)$	$P(X=6)$	$P(X=7)$	$P(X=8)$
$B\left(10,\frac{2}{5}\right)$	0.006	0.040	0.121	0.215	0.251	0.201	0.111	0.042	0.011
$B\left(100,\frac{1}{25}\right)$	0.017	0.070	0.145	0.197	0.199	0.160	0.105	0.059	0.029
Poisson μ= 4	0.018	0.073	0.147	0.195	0.195	0.156	0.104	0.060	0.030

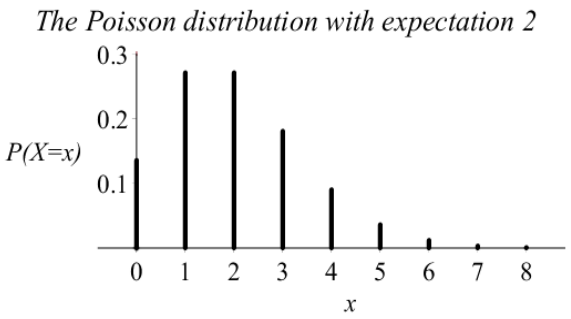


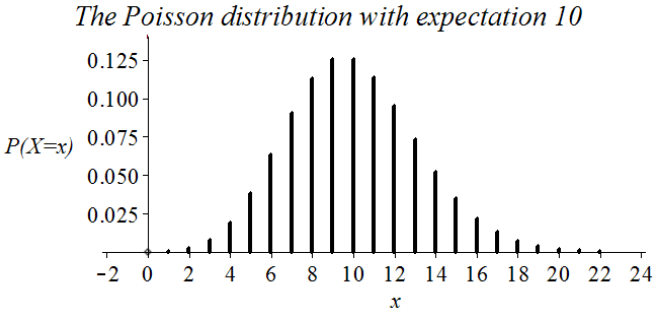




Property 4.5.9 indicates in which situation we could apply the Poisson distribution: often we count the **number of rare events**, that is, events that occur for each element at a small probability rate. Usually **the area and the time interval** for the occurrence of these events are **restricted**. Examples are the number of car thefts in a big city during one day, acute appendices surgery in a hospital during a week, or the number of red mushrooms in an acre of wood.
In many of these examples only the "mean number" (= the expectation μ) of events is known: often based on past experience (statistics): this value of μ is the parameter of the Poisson distribution to be used. Enlarging or decreasing the time interval or the area will cause a proportional change of the parameter μ: if the number of earthquakes in a country is, on average, 2 in one year, the mean number in two years is 4.

Some more shapes of the Poisson distribution are shown in the graphs below: note that the largest probabilities are close to μ.





In general the Poisson distribution is not symmetric, as the Poisson distribution shows, but for larger values of μ the graphs become more symmetric, e.g. if μ = 10, the graph looks quite "bell shaped" and the Empirical rule applies. ■

We summarize the discrete distributions and their characteristics in the following property:

Property 4.5.10 (Common discrete distributions and their characteristics)

Distribution	**Probability function**	$E(X)$	$var(X)$	**Example**
Homogeneous on 1, 2, …, N	$P(X=x)=\frac{1}{N}, x=1,2,..,N$	$\frac{N+1}{2}$	- -	result of one roll of a dice
Alternative (p)	$P(X=1)=p$ $P(X=0)=1-p$	p	$p(1-p)$	Dice results is 6 (X=1) or not (X=0)
Binomial $B(n,p)$	$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$, $x=0,\ 1,\ 2,\ \ldots,n$	np	$np(1-p)$	Number of sixes in 30 rolls of a dice
Geometric (p)	$P(X=x)=(1-p)^{x-1}p$, $x=1,\ 2,\ 3,\ \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	Number of rolls of a dice until 6 occurs.
Poisson (μ)	$P(X=x)=\frac{\mu^x e^{-\mu}}{x!}$, $x=0,1,\ldots$	μ	μ	Number of clients that enter an office in 10 minutes.
Hyper-geometric (R,N,n)	$P(X=x)=\frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}$, $x=0,1,\ldots,n$	np $p=\frac{R}{N}$	$np(1-p)\cdot\frac{N-n}{N-1}$	Number of girls if we choose 5 from a group of 10 boys and 12 girls.

4.6 Exercises

1. The probability function of X is given in the following table:

x	-5	-2	0	1	3
$P(X = x)$	0.1	0.2	0.2	0.4	0.1

Sketch the graph (bar graph) of the probability function and compute:

a. $P(X > 0)$,
b. $E(X)$,
c. $E(X^2)$ and
d. the variance and the standard deviation of X (using the results of b. and c.).

2. Give the distribution (probability function) of the variable X = "the first digit of the number plate of a passing car".
Sketch the probability function and determine $E(X)$, $E(X^2)$ and $var(X)$.

3. ("**One-armed bandit**") A simple model fruit machine consists of 3 rotating discs, each with 10 symbols: one of the symbols is J (*Jackpot*). After entering a Euro the discs start to rotate and after a while one disc after the other comes to a stand. On each disc one symbol is visible. If no J is visible, one loses the entered Euro, but if one J is visible, then the entered Euro is returned. If 2 J`s are visible the machine pays 10 Euro (leaving a profit of 9 Euro`s). How many Euro`s should the machine pay when 3 $J$`s are visible for a "fair" game? (A game is called "fair" if the expected profit is 0).

4. ("**Chuck-a-luck**") In this popular game at fairs one should bet a certain amount of money on one of the face up numbers of a dice: 1, 2, 3, 4, 5 or 6. Then 3 dice are rolled once and the number of dice with the chosen number face up is counted.
If the number of dice with your number face up is 0, you lose your money, but if it is at least 1, then your bet is returned and you are paid the amount of your bet times the number of dice with the chosen number.
Compute the expected profit when betting 1 Euro.

5. A vase contains N marbles with numbers $1, 2, \ldots, N$. Arbitrarily and without replacement n ($n \leq N$) marbles are drawn from the vase. Define X as the number of the marble with the highest number among the drawn marbles.
a. Determine $P(X = 7)$ for $n = 4$ and $N = 10$.
b. Give the probability function of X for arbitrary n and N ($1 \leq n \leq N$).

6. The random variable X has a binomial distribution. Compute the following probabilities, using the binomial table in this reader:
a. $P(X \leq 7)$, if $n = 10$ and $p = 0.3$
b. $P(X \geq 7)$, if $n = 10$ and $p = 0.3$
c. $P(X = 9)$, if $n = 15$ and $p = 0.6$. Check the value with the (exact) binomial formula.
d. $P(X < 12)$, if $n = 15$ and $p = 0.6$

7. A random variable X has a Poisson distribution with parameter $\mu = 3$.
Compute, using the table of (cumulative) Poisson probabilities:

a. $P(X = 5)$. Check the table result using the exact Poisson probability function.
b. $P(X < 2)$.
c. $P(X > 3)$.

8. Determine for each of the following situations whether the random variable has a homogeneous, geometric, binomial, hypergeometric or Poisson distribution, or none of them. Give for each part: 1. Your choice (including parameters!) with a brief motivation.
2. The probability $P(X = 2)$ and
3. $E(X)$.

a. A software designer takes care of a hotline for clients, who can call to ask questions about the use of software. From passed experience we know that the number of telephone calls is 30 per hour.
X is the number of telephone calls in a 10 minutes period.
b. Assume that a company has two vacancies in the board of directors. There are five equally suitable candidates, of whom two are females.
The two directors are chosen by drawing lots: X= "the number of chosen women".
c. A producer of computer chips draws a random sample of 100 chips out of the total (very large) production of one hour. In the total production 2% of the chips is defective. X is the number of defective chips among the chosen 100 chips.
d. Somebody tries to open a door using a bunch of 10 (different) keys. Only one of the keys can open the door. X is the number of trials if he tries to open the door by choosing an arbitrary key from the bunch ("with replacement").
e. Consider the key problem in d. again, but this time when he removes a key that he tried ("without replacement"). Again X is the number of trials to open the door.

9. X is binomially distributed with $n = 25$ and $p = 0.05$.
In this exercise we are going to check how good a Poisson approximation of a binomial probability is in this case.

a. Compute $P(X = 0)$ using the given binomial distribution..
b. Approximate $P(X = 0)$ with the proper Poisson distribution.

10. The variable X has a simple distribution: $P(X = c) = \frac{1}{2}$ and $P(X = 0) = \frac{1}{2}$.

a. Determine a formula for all moments $E(X^k)$, $k = 1,2,3, \ldots$
b. Use the result of a. to compute the expectation and the variance of X.

11. The variable X has the following probability function: $P(X = i) = c \cdot \left(\frac{1}{3}\right)^i$, $i = 0, 1, 2, \ldots$

- **a.** Show that $c = \frac{2}{3}$ and sketch the probability function.
- **b.** Which (well known) distribution does $Y = X + 1$ have (including parameter(s))?
- **c.** Use the result of b. to determine the variance of X.

12. (former exam exercise)
A company with 150 employees wants to change its telephone policy, by reducing the number of outgoing telephone lines. At the moment everybody has an outgoing line, but the number of used outgoing lines is on average only 3 (out of 150) during office hours.

- **a.** Which distribution could you use to model the number of calls outside the company at an arbitrary moment during office hours? First state the assumptions you made.
- **b.** Determine the smallest number of outgoing telephone lines as to ensure that the probability that the number of outgoing lines is insufficient is less than 5%.

13. (former exam exercise)
Give for each of the three described situations the (most) suitable distribution of X and determine $P(X > EX)$.

- **a.** A colporteur sells an energy contract to, on average, 15% of the visited clients.
  X is the number of sold contracts, when he visits 12 (potential) clients on a day.
- **b.** A hospital has two incubators for just born babies. The average number of requested incubators on an arbitrary day is 2 as well.
  X is the actual demand of incubators on a day.
- **c.** There are 3 Dutchmen in a group of 10 candidate astronauts. 4 out of 10 will be chosen for the next space flight.
  X is the number of Dutchmen of the (arbitrarily) chosen crew of 4.

14. It is known that 4% of all eggs of weight class 2 in supermarkets are outside class 2 weight bounds.

- **a.** Compute the probability that a box of 10 eggs contains at least one egg outside weight class 2. State your assumptions.
- **b.** Compute or approximate (using the same assumptions) the probability that in 10 boxes of 10 eggs at least 4 eggs are outside class 2.
- **c.** On buying a box of 10 eggs we will check whether all eggs are class 2 eggs. We will buy boxes of eggs until we have a box that contains eggs outside class 2.
  What is the expected number of boxes that we have to buy?

15. **a.** In a game we will have to repeat rolling a dice until we rolled 6 three times.
Compute the probability that we roll the third 6 in the tenth roll.

b. (Generalization of a.). We repeat a Bernoulli trial with success probability p until we have m successes. If X is the number of required trials, then determine S_X and the probability function of X.
(This distribution is the **negative binomial distribution**.)

16. M is the **median** of the distribution of X if $P(X \geq M) \geq \frac{1}{2}$ and $P(X \leq M) \geq \frac{1}{2}$.
Determine the median (or medians) if X has the following distribution:

a. the geometric distribution with parameter $p = \frac{1}{3}$,
b. the Poisson distribution with parameter 2,
c. the Poisson distribution with parameter 2.5 and
d. the $B(7, ½)$-distribution.

Some hints for solution of the exercises of chapter 4:

1. First write down the formulas of $E(X)$, $E(X^2)$ and $var(X)$ $(= E(X^2) - (EX)^2)$.
2. Idem
3. Compute the probability of each number of J's and the corresponding profit (payment – bet).
4. Similar as exercise 3.
5. Solve this using a combinatorial approach: compute the total and favorable number of drawing results.
6. Note that X can only attain integer values and that the probabilities $P(X \leq k)$ can be found in the table.
In c. and d.: $p > 0.5$. If the success probability is greater than 0.5, the probability of a failure is less than 0.5; transfer the event to the number of failures.
7. Poisson-tables are similar as the binomial tables: they contain cumulative prob. $P(X \leq c)$.
8. Memorize the types of distributions to choose from, and their (logical) expectations:
- Geometric: count the number of independent trials with success rate p until you succeed
- Binomial: number of successes (p) in n independent trials (draws with replacement)
- Hypergeometric: n draws without replacement, count the number of "successes": $p = \frac{R}{N}$
- Poisson: Number of rare events in an area/period, on average μ occur.
- Homogeneous: equal probabilities for all values of X

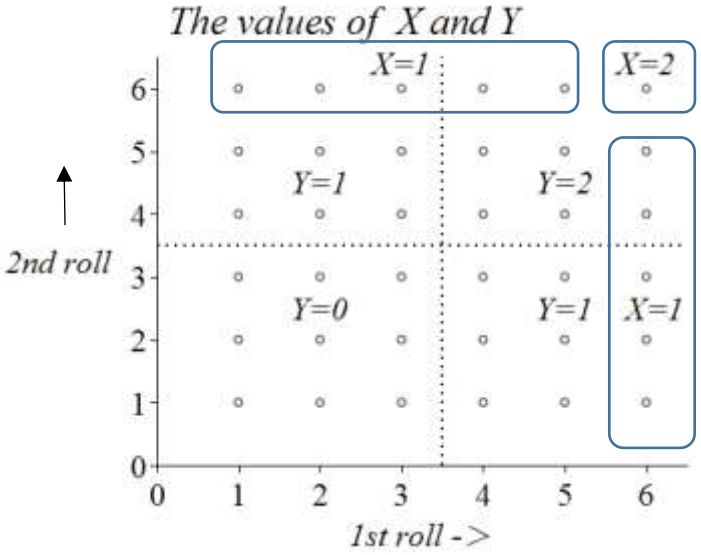
The formula of the Poisson probability function is to be found on the formula page.
9. -
10. Formula of $E(X^k)$ can easily be remembered as the weighted average of the values of X^k.
11. a. Recognize the summation as a geometric series, see appendix "Mathematical techniques"
b. Formula for $E(aX + b)$ and $var(aX + b)$ should be given without hesitation….
12. See 8.
13. See 8.
14. See 8.
15. The formula can be derived similarly as the geometric and binomial formula: take into account that the last trial always should be success: the 3^{rd} success occurs in the last trial.
16. For the Poisson- and binomial distribution, one can use the tables in this reader.

Chapter 5: Two or more discrete variables

5.1 Joint probability functions

In chapter 4 we discussed merely the distribution of one random variable, but in many situations more than one "quantitative aspects" play a role: several random variables can be defined. If two variables X and Y are defined for the same probability space we could be interested in the simultaneous occurrence of events $\{X \in B\}$ and $\{Y \in C\}$, where $B \subset \mathbb{R}$ and $C \subset \mathbb{R}$. Moreover, we are interested in the relation of the numbers$X$ and Y, or: are the events $\{X \in B\}$ and $\{Y \in C\}$ independent? Concepts and definitions are initially given for two random variables, but in most cases they are easily extended to more than two variables.

Example 5.1.1 We toss an unbiased dice twice. X and Y are defined as "the number of sixes (in two tosses)" and "the number of two tosses with a result larger than 3", resp.
Both X and Y can attain the values 0, 1 and 2: $S_X = S_Y = \{0, 1, 2\}$.
Since the 2 tosses result in 36 equally likely outcomes, every outcome of the experiment can be linked to a value of both X and Y (not necessarily the same), as shown in the diagram:



i \ j	0	1	2
0	$\frac{9}{36}$	$\frac{12}{36}$	$\frac{4}{36}$
1	0	$\frac{6}{36}$	$\frac{4}{36}$
2	0	0	$\frac{1}{36}$

Table of joint probabilities $P(X = i \text{ and } Y = j)$

$\{X = 1 \text{ and } Y = 2\}$ is the event that both toss results are larger than 3 and one of them is a 6, so the events occurs if the one of the outcomes (4,6), (5,6), (6,4) or (6,5) occurs.
So $P(X = 1 \text{ and } Y = 2) = \frac{4}{36}$.
Similarly, we can determine for all values of i and j the so called **joint probabilities** $\boldsymbol{P(X = i \text{ and } Y = j)}$**.** For $i \in S_X$ and $j \in S_Y$ we found the table above.
It is easily seen that the probabilities add up to 1: $\sum_{i=0}^{2} \sum_{j=0}^{2} P(X = i \text{ and } Y = j) = 1$.
The events $\{X = i \text{ and } Y = j\}$ are a partition of the full sample space of 36 sample points.
Using the table of joint probabilities, we can compute the probability of one 6 (so $X = 1$), by splitting up the event $\{X = 1\}$ into the events (a partition of $\{X = 1\}$):
$\{X = 1 \text{ and } Y = 0\}$, $\{X = 1 \text{ and } Y = 1\}$ and $\{X = 1 \text{ and } Y = 2\}$.

So $P(X=1) = P(X=1 \text{ and } Y=0) + P(X=1 \text{ and } Y=1) + P(X=1 \text{ and } Y=2)$
$= 0 + \frac{6}{36} + \frac{4}{36} = \frac{10}{36}$

Similarly: $P(X=0) = \sum_{j=0}^{2} P(X=0 \text{ and } Y=j) = \frac{25}{36}$.

And: $P(X=2) = \sum_{j=0}^{2} P(X=2 \text{ and } Y=j) = \frac{1}{36}$.

Hereby we found the distribution of X, by adding the the probabilities $P(X=i \text{ and } Y=j)$ in each row. Of course, we could have found the distribution directly by considering the two rolls of the dice as Bernoulli trials with probability of a 6 equal to $\frac{1}{6}$: $X \sim B\left(2, \frac{1}{6}\right)$.

The distribution of Y can be computed by adding the probabilities of each column.

i \ j	0	1	2	$P(X=i)$
0	$\frac{9}{36}$	$\frac{12}{36}$	$\frac{4}{36}$	$\frac{25}{36}$
1	0	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{10}{36}$
2	0	0	$\frac{1}{36}$	$\frac{1}{36}$
$P(Y=j)$	$\frac{9}{36}$	$\frac{18}{36}$	$\frac{9}{36}$	1

Anticipating on the definition of **independence of the variables X and Y** in section 5.3 we notice that the events $\{X=i\}$ and $\{Y=j\}$ in general are not independent:

Here e.g.: $\frac{9}{36} = P(X=0 \text{ and } Y=0) \neq P(X=0) \cdot P(Y=0) = \frac{25}{36} \cdot \frac{9}{36}$

Intuitively, the dependence is clear: if both rolls of the dice result in values 3 or smaller ($Y=0$), no sixes can be rolled ($X=0$).

We can compute the probability of the event that the number of sixes is smaller than the number of rolls larger than 3:

$$P(X<Y) = P(X=0 \text{ and } Y=1) + P(X=0 \text{ and } Y=2) + P(X=1 \text{ and } Y=2)$$
$$= \frac{12}{36} + \frac{4}{36} + \frac{4}{36} = \frac{20}{36}$$

■

Definition 5.1.2 If a pair (X,Y) of discrete random variables defined on the same probability space and range $S_X \times S_Y$, then for $(x,y) \in S_X \times S_Y$ $\boldsymbol{P(X=x \text{ and } Y=y)}$ is the **joint probability function** of X and Y.

Property 5.1.3 For each joint probability function $P(X=x \text{ and } Y=y)$ of X and Y we have:

1) $\boldsymbol{P(X=x \text{ and } Y=y) \geq 0}$.

2) $\boldsymbol{\sum_{x \in S_X} \sum_{y \in S_Y} P(X=x \text{ and } Y=y) = 1}$.

Conversely, if a two dimensional function satisfies these two conditions, it is a joint probability function. The probability function of (only) X is a **marginal probability function**, $P(Y=y)$ is the marginal probability function of Y. The computation of the marginal probability function of X and Y (example 5.1.1) can be given as follows:

Property 5.1.4

$$\boldsymbol{P(X=x)=\sum_{y\in S_Y} P(X=x \text{ and } Y=y)} \quad \text{and} \quad \boldsymbol{P(Y=y)=\sum_{x\in S_X} P(X=x \text{ and } Y=y)}$$

In example 5.1.1 these formulas are equivalent to addition of the rows (X) and addition of the columns (Y) of the table.
The event $\{X<Y\}$ in this example consisted of 3 pairs (x,y) of values that X and Y can attain. Generalizing this approach: if B is a subset of the xy-plane, so $B\subset\mathbb{R}^2$, then the probability that the pair (X,Y) attains values from B, can be computed using the joint probability function:

$$P((X,Y)\in B)=\sum_{(x,y)\in B} P(X=x \text{ and } Y=y)$$

Example 5.1.5 The grand masters (in chess) Timman and Karpov play a match of 6 chess games. Chess statisticians computed that in the past Timman won a game against Karpov with probability $p_1=\frac{2}{10}$. His probabilities of loss and a draw are $p_2=\frac{3}{10}$ and $p_3=\frac{5}{10}$, no matter who started with white or black.
How large is the probability that Timman wins the match of six games? This will be the case if the number of wins by Timman (X) is larger than the number of his losses (Y: the number of wins by Karpov). Using the random variables X and Y we will have to compute the probability:

$$P(X>Y)=\sum_{i>j} P(X=i \text{ and } Y=j)$$

The event $X>Y$ can occur, e.g., if Timman wins 3 times ($X=3$) and looses one time ($Y=1$): the number of draws is two in that case. The probability of $\{X=3 \text{ and } Y=1\}$ can be computed: if *w* denotes "a win" by Timman, *l* "loss" and *d* "draw", then (*l, d, d, w, w, w*) is one of the outcomes in this event, with probability:

$$P\big((l,d,d,w,w,w)\big)=p_2\cdot p_3\cdot p_3\cdot p_1\cdot p_1\cdot p_1=p_2{p_3}^2p_1^3,$$

applying independence of the games.
The number of orders of 3 wins, 1 loss and 2 draws is $\binom{6}{3}\binom{3}{1}\binom{2}{2}=\frac{6!}{3!1!2!}$
Since every of these outcomes has the same probability, we have:

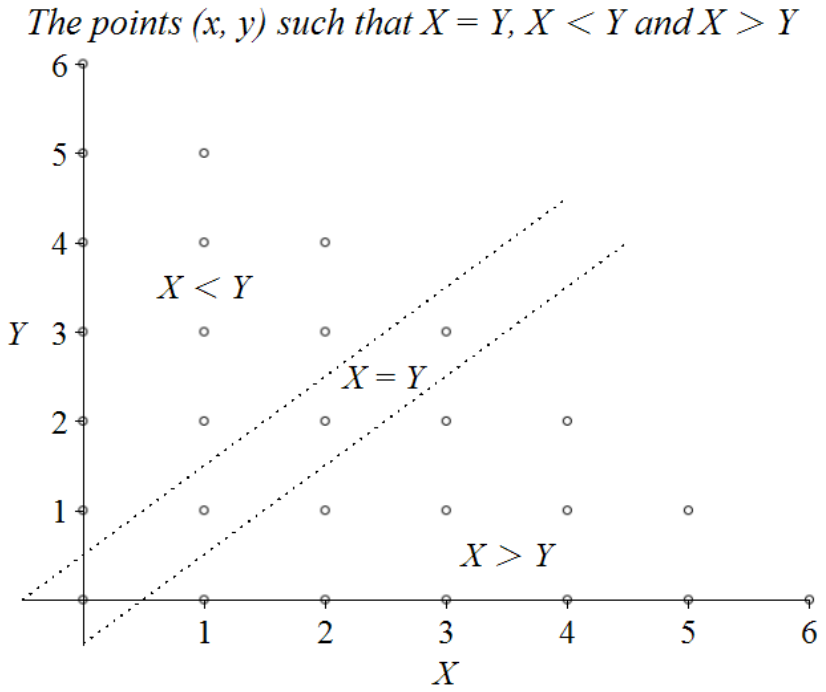
$$P(X=3 \text{ and } Y=1)=\frac{6!}{3!\,1!\,2!}\left(\frac{2}{10}\right)^3\left(\frac{3}{10}\right)\left(\frac{5}{10}\right)^2\approx 3.6\%.$$

In general we can express the joint probability function $P(X=i \text{ and } Y=j)$ of X and Y in a match of n games in the Timman`s win and loss probabilities p_1 and p_2, because $p_3=1-p_1-p_2$ and the number of draws is $n-i-j$.

$$P(X=i \text{ and } Y=j)=\frac{n!}{i!\,j!\,(n-i-j)!}p_1^i p_2^j(1-p_1-p_2)^{n-i-j}$$

This is a valid formula for $i\geq 0$ and $j\geq 0$ such that $i+j\leq n$.
For all remaining pairs of values (i,j)we have $P(X=i \text{ and } Y=j)=0$.

This joint probability function $P(X = i \text{ and } Y = j)$ is called a **multinomial (here: trinomial) probability function.** Comparing it to the binomial probability function we can see that in this case we have independent experiments (games) with **3** instead of **2** different outcomes.
In the diagram below we illustrated the range of only the pairs (i, j) with positive probability $P(X = i \text{ and } Y = j)$:



Using the diagram, the requested probability can be computed as follows:

$$P(X > Y) = \sum_{i>j} P(X = i \text{ and } Y = j)$$

$$= \sum_{i=1}^{6} P(X = i \text{ and } Y = 0) + \sum_{i=2}^{5} P(X = i \text{ and } Y = 1) + \sum_{i=3}^{4} P(X = i \text{ and } Y = 2)$$

$$= 0.25956 \approx 26\%.$$

(Of course the probabilities $P(X = Y)$ and $P(X < Y)$ can be computed in a similar manner).
The marginal probability functions of X and of Y could be computed by applying property 5.1.4, but it is easier to reason that X has a $B(n, p_1)$-distribution since the $n = 6$ independent games all can be won by Timman with probability p_1.
Similarly, the number of losses, Y, has a $B(n, p_2)$-distribution.
And the number of draws, $Z = n - X - Y$, has a $B(n, 1 - p_1 - p_2)$-distribution.
Since X and Y are clearly dependent (if $X = 6$, then $Y = 0$), we **cannot** use the equality
$P(X = i \text{ and } Y = j) = P(X = i) \cdot P(Y = j)$ ■

The joint probability function $P(X = i \text{ and } Y = j)$ in the previous example is sometimes given as $P(X = i \text{ and } Y = j \text{ and } Z = k)$: in that case we have an extra restriction:
$i + j + k = n$.
Definition 5.1.2 and properties 5.1.3 and 5.1.4 can easily be extended to joint probability functions $P(X_1 = x_1 \text{ and } X_2 = x_2 \text{ and } \ldots \text{and } X_n = x_n)$ of n (> 2) discrete random variables.

From these joint probability functions we can similarly derive the marginal distribution of, e.g., for X_1:

$$P(X_1 = x_1) = \sum_{x_2 \in S_{X_2}} \dots \sum_{x_n \in S_{X_n}} P(X_1 = x_1 \text{ and } X_2 = x_2 \text{ and} \dots \text{and } X_n = x_n)$$

Example 5.1.6 A company produces mega chips in batches of 20 pieces. Because of a design error each batch of 20 chips contains 7 defective chips. Before the design error was discovered, 4 of the 20 chips of the first batch were sold to customers (numbered 1 to 4). We can denote a 1 for a non-defective chip and 0 for a defective one. We will use a variable X_i for every customer i:

$$X_i = \begin{cases} 1, & \text{if the chip of customer } i \text{ works} \\ 0, & \text{if the chip is defective} \end{cases} \qquad (\text{for } i = 1, 2, 3, 4)$$

Given are the probabilities: $P(X_i = 1) = \frac{13}{20}$ and $P(X_i = 0) = \frac{7}{20}$, irrespectively the order of delivery to clients. (See for a motivation exercise 2.3 and its solution)
Every X_i has a $B\left(1, \frac{13}{20}\right)$ –distribution, or: an **alternative distribution with success probability $\boldsymbol{p = \frac{13}{20}}$**.
We cannot use the marginal distributions of the X_i's to compute the joint distribution: e.g. the probability that all 4 delivered chips work is **not** $\left(\frac{13}{20}\right)^4$.
However, if we apply the **product rule of dependent events** (property 3.1.5) to the events $\{X_1 = 1\}, \{X_2 = 1\}, \{X_3 = 1\}$ and $\{X_4 = 1\}$ we find:

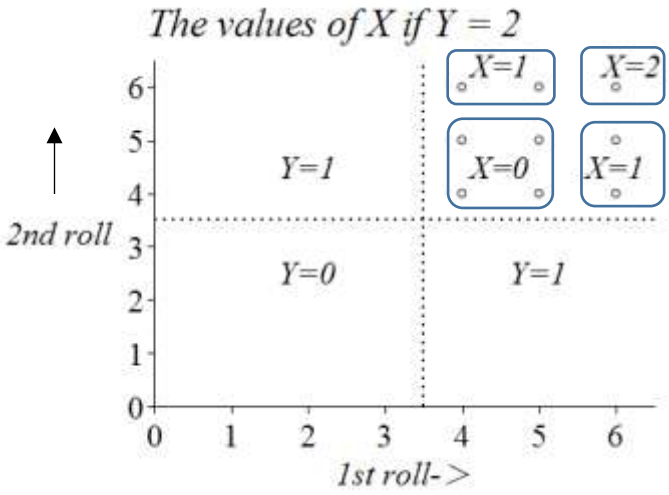
$$\begin{aligned} &P(X_1 = 1 \text{ and } X_2 = 1 \text{ and } X_3 = 1 \text{ and } X_4 = 1) \\ &\quad = P(X_1 = 1) \cdot P(X_2 = 1 | X_1 = 1) \cdot P(X_3 = 1 | X_1 = X_2 = 1) \cdot P(X_4 = 1 | X_1 = X_2 = X_3 = 1) \\ &\quad = \frac{13}{20} \cdot \frac{12}{19} \cdot \frac{11}{18} \cdot \frac{10}{17} \end{aligned}$$

Similarly, we can find the joint probability function $P(X_1 = x_1 \text{ and} \dots \text{and } X_4 = x_4)$ for all other values x_1 to x_4. ■

5.2 Conditional distributions

In the first section of this chapter we were interested in the joint probabilities, but sometimes we have information about one of the variables and we want to know how this information affects the distribution of the other variable. E.g., if in example 5.1.6 the first customer complains that his chip does not work ($X_1 = 0$), then this information affects the distribution of the variables X_2, X_3 and X_4.

Example 5.2.1 Returning to example 5.1.1 (roll a dice twice) we know that both rolls resulted in outcomes larger than 3 ($Y = 2$):
how does this information affect the probability to roll 0, 1 or 2 sixes (the distribution of X)?
If $Y = 2$, then only 9 (of 36) outcomes could have occurred (as illustrated in the accompanying diagram).
As before it seems reasonable to assume that these 9 sample points have the same probability. Then the probability of 2 sixes is 1 out of 9: $\frac{1}{9}$.



This conditional probability of the event $\{X = 2\}$, given the event $\{Y = 2\}$, is denoted as $P(X = 2|Y = 2)$.
Using the definition of conditional probability we can verify the correctness of the probability:

$$P(X = 2|Y = 2) = \frac{P(X = 2 \text{ and } Y = 2)}{P(Y = 2)} = \frac{1/36}{9/36} = \frac{1}{9}$$

Likewise:

$$P(X = 1|Y = 2) = \frac{4}{9}$$
$$P(X = 0|Y = 2) = \frac{4}{9}$$

These 3 probabilities is a probability distribution: X attains values 0, 1 and 2 with the given probabilities, adding up to 1. But all under condition that both rolls are larger than 3.
The 3 probabilities give the **conditional probability of X, given $Y = 2$**.
As before we can compute the expected value of X, given $Y = 2$, as the weighted average of the values of X:

$$0 \cdot \frac{4}{9} + 1 \cdot \frac{4}{9} + 2 \cdot \frac{1}{9} = \frac{2}{3}$$

This expectation is **not** $E(X)$, the unconditional expectation of X, since we used the condition $Y = 2$ to compute the probabilities.
Therefore we will use the notation: $E(X|Y = 2) = \frac{2}{3}$.
Similarly, we can use the joint and marginal probability functions of X and Y to determine the distribution and expected value of X under the condition $Y = 1$:

$$P(X = 0|Y = 1) = \frac{P(X = 0 \text{ and } Y = 1)}{P(Y = 1)} = \frac{2}{3}$$
$$P(X = 1|Y = 1) = \frac{P(X = 1 \text{ and } Y = 1)}{P(Y = 1)} = \frac{1}{3}$$
$$P(X = 2|Y = 1) = 0.$$

So, $E(X|Y = 1) = \sum_{i=0}^{2} i \cdot P(X = i|Y = 1) = 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} + 2 \cdot 0 = \frac{1}{3}$.
We found that, in this case ($Y = 1$), the expected number of sixes is smaller than if $Y = 2$.
The last given value of Y could be $Y = 0$, then:

$$P(X = 0|Y = 0) = \frac{P(X = 0 \text{ and } Y = 0)}{P(Y = 0)} = \frac{9/36}{9/36} = 1,$$

meaning: if $Y = 0$, X can only attain the value 0, so $E(X|Y = 0) = 0$. ■

The conditional distributions for two discrete random variables X and Y within the same probability space can in general be defined as follows:

Definition 5.2.2 If X and Y are discrete random variables, then the **conditional probability function of X, given $Y = y$**, is defined by:

$$\boldsymbol{P(X = x|Y = y) = \frac{P(X = x \text{ and } Y = y)}{P(Y = y)}}, \qquad \text{for } x \in S_X.$$

This conditional distribution is only defined for values y in S_Y, such that $P(Y = y) > 0$.

For fixed $y \in S_Y$ and variable $x \in S_X$, $P(X = x|Y = y)$ is really a probability function since:

1) $P(X = x|Y = y) = \frac{P(X=x \text{ and } Y=y)}{P(Y=y)} \geq 0$, for $x \in S_X$ and
2) $\sum_{x \in S_X} P(X = x\,|Y = y) = \sum_{x \in S_X} \frac{P(X=x \text{ and } Y=y)}{P(Y=y)} = \frac{P(Y=y)}{P(Y=y)} = 1$

For any probability function the expected value is defined as the "weighted average of the values" of the variable. This remains the case for conditional probability functions, but we will use the notation $E(X|Y = y)$ to make it clear that we are using the distribution under the condition $Y = y$.

Property 5.2.3 The **conditional expectation X, given $Y = y$,** is

$$\boldsymbol{E(X|Y = y) = \sum_{x \in S_X} x \cdot P(X = x|Y = y)}.$$

As is the case for any expectation the usual properties of expectation still apply, such as

$$E(aX + b|Y = y) = aE(X|Y = y) + b.$$

Example 5.2.4 Let us return to the examples 5.1.1 and 5.2.1 where we rolled a dice twice. We noticed that knowledge about the number of rolls larger than 3 (Y) affects the probability of specific numbers of sixes (X) in the same two rolls:

$$E(X) = \frac{1}{3} \quad \text{versus} \quad \begin{cases} E(X|Y = 0) = 0 \\ E(X|Y = 1) = \frac{1}{3} \\ E(X|Y = 2) = \frac{2}{3} \end{cases}$$

What is the relation between the (unconditional) expectation $E(X)$ and these three conditional expectations of X?

Well: if $Y = 0$, then $E(X|Y = 0) = 0$, and this occurs with probability $P(Y = 0) = \frac{1}{4}$.

Similarly $E(X|Y = 1) = \frac{1}{3}$ occurs with probability $P(Y = 1) = \frac{1}{2}$ and $E(X|Y = 2) = \frac{2}{3}$ with

probability $P(Y=2)=\frac{1}{4}$.
The weighted average of these conditional expectations can be computed:

$$E(X|Y=0)\cdot P(Y=0)+E(X|Y=1)\cdot P(Y=1)+E(X|Y=2)\cdot P(Y=2)$$
$$=0\cdot\frac{1}{4}+\frac{1}{3}\cdot\frac{1}{2}+\frac{2}{3}\cdot\frac{1}{4}=\frac{1}{3}=E(X)$$

Apparently we can conceive $E(X)$ as the weighted average of these conditional expectations $E(X|Y=y)$ with weighing factors the probabilities of the conditions, $P(Y=y)$. ■

Example 5.2.4. shows that the numerical values $E(X|Y=y)$ form the range of a random variable (e.g. let us call it V), such that:

$$P[V=E(X|Y=y)]=P(Y=y)$$

in this example: $P\left(V=\frac{2}{3}\right)=\frac{1}{4}$, $P\left(V=\frac{1}{3}\right)=\frac{1}{2}$ and $P(V=0)=\frac{1}{4}$.
The common notation of the "random conditional expectation" V is $E(X|Y)$.

> **Definition 5.2.5** If X and Y are discrete random variables, then the **conditional expectation $\boldsymbol{E(X|Y)}$** is a **random variable**, that attains values met $\boldsymbol{E(X|Y=y)}$ with probability $\boldsymbol{P(Y=y)}$.

$E(X|Y)$ is referred to as **"the conditional expectation of $\boldsymbol{X}$, given $\boldsymbol{Y}$."**
Note that $E(X|Y)$ is a random variable, where as $E(X|Y=y)$ is a numerical values for specific given value of Y.

Using the defined distribution of $E(X|Y)$ we can apply the definition of expectation:

$$E[\,E(X|Y)\,]=\sum_y E(X|Y=y)\cdot P(Y=y)$$

In example 5.2.5 we saw that the right hand side of the equation results in $E(X)$.
This is a general property, which can be proven using the definitions of both $E(X|Y=y)$ and $P(X=x|Y=y)$:

$$E[E(X|Y)]=\textstyle\sum_y E(X|Y=y)\cdot P(Y=y)$$
$$=\sum_y\left(\sum_x x\cdot P(X=x|Y=y)\right)\cdot P(Y=y)$$
$$=\sum_x\sum_y x\cdot\frac{P(X=x\text{ and }Y=y)}{P(Y=y)}\cdot P(Y=y)$$
$$=\sum_x x\cdot\sum_y P(X=x\text{ and }Y=y)=\sum_x xP(X=x)=E(X)$$

We have proven:

Property 5.2.6 $\quad\boldsymbol{E[\,E(X|Y)\,]=E(X)}$

This property can be applied to compute the expectation of a variable, for which the distribution is unknown, but its conditional distribution is.

Example 5.2.7 Let us assume that a proper probability model of the number of bicycle thefts in a particular town during a day is the Poisson distribution with a mean number of 8 thefts. Furthermore we know that only half of all bicycle thefts are reported to the police.
Modelling this situation we define N as the Poisson distributed number of thefts on an arbitrary day and X as the number of reported thefts on that day, so necessarily $X \leq N$.
If we assume, e.g., that the actual number of thefts on a given day is $N = 10$, then each of these thefts will, or will not, be reported: we observe 10 independent trials with "report probability" $\frac{1}{2}$.

The reported number X has, given N = 10, a $B\left(10, \frac{1}{2}\right)$-distribution.

Or, in general, X has, given $N = n$, $B\left(n, \frac{1}{2}\right)$-distribution and, consequently,

$$E(X|N = n) = \frac{1}{2}n.$$

The random variable $E(X|N)$ takes on the values $E(X|N = n) = \frac{1}{2}n$ with probability $P(N = n)$, so $E(X|N)$ is a function of N: $E(X|N) = \frac{1}{2}N$.
But, according to property 5.2.6:

$$E(X) = E[E(X|N)] = E\left[\tfrac{1}{2}N\right] = \tfrac{1}{2}E(N) = \tfrac{1}{2} \cdot 8 = 4,$$

So, the expected number of reported bicycle thefts is half of the number of the expected number of thefts.
On a day with 6 reported bicycle thefts the police is wondering how many bicycle thefts really occurred: how many of them should you expect? Or, in terms of the defined variables, what is $E(N|X = 6)$?
If there are $X = 6$ reported bicycle thefts, then $N = 6, 7, 8, \ldots$ thefts occurred, where:

$$P(N = n|X = 6) = \frac{P(X = 6 \text{ and } N = n)}{P(X = 6)} = \frac{P(X = 6|N = n) \cdot P(N = n)}{P(X = 6)}$$

In this formula:

- X has, given $N = n$, $B\left(n, \frac{1}{2}\right)$-distribution,
- N is Poisson distributed with "mean" $\mu = 8$ and
- $P(X = 6)$ can be computed with the **law of total probability**:

$$\begin{aligned} P(X = 6) &= \sum_{n=6}^{\infty} P(X = 6 \text{ and } N = n) = \sum_{n=6}^{\infty} P(X = 6|N = n) \cdot P(N = n) \\ &= \sum_{n=6}^{\infty} \binom{n}{6} \left(\tfrac{1}{2}\right)^n \cdot \frac{8^n e^{-8}}{n!} \\ &= \frac{e^{-4}}{6!} \sum_{n=6}^{\infty} \frac{\left(\frac{1}{2} \cdot 8\right)^n e^{-4}}{(n-6)!} \\ &\overset{n-6=k}{=} \frac{e^{-4}}{6!} \cdot 4^6 \sum_{k=0}^{\infty} \frac{4^k e^{-4}}{k!} = \frac{4^6 e^{-4}}{6!} \cdot 1 \\ &= \frac{4^6 e^{-4}}{6!} \end{aligned}$$

Conducting this computation for $X = k$ instead of $X = 6$ we find: $P(X = k) = \frac{4^k e^{-4}}{k!}$, for $k = 0, 1, 2, \ldots$.
Conclusion from the probability function we found: $\boldsymbol{X}$ has a **Poisson** distribution with mean $\boldsymbol{\mu = E(X) = 4}$.
We are now ready to apply the formula for $P(N = n|X = 6)$ we found before (actually applying Bayes' rule):

$$P(N = n|X = 6) = \frac{P(X = 6|N = n) \cdot P(N = n)}{P(X = 6)} = \frac{\binom{n}{6}\left(\frac{1}{2}\right)^n \cdot \frac{8^n e^{-8}}{n!}}{\frac{4^6 e^{-4}}{6!}} = \frac{4^{n-6} e^{-4}}{(n-6)!},$$

where $n = 6, 7, 8, \ldots$
We recognize a "shifted Poisson distribution" in this expression. Shifting it 6 units back means replacing $n - 6$ by k:

$$P(N - 6 = k|X = 6) = \frac{4^k\, e^{-4}}{k!}, \qquad \text{for } k = 0, 1, 2, \ldots,$$

Which shows that $N - 6$ has, given $X = 6$, a Poisson distribution with parameter $\mu = 4$.
So $E(N - 6|X = 6) = 4$ and thus $E(N\,|X = 6) = 6 + 4 = 10$.
Though the expected number of reported thefts is half of the actual number of bicycle thefts ($E(X|N = n) = \frac{1}{2}n$), the expected number of bicycle thefts is not twice the reported number, since $E(N|X = x) = x + 4$.
Or: $E(N|X) = X + 4$, and (prop. 6.2.6) $E(N) = E[E(N|X)] = E(X + 4) = E(X) + 4 = 8$ ■

Note 5.2.8: The notion of conditional variance falls outside this course, but can be applied in a similar way as the conditional expectation: in the previous example it is natural to state that $var(X|N = n) = np(1 - p) = \frac{1}{4}n$, since X has, given $N = n$, a $B\left(n, \frac{1}{2}\right)$ distribution.
But where we could conclude from $E(X|N) = \frac{1}{2}N$, that $E(X) = E\left(\frac{1}{2}N\right) = \frac{1}{2}E(N)$, we cannot apply the same approach for the variance:

$$var(X|N) = \frac{1}{4}N, \text{ but } \quad var(X) \neq E[var(X|N)] = E\left[\frac{1}{4}N\right] = \frac{1}{4}E(N) = \frac{1}{4} \cdot 8 = 2,$$

since we found in the example that X has a Poisson distribution with $\mu = 4 = var(X)$. ■

5.3 Independent random variables

We know that two events A and B are, by definition, independent if $P(AB) = P(A) \cdot P(B)$.
In example 5.1.1 we noticed that such an equality is not always the case for two events $\{X = i\}$ and $\{Y = j\}$, where X and Y are discrete random variables.
In example 5.1.6 we saw that for 4 random variables the events $\{X_1 = x_1\}$ to $\{X_4 = x_4\}$ can be dependent as well. If these events are dependent, then the **random variables** are **dependent** as well.

Definition 5.3.1 Two discrete random variables $\boldsymbol{X}$ **and** $\boldsymbol{Y}$ are **independent** if

$$\boldsymbol{P(X = x \text{ and } Y = y) = P(X = x) \cdot P(Y = y)} \text{ for } \textbf{all pairs } \boldsymbol{(x, y)} \in S_X \times S_Y$$

The equality in the definition is referred to as the **product rule for independent variables**. This definition is easily extended to n discrete random variables $X_1, \ldots, X_n$:

$P(X_1 = x_1 \text{ and} \ldots \text{and } X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$, for all n-tuples $(x_1, \ldots, x_n)$.

We note that, if X and Y are independent, then events such as $\{X > a\}$ and $\{Y > b\}$ are independent as well:

$$P(X > a \text{ and } Y > b) = \sum_{x>a} \sum_{y>b} P(X = x \text{ and } Y = y) \overset{\text{ind.}}{=} \sum_{x>a} \sum_{y>b} P(X = x) \cdot P(Y = y)$$
$$= \sum_{x>a} P(X = x) \cdot \sum_{y>b} P(Y = y) = P(X > a) \cdot P(Y > b)$$

Moreover, in case of independence the conditional distribution of X, given $Y = y$, is the same as the (unconditional) distribution of X:

$$P(X = x|Y = y) = \frac{P(X = x \text{ and } Y = y)}{P(Y = y)} \overset{\text{ind.}}{=} \frac{P(X = x) \cdot P(Y = y)}{P(Y = y)} = P(X = x)$$

(*Interpretation: knowing that $Y = y$ does not affect the probability of $X = x$.*)

From the definition of independence we can conclude that the joint probability function is completely defined if the marginal probability functions of the independent X and Y are known.
For dependent random variables this is not the case: the joint distribution determines the marginal distributions, but not reversely.

Example 5.3.2 (Checking independence)
The joint distribution of X and Y is given in the table and by adding the rows and columns we determined the marginal distributions of X and Y.
On the diagonal of the table, for $i = j = -1, 0, 1$:
$P(X = i \text{ and } Y = i) = \frac{1}{9} = \frac{1}{3} \cdot \frac{1}{3}$
$= P(X = i) \cdot P(Y = i)$

i \ j	-1	0	1	$P(X = i)$
-1	$\frac{1}{9}$	$\frac{2}{9}$	0	$\frac{1}{3}$
0	0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{3}$
1	$\frac{2}{9}$	0	$\frac{1}{9}$	$\frac{1}{3}$
$P(Y = j)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

But if $i \neq j$, this is not true,
e.g. if $i = -1$ and $j = 0$ we find:
$P(X = -1 \text{ and } Y = 0) = \frac{2}{9}$
$\neq \frac{1}{3} \cdot \frac{1}{3} = P(X = -1) \cdot P(Y = 0)$
(If the joint probability is 0, as is the case for 3 pairs (i, j), the inequality is evident)
Conclusion: X and Y are dependent.
Besides, it can be checked that $E(X|Y = j) \neq E(X)$ for $j = -1, 0, 1$. ■

Usually we do not have to "prove" independence of two variables using the definition, but **independence is assumed** since the random variables are related to supposedly independent experiments. Joint probabilities can be determined using the product rule of independent variables, as is shown in the following example.

Example 5.3.3 X and Y are the "Number of acute appendicitis operations" and the "Number of kidney stone operations" on a day in a hospital. On average the hospital has 3 and 4 of these operations per day, respectively. To avoid capacity problems the management wants to know how large the probability is that for each kind of the operations at least 5 must be conducted during an arbitrary day. To answer this question we will first have to formulate a **probability model** or: "state reasonable model assumptions".

1. Both X and Y have Poisson distributions with expectations 3 and 4.
2. X and Y are independent.

The requested probability can now be computed using the Poisson table:

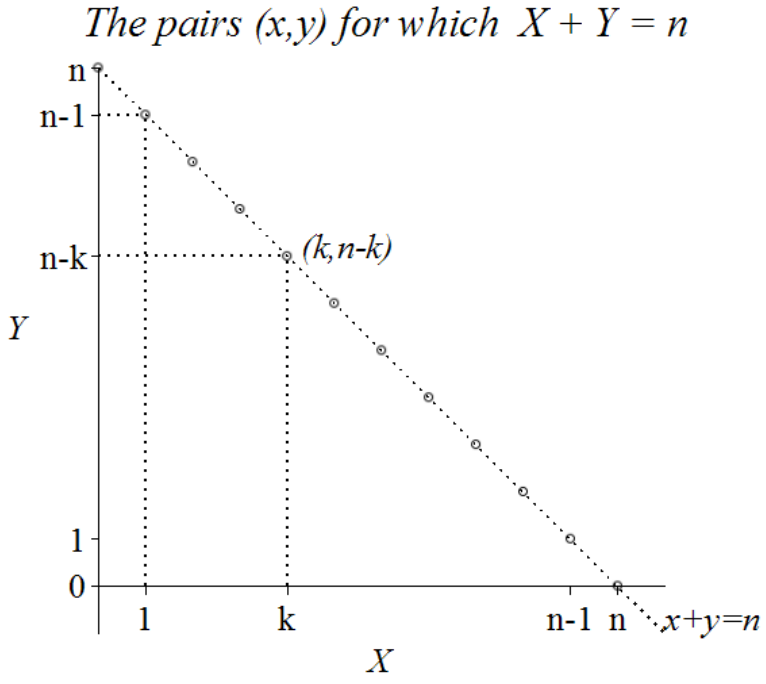
$$\begin{aligned} P(X \geq 5 \text{ and } Y \geq 5) &\overset{\text{ind.}}{=} P(X \geq 5) \cdot P(Y \geq 5) && \text{(independence)} \\ &= \big(1 - P(X \leq 4)\big) \cdot (1 - P(Y \leq 4)) && \text{(complement rule)} \\ &= (1 - 0.815) \cdot (1 - 0.629) && \text{(using the Poisson tables.)} \\ &\approx 6.9\% \end{aligned}$$

■

5.4 Functions of discrete random variables

In the last example we were interested what the probability is that both X and Y are at least 5. This probability can also be denoted as $P(min(X,Y) \geq 5)$.
If we determine the probabilities $P(min(X,Y) = k)$ for all possible values of k, we are actually determining the distribution of a new random variable $W = min(X,Y)$, which can be seen as a function $g(X,Y)$ of the two variables X and Y.
W is the random variable that realizes the numerical value $min(x,y)$, if X and Y take on the values x and y.
Another function $g(X,Y)$ of X and Y, which can be of interest in example 5.3.3, is the total number of operations (of both types): $g(X,Y) = X + Y$.

Example 5.4.1 X and Y are the variables in example 5.3.3: independent and both Poisson distributed with expectations 3 and 4. The total number of operations on a day is $Z = X + Y$. Since $S_X = S_Y = \{0, 1, 2, \dots\}$ is the range of both, the range of Z is $S_Z = \{0, 1, 2, \dots\}$ as well.



The probability function $P(Z = n)$ of Z can be determined: consider the event $\{X + Y = k\}$ as a partition into sub-events $\{X = 0 \text{ and } Y = n\}, \{X = 1 \text{ and } Y = n - 1\}, \ldots,$ $\{X = n \text{ and } Y = 0\}$, the grid points on the line $x + y = n$ as shown in the graph above.
From this it follows:

$$P(Z = n) = \sum_{k=0}^{n} P(X = k \text{ and } Y = n - k)$$

$$\overset{ind.}{=} \sum_{k=0}^{n} P(X = k) \cdot P(Y = n - k) \text{ (using the assumed independence of } X \text{ and } Y).$$

$$= \sum_{k=0}^{n} \frac{3^k\, e^{-3}}{k!} \cdot \frac{4^{n-k}\, e^{-4}}{(n-k)!} = e^{-7} \sum_{k=0}^{n} \frac{1}{k!\,(n-k)!} 3^k 4^{n-k} = \frac{e^{-7}}{n!} \sum_{k=0}^{n} \binom{n}{k} 3^k 4^{n-k}$$

In the last sum we recognize Newton`s Binomial Theorem, $\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a + b)^n$:

$$P(Z = n) = \frac{e^{-7}}{n!} \cdot (3 + 4)^n = \frac{7^n e^{-7}}{n!}, \quad \text{for } n = 0, 1, 2, \ldots.$$

Apparently Z has a Poisson distribution as well, and the expectation is $7 = E(Z) = E(X) + E(Y) = 3 + 4$. Using this distribution we can compute all relevant probabilities with respect to the total number of operations and e.g. we know that $var(Z) = 7$. ■

Generalizing this approach for two X and Y, having ranges S_X and S_Y, we can derive the distribution of $X + Y$ from the joint probability function of X and Y.

$$P(X + Y = n) = \sum_{k \in S_X \text{ with } n-k \in S_Y} P(X = k \text{ and } Y = n - k)$$

Moreover, if X and Y are independent, then we find:

Property 5.4.2 (Convolution sum)

If X and Y are independent discrete random variables, with integer numbered ranges, then:

$$\boldsymbol{P(X + Y = n) = \sum_{k \in S_X} P(X = k) \cdot P(Y = n - k)}$$

The summation on the right hand side is referred to as the "convolution sum" and the addition $X + Y$ as the convolution of X and Y.
The graph in example 5.4.1 shows that the summation is conducted on all pairs $(k, n - k)$ of values of (X, Y) on the line $x + y = n$, provided that $k \in S_X$ and $n - k \in S_Y$.
For other functions $g(X, Y)$ of X and Y, such as $X \cdot Y$, $min(X, Y)$ or $max(X, Y)$ we can determine the distribution of $Z = g(X, Y)$ similarly.
If we only want to know the expected value of Z, so $E[g(X, Y)] = Eg(X, Y)$, we can use the following property (without proof). Analogously to the computation of $Eg(X)$ in chapter 4, we can compute $Eg(X, Y)$ as the weighted average of the function values $g(x, y)$, weighing them with the corresponding joint probabilities $P(X = x \text{ and } Y = y)$.

Property 5.4.3 For a pair (X, Y) of discrete random variables we have:

$$\boldsymbol{Eg(X, Y) = \sum_{x \in S_X} \sum_{y \in S_Y} g(x, y) P(X = x \text{ and } Y = y)}$$

Example 5.4.4 In example 5.1.6 we considered random draws without replacement from a batch of 20 mega chips of which 13 were, and 7 were not, meeting quality specifications. We restricted ourselves to alternatives X_1 and X_2, variables that can take on only the values 1 or 0: 1 if the first resp. second is good and 0 if not.
X_1 and X_2 have both a $B\left(1,\frac{13}{20}\right)$-distribution, so $EX_1 = EX_2 = \frac{13}{20}$.
X_1 and X_2 are however dependent and their joint probability function should therefore be computed according the product rule of dependent events:

$$P(X_1 = i \text{ and } X_2 = j) = P(X_1 = i) \cdot P(X_2 = j | X_1 = i)$$

For example:

$$P(X_1 = 1 \text{ and } X_2 = 1) = P(X_1 = 1) \cdot P(X_2 = 1 | X_1 = 1) = \frac{13}{20} \cdot \frac{12}{19}$$

And applying property 5.4.3:

$$E(X_1 X_2) = \sum_{i=0}^{1} \sum_{j=0}^{1} i \cdot j \cdot P(X_1 = i \text{ and } X_2 = j) = 1 \cdot 1 \cdot P(X_1 = 1 \text{ and } X_2 = 1) = \frac{13}{20} \cdot \frac{12}{19}$$

Furthermore:

$$E(X_1 + X_2) = \sum_{i=0}^{1} \sum_{j=0}^{1} (i + j) \cdot P(X_1 = i \text{ and } X_2 = j)$$
$$= (0+0) \cdot P(X_1 = 0 \text{ and } X_2 = 0) + (0+1) \cdot P(X_1 = 0 \text{ and } X_2 = 1) + (1+0) \cdot P(X_1 = 1 \text{ and } X_2 = 0) + (1+1) \cdot P(X_1 = 1 \text{ and } X_2 = 1)$$
$$= 0 \cdot \frac{7}{20} \cdot \frac{6}{19} + 1 \cdot \frac{7}{20} \cdot \frac{13}{19} + 1 \cdot \frac{13}{20} \cdot \frac{7}{19} + 2 \cdot \frac{13}{20} \cdot \frac{12}{19} = \mathbf{\frac{26}{20}}$$

Since $EX_1 = EX_2 = \frac{13}{20}$, we showed for this example:

$$E(X_1 + X_2) = EX_1 + EX_2,$$

but

$$\frac{13}{20} \cdot \frac{12}{19} = E(X_1 \cdot X_2) \neq EX_1 \cdot EX_2 = \left(\frac{13}{20}\right)^2$$

■

The equality $E(X + Y) = EX + EY$ is a general property which can be proven using property 5.4.3. It is also valid if X and Y are dependent, as the previous example illustrates.

Property 5.4.5 For (discrete) random variables X and Y we have:

a. $\boldsymbol{E(X + Y) = E(X) + E(Y)}$.
b. **If X and Y are independent, then**: $\boldsymbol{E(XY) = E(X) \cdot E(Y)}$.

Proof:
a. We notice that according property 5.4.3, if $g(x, y) = x$:

$$E(X) = \sum_{x \in S_X} \sum_{y \in S_Y} x \cdot P(X = x \text{ and } Y = y)$$

So: $E(X + Y) = \sum_{x \in S_X} \sum_{y \in S_Y} (x + y) \cdot P(X = x \text{ and } Y = y)$

$$= \sum_{x \in S_X} \sum_{y \in S_Y} x \cdot P(X = x \text{ and } Y = y) + \sum_{x \in S_X} \sum_{y \in S_Y} y \cdot P(X = x \text{ and } Y = y)$$
$$= E(X) + E(Y)$$

b. $$E(XY) = \sum_{x \in S_X} \sum_{y \in S_Y} xy \cdot P(X = x \text{ and } Y = y)$$
$$\overset{ind.}{=} \sum_{x \in S_X} \sum_{y \in S_Y} xy \cdot P(X = x) \cdot P(Y = y)$$
$$= \sum_{x \in S_X} x \cdot P(X = x) \cdot \sum_{y \in S_Y} y \cdot P(Y = y) = E(X) \cdot E(Y)$$ ■

If we apply property 5.4.5b to example 5.4.1 (where X and Y are independent!) we get: $E(XY) = E(X) \cdot E(Y) = 3 \cdot 4 = 12$.
In the same example we showed that the sum of two independent, Poisson distributed random variables is Poisson distributed, with new parameter $\mu = E(X + Y) = E(X) + E(Y)$.

A similar property could be derived for the total number of kidney stone operations in a week (7 days), so the convolution of 7 independent, Poisson distributed numbers of operations, all with mean 4. For that goal we can easily extend 5.4.2 to n variables $X_1, X_2, \ldots, X_n$ (the proof uses induction w.r.t. n)

Property 5.4.6 $\boldsymbol{E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)}$.

Property 5.4.7 If $X_1, X_2, \ldots, X_n$ are independent, then so are $\boldsymbol{g(X_1, \ldots, X_{n-1})}$ and $\boldsymbol{X_n}$.

The last property implies, for example, that if $X_1, X_2, \ldots, X_n$ are independent, then $\sum_{i=1}^{n-1} X_i$ and X_n are independent as well. This consequence of 5.5.7 can be used to prove in general:

Property 5.4.8 If $X_1, X_2, \ldots, X_n$ are independent and en X_i has a Poisson distribution with parameter μ_i for each $i = 1, 2, \ldots, n$, then:

$$\boldsymbol{\sum_{i=1}^{n} X_i \text{ has a Poisson distribution with parameter } \mu = \sum_{i=1}^{n} \mu_i}$$

Example 5.4.9 (the expectation of the binomial and the hypergeometric distribution)
A population with N elements consists of R elements with a specific property (success) and the remaining $N - R$ elements without (failure).
Then the success proportion is $p = \frac{R}{N}$ and the failure rate $1 - p = \frac{N-R}{N}$.
For both drawing **with** and **without replacement** we can define alternatives for each draw i:

$$X_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ draw results in a success} \\ 0 & \text{if the } i^{\text{th}} \text{ draw results in a failure} \end{cases} \qquad (i = 1, 2, \ldots, n)$$

The marginal distribution of each X_i is for both cases (with and without replacement) $B(1, p)$, so $E(X_i) = p$ and $var(X_i) = p(1 - p)$.
The number of successes, X, can for both cases be expressed in the X_i's: $X = \sum_{i=1}^{n} X_i$
So the expected number of successes is according 5.4.6

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

We note that for his derivation it is not decisive whether or not the X_i`s are independent: When drawing **with replacement** $X_1, X_2, \ldots, X_n$ are independent and $\sum_{i=1}^{n} X_i$ has a $B(n,p)$-distribution and, when drawing **without replacement**, $X_1, X_2, \ldots, X_n$ are dependent and $\sum_{i=1}^{n} X_i$ has a hypergeometric distribution (with parameters N, R en n). ■

In the previous example we modelled the binomially and hypergeometrically distributed numbers, using alternatives for each of the trials. This approach will be applied more often as properties can easier be derived using the summation of n alternatives, e.g. for finding the binomial and hypergeometric variance formulas in the next section.

5.5 Correlation

In example 5.2.7 we saw that X, the number of reported bicycle thefts on a day in a town, is related to actual number of bicycle thefts $Y : X \leq Y$.
Large values of X coincide with large values of Y and small values of X mean small expected values of Y. We would like to characterize the strength of this relation in one numerical value: a **measure of relation** (**dependence**). We will do so by considering the deviations w.r.t. the expected values μ_X and μ_Y. The covariance is defined as "the mean product of deviations of X and Y, as follows:

Definition 5.5.1 The **covariance** of two random variables X and Y is defined as

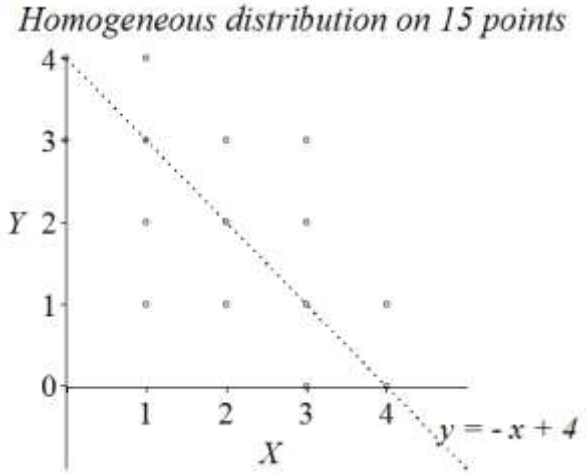
$$\boldsymbol{cov(X,Y) = E(X-\mu_X)(Y-\mu_Y)}$$

According property 5.4.3, where $g(X,Y) = (X-\mu_X)(Y-\mu_Y)$ is the function of X and Y, this value can be computed as the mean of products of deviations:

$$E(X-\mu_X)(Y-\mu_Y) = \sum_{x \in S_X} \sum_{y \in S_Y} (x-\mu_X)(y-\mu_Y) \cdot P(X = x \text{ and } Y = y)$$

The next example clarifies which information about the relation of X and Y can be deduced from the value of the covariance.

Example 5.5.2 The range of (X,Y) consists of 15 sample points, all having a probability $\frac{1}{15}$ of occurrence. These points are grouped around the line $y = -x + 4$, as the graph below shows.



The relation of X and Y is said to have "negative correlation": if X attains (relatively) large values, then Y takes on relatively small values. Reversely, if X is small, then Y is large.
The line which fits the "cloud" of 15 points best (minimizing the total distance of points to the line) has a negative slope.
The marginal probability functions of X and Y are the same and both symmetric:

x	0	1	2	3	4	Total
$P(Y=x)=P(X=x)$	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{2}{15}$	1

Using the symmetry of the distributions, we can state that $\mu_X = \mu_Y = 2$
X and Y are not independent, since, e.g.:

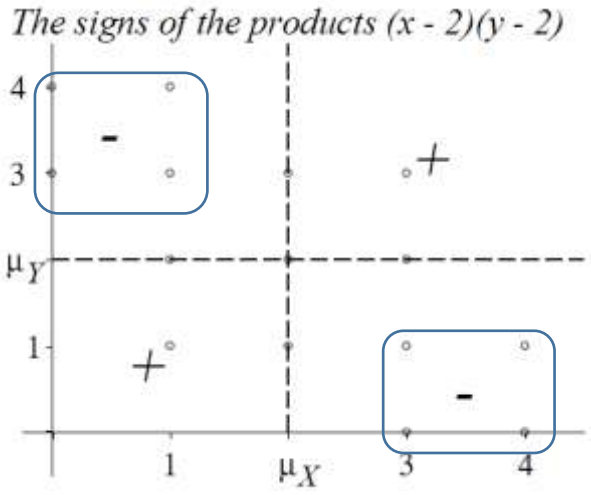
$$P(X=4 \text{ and } Y=4) = 0, \text{ where } P(X=4)\cdot P(Y=4) = \left(\tfrac{2}{15}\right)^2.$$

We will compute $cov(X,Y)$ as the weighted average of products $(x-\mu_X)(y-\mu_Y)$ with probabilities $(X=x \text{ and } Y=y) = \frac{1}{15}$, so

$$cov(X,Y) = \sum_{(x,y)} (x-\mu_X)(y-\mu_Y)\cdot\tfrac{1}{15}$$

If we split the 15 points in the first quadrant in 4 parts, using the lines $x=\mu_X$ and $y=\mu_Y$, as shown in the graph below, then we observe for the following points:

- If $(x,y)=(3,3)$: $(x-\mu_X)(y-\mu_Y) = (3-2)\cdot(3-2) = +1$
- If $(x,y)=(0,4)$: $(x-\mu_X)(y-\mu_Y) = (0-2)\cdot(4-2) = -4$
- If $(x,y)=(3,1)$: $(x-\mu_X)(y-\mu_Y) = (3-2)\cdot(1-2) = -1$



The determination of all the products and their signs show that the value of $cov(X,Y)$ is negative for his example:
there are 8 points with a negative product and only 2 with a positive product.
Moreover, the absolute value of the positive products are the smallest possible.

Computation leads to the value

$$cov(X,Y) = -\frac{16}{15}$$

So: X and Y have a negative correlation. ■

In example 5.5.2 the covariance is negative, because the overall grouping of the points in the xy-plane lies around a line with negative slope.
Similarly, e.g. by replacing Y in example 5.5.2 by $Z = 4 - Y$ (the points of the graph in the example are reflected about the line $y=2$), we will see that the covariance is positive and the points in the range of (X,Y) will lie "with large probability near" a line with positive slope.
In example 5.2.4 (bicycle thefts which are partially reported) we would expect a positive covariance between the number of reported thefts (X) and the number of actual thefts (Y).
Computation leads to the value $cov(X,Y) = 4$.

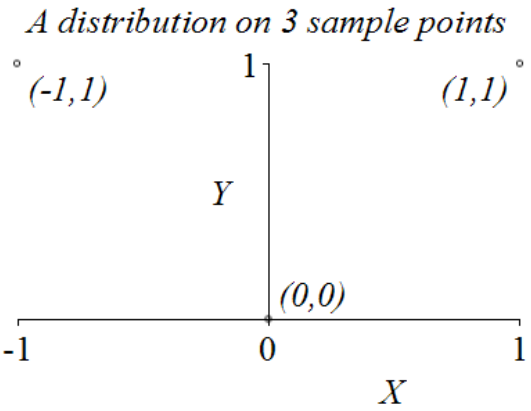
We will now show that in case of independence, there is no correlation: $cov(X,Y) = 0$. In that case X and Y are not correlated. Furthermore the second part of the property gives a formula that is helpful in computing the covariance.

Property 5.5.3 **a.** $\boldsymbol{cov(X,Y) = E(XY) - \mu_X\mu_Y}$.
b. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **independent**, then $\boldsymbol{cov(X,Y) = 0}$.

Proof:
a. $cov(X,Y) = E(X-\mu_X)(Y-\mu_Y)$
$= E(XY - X\cdot\mu_Y - \mu_X\cdot Y + \mu_X\mu_Y)$
$= E(XY) - \mu_Y\cdot E(X) - \mu_X\cdot E(Y) + \mu_X\mu_Y$ (properties 4.4.3 and 5.4.5)
$= E(XY) - \mu_X\mu_Y$
b. Consequence of a. and property 5.4.5 b: $E(XY) = E(X)\cdot E(Y)$ ■

The statement in the b-part of this property cannot be reversed: **"no correlation" does not imply independence** as the following example illustrates. But reversely, if the variables are correlated ($cov(X,Y) \neq 0$), then X and Y are **dependent** (not independent)

Example 5.5.4
The joint distribution of X and Y is given by three probabilities
$P(X=-1 \text{ and } Y=1) = P(X=0 \text{ and } Y=0) = P(X=1 \text{ and } Y=1) = \frac{1}{3}$



The marginal distributions are:

i	-1	0	1
$P(X=i)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$P(Y=i)$	0	$\frac{1}{3}$	$\frac{2}{3}$

So $\mu_X = 0$ (symmetry) and
$\mu_Y = \sum_i i\cdot P(Y=i) = 0\cdot\frac{1}{3} + 1\cdot\frac{2}{3} = \frac{2}{3}$

X and Y are dependent: e.g. $P(X=0 \text{ and } Y=1) = 0$, but $P(X=0)\cdot P\,(Y=1) = \frac{1}{3}\cdot\frac{2}{3}$.
(Intuitively: if $Y=1$, X can only be -1 or 1, and if $Y=0$, X can only be 0: X and Y are dependent.)
$E(XY) = -1\cdot 1\cdot\frac{1}{3} + 0\cdot 0\cdot\frac{1}{3} + 1\cdot 1\cdot\frac{1}{3} = 0$, so using property 5.5.3a we find:

$$cov(X,Y) = E(XY) - \mu_X\mu_Y = 0 - 0\cdot\frac{2}{3} = 0$$

X and Y are **not correlated, but they are dependent.** ■

The covariance, the mean product of deviations can be zero for dependent variables, but we noticed that the covariance deviates from 0 especially if the points (x,y) are close to a line (with large probability). This observation reflects the idea that covariance is **measure of linear relation.** So, if the points are close to a line $y = ax + b$ this is noticed by the covariance, not if the type of relation of X and Y is quadratic (points close to a parabola) or circular, etc.

When we wonder whether the absolute value of $cov(X,Y)$ is informative, we must conclude that it is not: the absolute value depends on the unit of measurement that is chosen as is shown in the c-part of the following property.

Property 5.5.5 (Properties of the covariance)

a. $\boldsymbol{cov(X,X) = var(X)}$.
b. $\boldsymbol{cov(X,Y) = cov(Y,X)}$.
c. $\boldsymbol{cov(aX+b,Y) = a \cdot cov(X,Y)}$, for $a \in \mathbb{R}$ and $b \in \mathbb{R}$.
d. $\boldsymbol{cov(X+Y,Z) = cov(X,Z) + cov(Y,Z)}$.
e. $\boldsymbol{var(X+Y) = var(X) + var(Y) + 2cov(X,Y)}$ and
$\boldsymbol{var(X-Y) = var(X) + var(Y) - 2cov(X,Y)}$

Proof:
a. $cov(X,X) = E(X-\mu_X)(X-\mu_X) = E(X-\mu_X)^2 = var(X)$.
b. This is a consequence of the symmetry in the definition of $cov(X, Y)$: interchanging X and Y leads to the same expression
c. $cov(aX+b,Y) = E[\big(aX+b-E(aX+b)\big) \cdot (Y-\mu_Y)]$
$= E[a \cdot (X-\mu_X) \cdot (Y-\mu_Y)]$
$= a \cdot E(X-\mu_X)(Y-\mu_Y)$
$= a \cdot cov(X,Y)$
d. Follows directly by applying the definition of covariance.
e. $var(X+Y) = cov(X+Y, X+Y)$, according a.
$= cov(X,X) + cov(X,Y) + cov(Y,X) + cov(Y,Y)$, using d. and b.
$= var(X) + 2 \cdot cov(X,Y) + var(Y)$, now using a. and b.
So $var(X-Y) = var\big(X + (-Y)\big)$
$= var(X) + var(-Y) + 2cov(X,-Y)$
$= var(X) + var(Y) - 2cov(X,Y)$,
since $var(-Y) = (-1)^2 var(Y) = var(Y)$ ■

Property 5.5.5c implies that $cov(X,Y)$ **depends** on the chosen **unit of measurement:** if X is a random length in meters and we decide to give the same lengths in centimetres (we have $100X$ instead of X), then the value of the covariance increases:
$cov(100X,Y) = 100 \cdot cov(X,Y)$. Therefore we want a **measure of linear relation** which does not depend on the unit, by dividing the covariance by both the standard deviations of X and Y (provided that $\sigma_X > 0$ and $\sigma_Y > 0$).

Definition 5.5.6 The **correlation coefficient** $\rho(X,Y)$ of two random variables X and Y is defined by

$$\boldsymbol{\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}}$$

Example 5.5.7 In a large survey on the relation between the use of toothpaste with Fluor and tooth decay (caries), found by the dentist during periodic checks, we will define an **indicator variable** X to be 1, if the dentist has to repair a hole in a tooth and $X = 0$ if not.
Y is the indicator variable for using toothpaste with Fluor ($Y = 1$) or without Fluor ($Y = 0$).

i \ j	0	1
0	$\frac{1}{9}$	$\frac{1}{3}$
1	$\frac{1}{3}$	$\frac{2}{9}$

After extensive research the following (estimates of) probabilities
$P(X = i \text{ and } Y = j)$ were found:

If we treat the estimates as the real probabilities, then the marginal probability functions of X and Y are the same, e.g. $P(X = 1) = \frac{5}{9}$ and $P(X = 0) = \frac{4}{9}$, so we find:

$$\mu_X = \mu_Y = \frac{5}{9} \text{ and } \sigma_X = \sigma_Y = \sqrt{\frac{20}{81}}.$$

We will compute the covariance with the computational formula $cov(X,Y) = E(XY) - \mu_X\mu_Y$

$$E(XY) = \sum_x \sum_y x \cdot y \cdot P(X = x \text{ en } Y = y) = 1 \cdot 1 \cdot \frac{2}{9} = \frac{2}{9}$$

and $$cov(X,Y) = E(XY) - \mu_X\mu_Y = \frac{2}{9} - \frac{5}{9} \cdot \frac{5}{9} = -\frac{7}{81}$$

The correlation coefficient is:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X\sigma_Y} = \frac{-\frac{7}{81}}{\sqrt{\frac{20}{81}} \cdot \sqrt{\frac{20}{81}}} = -\frac{7}{20}$$

Covariance and correlation coefficient are both negative: X and Y are negatively correlated, meaning that caries ($X = 1$) coincides more often with no Fluor ($Y = 0$), and reversely. ■

If the correlation coefficient of X and Y is not 0, as is the case in the previous example, then in conclusion there is dependence of the phenomena "use of Fluor" and "caries". But we cannot draw the conclusion that use of Fluor containing toothpaste **causes** less caries.
The correlation of the phenomena might be explained by the fact that people who take good care of their teeth (e.g. by not eating sweets and brushing their teeth regularly), also choose often for toothpaste with Fluor because of its alleged advantages.
Famous is the example of the correlation between the decrease of storks (the birds of whom children are told to carry in babies) and the decrease of new born babies in the Netherlands in the late 60`s. The phenomena did occur simultaneously, but one did not cause the other.
So, **beware of interpreting correlation of two variables as a causal relation**
The strength of the dependence is not given by the covariance, but the correlation coefficient does give us this information: the following property states that $\rho(X,Y)$ takes on its extreme values $+1$ or -1, if there is a strict linear relation between (the values) of X and Y:

Property 5.5.8 (properties of a correlation coefficient)

a. $\boldsymbol{\rho(aX + b, Y)} = \begin{cases} \boldsymbol{\rho(X,Y)} & \text{if } a > 0 \\ \boldsymbol{-\rho(X,Y)} & \text{if } a < 0 \end{cases}$

b. $\boldsymbol{-1 \le \rho(X,Y) \le 1}$

c. If $Y = aX + b$, then $\boldsymbol{\rho(X,Y)} = \begin{cases} \boldsymbol{+1} & \text{if } a > 0 \\ \boldsymbol{-1} & \text{if } a < 0 \end{cases}$
and reversely, if $\rho(X,Y) = 1$, then $Y = aX + b$, with $a > 0$.
and if $\rho(X,Y) = -1$, then $Y = aX + b$, with $a < 0$

Proof:

a. $cov(aX + b, Y) = a \cdot cov(X,Y)$ according property 5.5.5c and property 4.4.9, that is, $var(aX + b) = a^2 var(X)$, so: $\rho(aX + b, Y) = \frac{cov(aX+b,Y)}{\sqrt{var(aX+b) \cdot var(Y)}} = \frac{a cov(X,Y)}{|a| \cdot \sigma_X \cdot \sigma_Y} = \frac{a}{|a|}\rho(X,Y)$

b. We will use that $var\left(\frac{X}{\sigma_X}\right) = \left(\frac{1}{\sigma_X}\right)^2 var(X) = 1$ and according 5.5.5e for $\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}$:

$$var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = var\left(\frac{X}{\sigma_X}\right) + var\left(\frac{Y}{\sigma_Y}\right) + 2cov\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) = 1 + 1 + 2 \cdot \rho(X,Y)$$

Since $var\left(\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}\right) \geq 0$, we have $2+2\rho(X,Y) \geq 0$, so $\rho(X,Y) \geq -1$

Similarly we can use $var\left(\frac{X}{\sigma_X}-\frac{Y}{\sigma_Y}\right) \geq 0$, to show that $\rho(X,Y) \leq 1$.

c. By $Y = aX + b$ we mean: $P(Y = aX + b) = 1$. The pairs of values (x, y), that can occur (with positive probability) lie on the line. See also exercise 11. ■

The strength of the dependence or correlation can be classified as follows:

- $\rho = 0$: *no correlation.*
- $\rho > 0$: *positive correlation* and $\rho < 0$: *negative correlation.*
- $|\rho| = 1$: *strict linear correlation.*
- $0.9 \leq |\rho| < 1$: *strong correlation.*
- $0 < |\rho| < 0.9$: *weak* (≤ 0.3) or *moderate* (≥ 0.3) *correlation.*

In example 5.5.7 found $\rho(X,Y) = -0.35$: a moderately negative correlation of X and Y.

In property 5.5.5e we noticed that the variance of $X + Y$, in general, is not equal to the addition of $var(X)$ and $var(Y)$. But the equality is valid if X and Y are independent. These properties can be extended to n random variables (without formal proof):

Property 5.5.9

a. $$\boldsymbol{var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) + \sum_{i \neq j}\sum cov(X_i, X_j)}$$

b. If $X_1, \ldots, X_n$ are independent, then: $$\boldsymbol{var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i)}$$

Note 5.5.10: these properties for the variance of a sum of variables are much more simple than the ones you would get if $E|X-\mu|$ would have been chosen as a measure of variation. The $n \times n$ terms in the right hand side of the equation in a. is often given in a so called **covariance matrix**, having variances on the main diagonal ($= cov(X_i, X_i)$):

$$\begin{bmatrix} var(X_1) & cov(X_1,X_2) & \cdots & cov(X_1,X_n) \\ cov(X_2,X_1) & var(X_2) & & cov(X_2,X_n) \\ \vdots & & \ddots & \vdots \\ cov(X_n,X_1) & cov(X_n,X_2) & \cdots & var(X_n) \end{bmatrix}$$

■

We will now apply property 5.5.9 to derive the variance formulas of the binomial and the hypergeometric distributions (referring to section 4.5).
If the trials are independent, as is the case when drawing **with replacement** from a dichotomous population, we have seen in example 5.4.9, that we can define alternatives X_i for each Bernoulli trial: X, the total number of successes in n trials, can be expressed in the X_i`s:

$$X = \sum_{i=1}^{n} X_i,$$

We saw before that X has a $B(n,p)$-distribution: $E(X) = \sum_{i=1}^{n} E(X_i) = n \cdot p$.
Because of the independence of the X_i's we can apply property 5.5.9b:

$$var(X) = var\left(\sum_{i=1}^{n} X_i\right) \overset{\text{ind.}}{=} \sum_{i=1}^{n} var(X_i) = n \cdot p(1-p)$$

(As before we used the 1-0 distribution of the X_i`s: $var(X_i) = E(X_i^2) - (EX_i)^2 = p - p^2$)

If we draw n times **without replacement** from a dichotomous population with R "successes" (red balls) and $N - R$ "failures" (white balls), then we define for the i^{th} draw the alternative X_i, having values 1 and 0 for a success and a failure, resp. Clearly, the X_i's are **dependent**, but the unconditional distributions of the X_i's remain $B(1,p)$-distributions, where $p = \frac{R}{N}$ is the success probability. Now we will have to apply property 5.5.9a to find the variance:

$$var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) + \sum_{i \neq j} \sum cov(X_i, X_j)$$

In this expression all variances are the same: $var(X_i) = p(1-p) = \frac{R}{N} \cdot \left(1 - \frac{R}{N}\right)$
Because of symmetry all covariance`s (there are $n^2 - n$ covariance`s in the last summation) are the same as well. Computing one of them is sufficient:

$$cov(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2),$$

where:

$E(X_1) = E(X_2) = \frac{R}{N}$ and
$E(X_1 X_2) = \sum\sum i \cdot j \cdot P(X_1 = i \text{ and } X_2 = j) = 1 \cdot 1 \cdot P(X_1 = 1 \text{ and } X_2 = 1)$
$= P(X_1 = 1) \cdot P(X_2 = 1 | X_1 = 1) = \frac{R}{N} \cdot \frac{R-1}{N-1}$ (The probability of 2 successes")

Substituting these results we find:

$$var(X) = var\left(\sum_{i=1}^{n} X_i\right) = n \cdot var(X_1) + (n^2 - n) cov(X_1, X_2)$$
$$= n \cdot \frac{R}{N}\left(1 - \frac{R}{N}\right) + (n^2 - n)\left[\frac{R}{N} \cdot \frac{R-1}{N-1} - \frac{R}{N} \cdot \frac{R}{N}\right]$$
$$= \cdots = n \cdot \frac{R}{N}\left(1 - \frac{R}{N}\right) \cdot \frac{N-n}{N-1}$$

In chapter 4 we noticed that with p instead of $\frac{R}{N}$ in the formula above, it resembles the variance formula of the binomial distribution.
The extra factor $\frac{N-n}{N-1}$ is referred to as the **correction factor for a finite population**: the factor tends to 1 for large populations ($N \to \infty$). Then the hypergeometric distribution is well approximated by the binomial distribution (property 4.5.5).

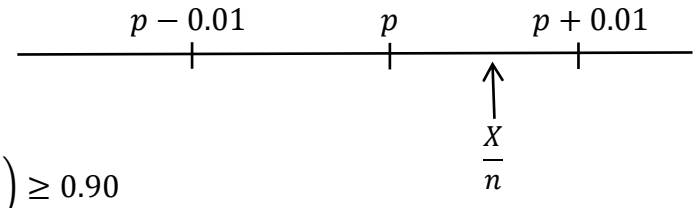
5.6 The weak law of large numbers

Example 5.6.1 For mass production of e.g. chips, resistances or sensors, a quality check is often performed by taking a random sample from the production and determining the fraction of rejected products. The **empirical law of large numbers** (chapter 1) told us that the proportion of successes "on the long run" will approximate the probability p of a rejected product very closely. Experimental practice tells us so, but it does not quantify the notion "very close to p", or what sample size should be chosen.
Now we developed probability models which we can use to describe these situations very well. In this case we can define a random variable X as "the number of defective products in a random sample of n products". The $B(n,p)$-distribution of X is a correct choice assuming that the sampling is with replacement, or a good approximating distribution if the total production is (very) large.
The relative frequency, the **sample proportion** $\frac{X}{n}$, is "on average" equal to the population proportion p. This is confirmed by computing the expected value and the variance:

- $E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n} \cdot np = p$
- $var\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right)^2 var(X) = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}$

In conclusion: the expected value of $\frac{X}{n}$ is p and the variation of $\frac{X}{n}$ (around p) decreases as n increases. We can quantify this process, by computing the minimum value of n, such that the probability, that $\frac{X}{n}$ deviates no more than 0.01 from p, is at least 90%.

In a sketch:



So: $P\left(\left|\frac{X}{n} - p\right| < 0.01\right) \geq 0.90$
Or: $P\left(\left|\frac{X}{n} - p\right| \geq 0.01\right) \leq 0.10$

If we apply Chebyshev`s rule: $P\left(\left|\frac{X}{n} - p\right| \geq 0.01\right) \leq \frac{var\left(\frac{X}{n}\right)}{0.01^2}$
The condition is fulfilled if $\frac{var\left(\frac{X}{n}\right)}{0.01^2} = \frac{p(1-p)}{0.01^2 n} \leq 0.10$, so if $n \geq 100000 \cdot p(1-p)$
Because $f(p) = p(1-p)$ is at most $\frac{1}{4}$ for $0 \leq p \leq 1$, we find $n \geq 25000$.

Stating the previous more generally, than for each (small) interval $(p-c, p+c)$ we have, according Chebyshev`s rule:

$$P\left(\left|\frac{X}{n} - p\right| \geq c\right) \leq \frac{p(1-p)}{c^2 n}$$

Conclusion: for any of these intervals we can find a sample size n such that the probability that $\frac{X}{n}$ deviates more than c from p, is as small as we wish. ■

It should be noted that the sample proportion $\frac{X}{n}$ can be interpreted as a sample mean: a $B(n,p)$-distributed variable X can be written as the summation of n independent alternatives $X_1,\ldots,X_n$, all with success probability p.
So $\frac{X}{n}=\frac{\sum_{i=1}^{n}X_i}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i$, which is often denoted $\overline{X}$ or $\overline{X}_n$.

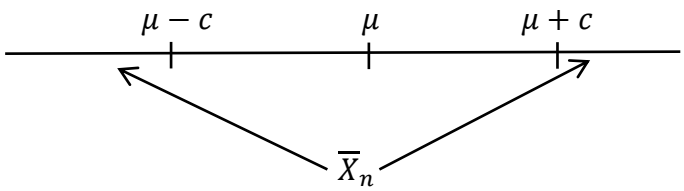
Sample means $\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$ play an important role in statistics: if we draw a **random sample** of size n from a population with mean μ and variance σ^2, the observed values in the sample are usually modelled as independent variables $X_1,\ldots,X_n$ all with te same (population) distribution having mean μ and variance σ^2. In example 5.6.1 is $\mu=p$ and $\sigma^2=p(1-p)$.

Property 5.6.2 The weak law of large numbers

If $X_1,X_2,\ldots$ are independent and all have the same distribution with expectation μ and variance σ^2, then for the mean $\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$ and every constant $c>0$ we have:

$$\lim_{n\to\infty} P(|\overline{X}_n-\mu|\geq c)=0$$

In a sketch:



Proof: We will use Chebyshev`s inequality, where X is substituted by $\overline{X}_n$.
For the mean $\overline{X}_n$ we have:

$$E(\overline{X}_n)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\cdot n\cdot\mu=\mu \quad \text{and}$$

$$var(\overline{X}_n)=var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\left(\frac{1}{n}\right)^2 var\left(\sum_{i=1}^{n}X_i\right)\overset{\text{o.o.}}{=}\left(\frac{1}{n}\right)^2\sum_{i=1}^{n}var(X_i)=\frac{1}{n^2}\cdot n\cdot\sigma^2=\frac{\sigma^2}{n}$$

Chebyshev: $P(|\overline{X}_n-\mu|\geq c)\leq\frac{var(\overline{X}_n)}{c^2}=\frac{\sigma^2}{nc^2}$
So $\lim_{n\to\infty}P(|\overline{X}_n-\mu|\geq c)=0$ ■

The limit is sometimes expressed in words by: "$\overline{X}_n$ **converges in probability to μ**".

Applied to example 5.6.1 the consequence is that the relative frequency $f_n(A)=\frac{X}{n}=\overline{X}_n$ of the event $A=$ "product is defective" converges in probability to the unknown p.
The weak law of large numbers confirms mathematically what the experimental law in section 1.3 conjectured: the relative frequency $f_n(A)$ "converges" to p.

5.7 Exercises

1. We toss a fair coin 4 times.
X is the number of tails in all 4 tosses and Y is the number of tails in the last two tosses.

- **a.** Determine the joint probability function of X and Y.
  For that goal, first make a list of all 2^4 outcomes of the 4 tosses and register for each of the outcomes the observed values of X and Y.
- **b.** Determine the probability function of X, given $Y = 1$, and $E(X|Y = 1)$.
- **c.** Compute $P(Y = 1|X = 3)$.

2. The joint probability function of X and Y is given by the formula

$$P(X = i \text{ and } Y = j) = \left(\frac{1}{3}\right)^{i-j}\left(\frac{2}{3}\right)^{1+j}, \qquad \text{where } i = 1, 2, 3, \ldots \text{ and } j = 0, 1$$

- **a.** Sketch all possible values of the pair (X, Y) by its grid points in the xy-plane.
- **b.** Show the marginal distribution of X is geometric and give $E(X)$.
- **c.** Determine the marginal probability function of Y, so $P(Y = 0)$ and $P(Y = 1)$.
- **d.** Are X and Y independent? (Motivate your answer).

3. A small factory works in a morning and an evening shift.
For an arbitrary day random variables X and Y are defined as:
$X =$ "the number of absent employees in the morning shift" and
$Y =$ "the number of absent employees in the evening shift".
The Human Resources department provided the statistics (over many years) for the joint probability function $P(X = x \text{ and } Y = y)$ of X and Y:

x \ y	0	1	2	3
0	0.05	0.05	0.10	0
1	0.05	0.10	0.25	0.10
2	0	0.15	0.10	0.05

- **a.** Determine the (marginal) distributions of X and of Y.
- **b.** Compute the expectation and the variance of X and Y.
- **c.** Consider $Z = 8 \cdot Y$, the number of lost labour hours caused by absenteeism in the evening shift. Give the probability function of Z, $E(Z)$ and $var(Z)$.
  According the rules for expectation and variance we have: $E(Z) = 8E(Y)$ and $var(Z) = 64var(Y)$. Check the correctness.
- **d.** Determine the probability function of $T = X + Y$ and compute $E(T)$ and $var(T)$
- **e.** Check the equality $E(X + Y) = E(X) + E(Y)$.
- **f.** Check that $var(X + Y) \neq var(X) + var(Y)$ for this distribution.
  Explain why $var(X + Y)$ for this case is not the same as $var(X) + var(Y)$.

4. We define N as the number of tosses with an unbiased coin until we toss a tail. If N realizes the number n (e.g. 10) we will toss the coin another n times. The number of tails in the second series of n tosses is the random variable X.

a. Determine $P(N = 10)$, $P(X = 4|N = 10)$ and $P(X = 4 \text{ and } N = 10)$.
b. Determine: the distribution of N,
the conditional distribution of X, given $N = n$ and
the joint distribution of X and N.
c. Determine $E(X|N = 10)$, $E(X|N = n)$ and,
using property $E(X) = E[E(X|N)]$, the expectation $E(X)$.
d. Determine $P(X = 0)$.
e. Determine the conditional probability function of N, given $X = 0$, and $E(N|X = 0)$.

5. Assume that the number of accidents, N, during the evening rush hour in a town has a Poisson distribution with parameter μ. In this simple model the damage to cars at each accident is either € 1000, € 2000, € 3000 or € 4000, which occurs with probabilities 0.1, 0.3, 0.4 and 0.2, respectively.
Let S be the total damage during that rush hour and let X_i be the damage of the i^{th} accident if $i = 1, 2, \ldots$. We assume that the amounts of damage (X_i`s) are independent. The goal of this exercise is to find the value of $E(S)$.

a. Describe S as a function of the X_i's if we know that the number of accidents is $N = n$.
b. Compute EX_i.
c. Determine $E(S|N = n)$.
d. Determine ES (use property 5.2.6: $E(X) = E[\, E(X|Y)\,]$).
e. Is $ES = E(X_i) \cdot E(N)$?
Note: $var(S) \neq var(X_i) \cdot E(N)$ as can be found in many books, e.g. on supply chain management.

6. X is the number of customers entering an office of a bank during half an hour: assume X has a Poisson distribution with parameter μ = 10.
Furthermore, Y is the number of the entered customers with a service demand that takes longer than 3 minutes. We will assume that Y has, given $X = x$, a $B(x, 0.3)$-distribution, a binomial distribution with parameters x and 0.3.

a. Compute $P(X = 8 \text{ and } Y = 2)$. (Round answers in 3 decimals).
b. Compute $E(Y|X = 8)$ and express $E(Y|X = x)$ in x.
c. Compute $E(Y)$.

7. The products on two conveyor belts, A and B, are checked: their quality is either good or bad. On each belt the products can be numbered: 1, 2, 3, … The quality of each product is good at a rate $P(good) = 0.9$, independent of the quality of other products.
Let X_1 and X_2 be the number of the first bad product in belt A and belt B, respectively.

a. Compute $P(X_1 = 10)$.
b. Compute $P(20 \leq X_1 \leq 30)$.
c. Compute $P(X_1 = X_2)$.
d. Compute $P(X_1 + X_2 = 20)$, using the convolution formula.

8. X and Y are independent random variables, geometrically distributed with parameter p.
In this exercise we will, step by step, show that the distribution of the minimum of X and Y is geometric as well.

a. Compute $P(X > i \text{ and } Y > i)$.

b. Compute $P(min(X,Y) > i)$.
c. Determine $P(min(X,Y) = i)$.
d. Which (geometric) distribution does $min(X,Y)$ have? Determine $E[min(X,Y)]$.

9. X and Y are independent and both have a geometric distribution with parameter p.

a. Express $E(X+Y)$ and $var(X+Y)$ in p.
b. Find the joint probability function of X and Y and sketch the range $S_X \times S_Y$ in $\mathbb{R}^2$.
c. Determine the probability function of $X+Y$, by applying the Convolution sum.

10. Consider the four joint distributions of X and Y below, given by $P(X = i \text{ and } Y = j)$ in the table:

Distribution 1

i \ j	0	1	2
0	0.2	0.1	0
1	0.1	0.2	0.1
2	0	0.1	0.2

Distribution 2

i \ j	0	1	2
0	$\frac{1}{20}$	$\frac{1}{5}$	$\frac{1}{20}$
1	$\frac{1}{5}$	0	$\frac{1}{5}$
2	$\frac{1}{20}$	$\frac{1}{5}$	$\frac{1}{20}$

Distribution 3

i \ j	0	1	2
0	0.09	0.12	0.09
1	0.12	0.16	0.12
2	0.09	0.12	0.09

Distribution 4

i \ j	0	1	2
0	0	0	0.3
1	0	0.4	0
2	0.3	0	0

a. Check (determine) that the marginal probability functions of X and of Y are the same for all four distributions. And determine $E(X)$, $var(X)$, $E(Y)$ and $var(Y)$.
b. Examine the dependence of the 4 distributions well, e.g. by drawing the points with a non-zero probability in the xy-plane. Then choose for all 4 distributions a value for $\rho(X,Y)$, taken from the set $\{-2, -1, -\frac{2}{3}, 0, \frac{2}{3}, 1, 2\}$.
c. For which of the distributions are X and Y independent. Motivate your answers briefly.
d. Determine for distribution 1: $E(XY)$, $cov(X,Y)$ and $\rho(X,Y)$
e. Determine for distribution 1 as well: $cov(3X, 2-Y)$ and $\rho(3X, 2-Y)$
f. Show that for distribution 2 we have: $\rho(X,Y) = 0$.
Can we conclude from this value that X and Y are independent? Why (not)?
g. Determine for distribution 1: the distribution of X, given $Y = 0$, and $E(X|Y=0)$.
Repeat this for the condition $Y = 1$ and for $Y = 2$.
Finally, check the equality $E(X) = \sum_y E(X|Y=y) \cdot P(Y=y)$.

11. Use the properties of covariance and correlation coefficient to determine $\rho(X,Y)$ if $Y = -3X + 4$ (or, more precise: $P(Y = -3X + 4) = 1$)

12. $X_1, X_2, \ldots, X_n$ are independent and all have the same distribution with expectation 1 and variance 2.

a. Determine with rules for covariance and correlation coefficient $cov(X_1, X_1 + X_2)$ and $\rho(X_1, X_1 + X_2)$.
b. Compute the smallest value of n such that $\rho(X_1, X_1 + X_2 + \ldots + X_n) < \frac{1}{3}$.

13. (former exam exercise)
Ten balls, numbered 1 to 10, are placed in an arbitrary order (positions 1 to 10)
Define: $X_i = \begin{cases} 1 & \text{if the ball with number } i \text{ is in position } i \\ 0 & \text{otherwise} \end{cases}$, $i = 1, 2, \ldots, 10$,
so $S = X_1 + X_2 + \cdots + X_{10}$ is the number of balls in the "right" position.

a. Compute $E(X_1)$, $var(X_1)$ and $cov(X_1, X_2)$.
b. Compute $E(S)$ and $var(S)$.

14. (former exam exercise)
A surgeon operates, on average, 2 persons who suffer from (acute) appendicitis and 3 who suffer from kidney stones in a week. Last week he had a total of 7 of these operations.

a. If X and Y are independent and have a Poisson distributions with parameters μ_1 and μ_2, respectively, <u>give</u> the distribution $X + Y$ (do not repeat the derivation!).
b. Show that X, given $X + Y = n$, has a $B\left(n, \frac{\mu_1}{\mu_1 + \mu_2}\right)$-distribution.
c. Determine the expected number appendicitis operations last week, given the total of 7 operations. Which (reasonable?) assumptions are necessary to apply the property in a.?

Some hints for solution of the exercises of chapter 5:

1. Make a table with 3 columns, in the first the all 4 digits outcomes, in the second the value of X and in the third the value of Y (e.g. for 0010 $X = 1$ and $Y = 1$). Then you can determine the joint probabilities $P(X = i \text{ and } Y = j)$ by counting the number of occurrences of a specific combination, such as $(X, Y) = (1, 1)$
2. Note that Y can only take on the values 0 and 1: substituting $j = 0$ you will find all probabilities $P(X = i \text{ and } Y = 0)$ (the grid points on the X-axis as from (1, 0)): The summation of these probabilities will give you $P(Y = 0)$. Note that you need the geometric series to complete the summation (see appendix mathematical techniques). Similar for $Y = 1$.
3. Marginal distribution of X: add the probabilities in each row. Make a 4 rows table to compute EX and $E(X^2)$. The first two rows are for the values x of X and their probabilities $P(X = x)$; row 3 and 4 contain $x \cdot P(X = x)$ and $x^2 \cdot P(X = x)$. Addition of these rows will give you $E(X)$ and $E(X^2)$. Furthermore apply $var(X) = E(X^2) - (EX)^2$. **d**. if e.g. $T = 3$, (X, Y) could be (1, 2) or (2, 1). (note that (0, 3) has probability 0).
4. Note that X depends on N, but N is variable.
5. The problem in this exercise is that S is the sum of X_1 to X_N, but the number N is a random variable. The exercise shows that the intuitive approach $E(S) = E(X) \cdot E(N)$ is justified.
6. Similar as 5.
7. **b**. Use $P(X > k) = (1 - p)^k$ for the geometric distribution: $P(X_1 \geq 20) = P(X_1 > 19)$
8. **a**. Use the property $P(X > k) = (1 - p)^k$ for the geometric distribution.
9. –
10. –
11. Substitute $Y = -3X + 4$ in the numerator and the denominator of the formula of $\rho(X, Y)$. Do **not** use $cov(X, Y) = E(XY) - EX \cdot E(Y)$, but $cov(aX + b, Y) = \cdots$.
12. Substitute $X = X_1$ and $Y = X_1 + X_2$ in the numerator and the denominator of the formula of $\rho(X, Y)$. Do **not** use $cov(X, Y) = E(XY) - E(X) \cdot E(Y)$, but $cov(X, Y + Z) = \cdots$.
13. What is the value of $P(X_i = 1)$?
14. –

Chapter 6 Continuous random variables

6.1 Density function, expectation and variance of a continuous variable

In chapters 4 and 5 we discussed mainly the distribution of integer valued variables, quantities that can take on a finite or a numerable number of values. We used the probability function $P(X = x)$ to specify the distribution of one variable X.
Many stochastic experiments lead, however, to real valued results, e.g.:

- Measure the IQ of an arbitrary student with a standard IQ-test: $S = [50{,}200]$.
- Observe the duration (in seconds) of a telephone conversation: $S = [0\,, \infty)$.
- Observe the change in value (in %) of a stock fund in one year: $S = \mathbb{R}$.
- Use a (pseudo) random number generator to produce a random number between 0 and 1: $S = [0{,}1]$.
- Determine the covered distance by a midfielder of Ajax during a Champions League game: $S = [0\,, \infty)$.

Of course, in these experiments the precision of measurements plays a role: a telephone call might be measured in an integer number of seconds or in thousands of seconds, etc. But if we state a probability model we usually do not care about the precision of measurement: we are trying to model the physical reality without (changeable) restrictions.

For continuous random variables, such as in the examples above, in general, we have $\boldsymbol{P(X = x) = 0}$ for each real value x.

Example 6.1.1 Age, measured in years, is a time variable that is continuous, with a continuous range $[0, 130]$ of positive values. Nevertheless, our age is usually given by an integer number: if your age is 20 and you wonder what the probability is that an arbitrary Dutch citizen has the same age, then we could find a positive probability:
$P(X = 20) = 1.75\%$, but in this case we have a discrete distribution of integer valued ages:

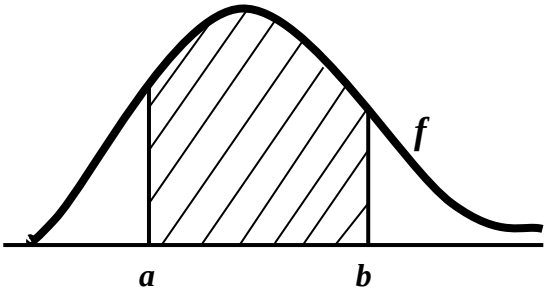
$$S_X = \{0, 1, 2, 3, \ldots\ldots, 130\}\,.$$

If X is modelled as a continuous variable, the event {an arbitrary person has age 20} would be given by an interval of values, $20 \leq X < 21$.
Clearly, choosing the interval of time smaller will reduce the probability. Considering the probabilities that an arbitrary Dutch citizen is born in the same year, the same month, the same day, the same minute, the same millisecond, as you are, converges to "approximately zero". ■

For continuous variables we cannot define the probability model by a probability function: we need another kind of model where probabilities that the variable X attains values in an interval of real values lead to positive values.
Such a model is given by the **(probability) density function** (abbreviated by "density" or "**pdf**") of a continuous random variable:

Definition 6.1.2 The **density function** of a continuous random variable X is a

$$\text{non} - \text{negative function } \boldsymbol{f}, \text{such that } \boldsymbol{P(a \le X \le b) = \int_a^b f(x)dx}$$



According the Fundamental theorem of Calculus we can express this integral in the anti-derivative $F(x)$ of $f(x)$:

$$\boldsymbol{P(a \le X \le b) = \int_a^b f(x)dx = F(x)|_{x=a}^{x=b} = F(b) - F(a)}$$

In this definition probabilities are area's: the probability that X takes on values between a and b is equal to the area below the graph of density function f above the interval $[a, b]$. Whether we choose an open interval (a, b) or a closed interval $[a, b]$ does not affect the area nor the probability:

$$P(a \le X \le b) = P(a < X < b)$$

For a continuous variable X the area of the line above $X = a$ on the X-axis below f is 0:

$$P(X = a) = \int_a^a f(x)dx = F(a) - F(a) = 0$$

It is clear that $f(x)$ should not be negative on an interval, since a negative probability must be excluded, and that the total area under the graph of f should be 1 (the total probability is 100%). Reversely, a function $f(x)$, that meets these two conditions, can be considered to be a density function.

Property 6.1.3 f is a density function if **a.** $\boldsymbol{f(x) \ge 0}$ and

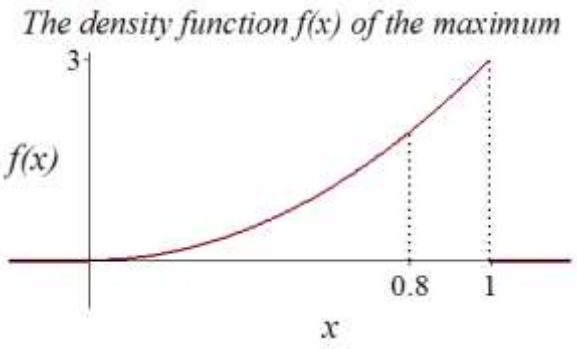
b. $\boldsymbol{\int_{-\infty}^{\infty} f(x)dx = 1}$

Formally f could be undefined (or even negatively defined) on a finite number of values x, since this does not affect the computation of areas, being probabilities.

Example 6.1.4

Use the random number button on your calculator to generate three of these numbers. If we choose the largest of three numbers, we have defined a random variable X in this stochastic

experiment. This maximum X is a continuous random variable with a range $S_X = [0, 1]$.
The density function of X can be derived (as we will see later in exercise 7.2 in the next chapter):



$$f(x) = \begin{cases} 3x^2 & \text{if } 0 \le x \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

This defines a density function, since the conditions of property 6.1.3 are fulfilled:

$$1.\ f(x) \ge 0 \quad \text{and} \quad 2.\ \int_{-\infty}^{\infty} f(x)dx = \int_{0}^{1} 3x^2 dx = [x^3]_{x=0}^{x=1} = 1$$

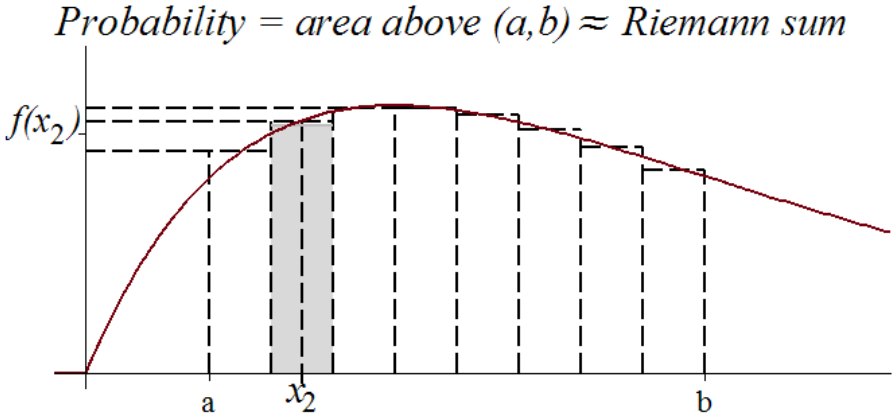
The probability that the maximum of 3 random numbers is larger than 0.8 can be computed:

$$P(X > 0.8) = \int_{0.8}^{1} 3x^2 dx = [x^3]_{x=0.8}^{x=1} = 1^3 - 0.8^3 = 48.8\%$$ ■

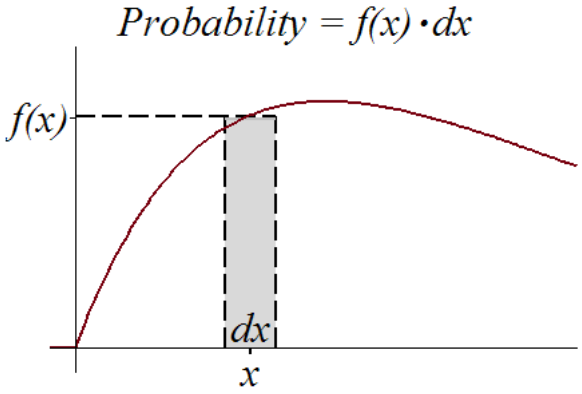
Example 6.1.4 illustrates that $f(x)$ is **not a probability** but a **probability *density***:
$f(x)$ can attain values larger than 1, but if we consider equally large intervals, such as [0, 0.2] and [0.8, 1.0], we conclude that the probability that X takes on a value from the last interval is larger than from the first, since the density function on [0.8, 1.0] is larger:

$$P(0.8 \le X \le 1) > P(0 \le X \le 0.2)$$

The concept of density functions and interpretation of probabilities as areas are supported by the definition of an integral as a "limit of a **Riemann sum** $\sum f(x)\Delta x$", which is illustrated in the graph below:
we can split the interval $[a, b]$ into n small intervals with equal widths $\Delta x = \frac{b-a}{n}$
(sometimes "small" dx is used instead of Δx).
The intervals $[a, a + \Delta x), [a + \Delta x, a + 2 \cdot \Delta x), \ldots, [a + (n-1) \cdot \Delta x, b]$ have midpoints $x_1, x_2, \ldots, x_n$.



The probability that X attains a value in the i^{th} interval ($i = 2$ in the graph), equals approximately the area of the rectangle: area = length×width = $f(x_i) \times \frac{b-a}{n} = f(x_i) \times \Delta x$.
If the limit of the Riemann sum $\sum_{i=1}^{n} f(x_i) \times \frac{b-a}{n}$ exists, as n approaches infinity ($n \to \infty$), we denote the limit as the integral $\int_a^b f(x)dx$.



For a (very) small interval with interval width $dx = \frac{b-a}{n}$ we have:

$$P\left(x - \tfrac{1}{2}dx \le X \le x + \tfrac{1}{2}dx\right) \approx f(x)dx$$

Using the analogy of discrete and continuous distributions it is not difficult to see that definitions and properties show similarities:

$$\sum_{a \le x \le b} P(X = x) \quad \text{corresponds with} \quad \int_a^b f(x)dx$$

$$\sum_{x \in S_X} P(X = x) = 1 \quad \text{corresponds with} \quad \int_{-\infty}^{\infty} f(x)dx = 1$$

and: $E(X) = \sum_{x \in S_X} x \cdot P(X = x)$ corresponds with $\boldsymbol{E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx}$

In the last expression we find that for continuous variables we can interpret $E(X)$ (again) as the weighted average of the x-values with weighing factor the "probability" $f(x)dx$.
The summation is replaced by an integral.

Definition 6.1.5 The **expectation (expected value)** of a continuous random variable X is

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

provided that this integral is absolute convergent: $\int_{-\infty}^{\infty} |x| \cdot f(x)dx < \infty$.

As before, we use notations EX, μ and μ_X for expectations as well.
The label (index) X (or Y or Z) in μ_X will be used if we want to avoid confusion.
For the same reason we will use labeled density functions, such as $f_X(x)$ and f_Z, whenever this is appropriate.

Example 6.1.6 In example 6.1.4 we gave the density function of X, the maximum of three random numbers:

$f(x) = 3x^2$, if $0 \le x \le 1$ and $f(x) = 0$ for all other values of X

What is the expectation and the variance of this maximum?

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_{0}^{1} x \cdot 3x^2 dx = \left[\tfrac{3}{4}x^4\right]_{x=0}^{x=1} = \frac{3}{4}$$

The variance is defined in the same way $var(X) = E(X - \mu)^2$ as in the discrete case and the computational formula as well:

$$var(X) = E(X^2) - (EX)^2$$

In this formula $E(X^2)$ is "the weighted average of the squares of the values X":

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx = \int_{0}^{1} x^2 \cdot 3x^2 dx = \left[\tfrac{3}{5}x^5\right]_{x=0}^{x=1} = \frac{3}{5}$$

So: $var(X) = E(X^2) - (EX)^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{3}{80}$ ■

Property 6.1.7 For every real valued function g we have:

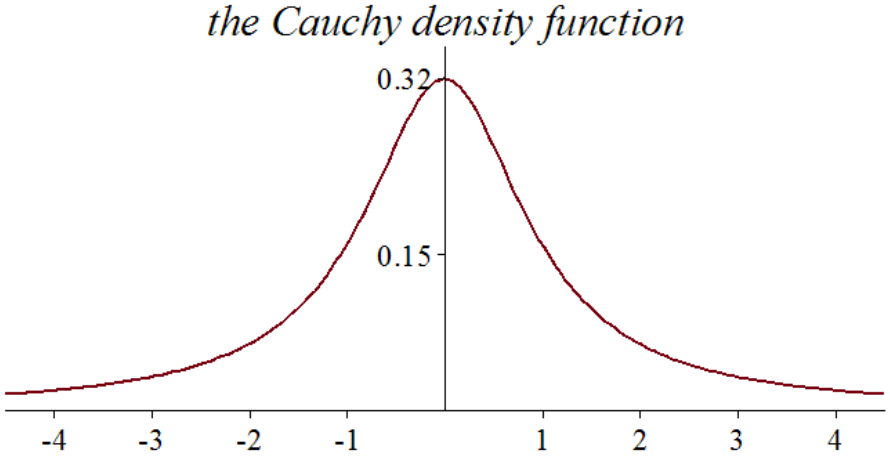
$$\boldsymbol{Eg(X) = \int_{-\infty}^{\infty} g(x)f(x)dx}$$

The computation of the variance can be executed in two ways, both applying property 6.1.7:

- Direct, with the definition: $\boldsymbol{var(X) = E(X - \mu)^2} = \int_{-\infty}^{\infty}(x-\mu)^2 \cdot f(x)dx$
- Usually we will preferably apply $\boldsymbol{var(X) = E(X^2) - \mu^2}$ for computational simplicity, so then we will apply $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$.

Example 6.1.8 The **Cauchy distribution** is defined by its density function

$$\boldsymbol{f(x) = \frac{1}{\pi(1 + x^2)}}, \qquad \text{for } x \in \mathbb{R}$$



This is indeed a density function since: 1) $f(x) \geq 0$ and

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

The last equality follows from the fact that $\frac{1}{1+x^2}$ is the derivative of $arctan(x)$, the inverse of the tangent function:

$$\int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)}dx = \frac{1}{\pi}\cdot[\arctan(x)]_{x\to-\infty}^{x\to\infty} = \frac{1}{\pi}\cdot\left[\frac{1}{2}\pi-\left(-\frac{1}{2}\pi\right)\right] = 1$$

Since the density function is symmetric one would guess that $E(X) = 0$, but this is not the case here! EX does not exist: the integral is not (absolute) convergent:

$$\int_{-\infty}^{\infty} |x|\cdot\frac{1}{\pi(1+x^2)}dx = 2\int_{0}^{\infty} x\cdot\frac{1}{\pi(1+x^2)}dx \text{ (since the function is even)}$$
$$= \left[\frac{1}{\pi}\cdot\ln(1+x^2)\right]_{x=0}^{x\to\infty} = \infty$$

Because $\mu = EX$ does not exist, $var(X) = E(X-\mu)^2$ does not exist either. ■

6.2 Distribution function

Example 6.2.1 A large insurance company in Apeldoorn, Holland, wants a probability model of the duration X (in seconds) of telephone calls by customers. A discrete model based on the rounded number of seconds seems inappropriate, since many outcomes are possible. Furthermore, from observations it seems clear that the proportion of telephone calls, longer than t seconds, decreases exponentially as t increases:

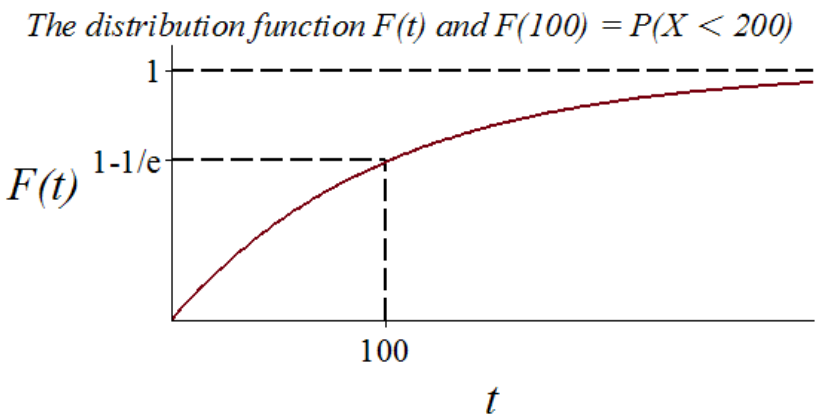
$$P(X > t) = e^{-0.01t}, \quad \text{for } t \geq 0.$$

This relation between the proportion of calls and the time t is fitting well to the data and can be used to determine many different probabilities:

- $P(X > 100) = e^{-1} \approx 36.8\%$,
- $P(X > 0) = e^{0} = 1$,
- $P(100 < X \leq 200) = P(X > 100) - P(X > 200) = e^{-1} - e^{-2} \approx 23.3\%$ and
- According the complement rule: $P(X \leq t) = 1 - P(X > t) = 1 - e^{-0.01t}$ $(t \geq 0)$.

The last probability is a function of t, that is $F: t \to P(X \leq t)$, and defines the **distribution function** of X. The frequency interpretation tells us that $F(t)$ is (approximately) the proportion of calls of t seconds and less if we consider many of these calls.

$$F(t) = P(X \leq t) = 1 - P(X > t) = \begin{cases} 1 - e^{-0.01t}, & \text{if } t \geq 0 \\ 0, & \text{if } t < 0 \end{cases}$$



We can express probabilities of the event "X in $[a,b]$" in F, such as:

$$P(\mathrm{X} \leq 100) = \mathrm{F}(100) = 1 - \mathrm{e}^{-\frac{1}{100}\times 100} = 1 - \frac{1}{\mathrm{e}} \approx 63.2\%$$

And
$$\begin{aligned} P(100 < X \leq 200) &= P(X \leq 200) - P(X \leq 100). \\ &= F(200) - F(100) \\ &= (1 - e^{-2}) - (1 - e^{-1}) \approx 23.3\% \end{aligned}$$ ■

For every random variable X, defined on a probability space (S, P), we could derive such a distribution function F.

> **Definition 6.2.2** The function F, defined by $F(x) = P(X \leq x)$ with $x \in \mathbb{R}$, is the (cumulative) **distribution function (cdf)** of the random variable X.

If there are two or more random variables involved, we can "label" the distribution functions: $F_X(x)$ and $F_Y(y)$ are the **marginal distribution functions** of the random variables X and Y.

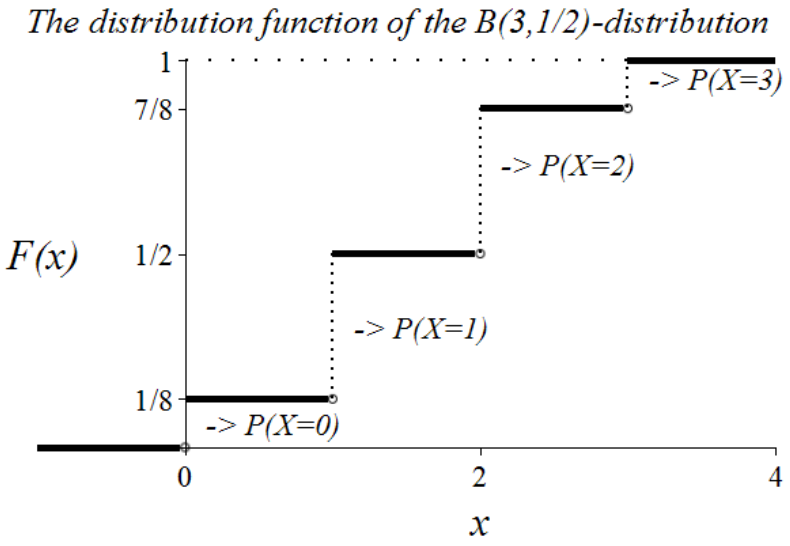
Example 6.2.3 In example 4.2.1 X had a $B\left(3, \frac{1}{2}\right)$-distribution.
The distribution function can be found for such a discrete variable as well, e.g., by :

$$F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}$$

and for non-integer values of x:

$$F(1.7) = P(X \leq 1.7) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}$$

So for any $x \in [1, 2)$ we find: $F(x) = P(X \leq x) = P(X \leq 1) = \frac{1}{2}$.



In the points x of discontinuity the "jump height" of the graph equals $P(X = x)$. ■

We noticed that the cumulative binomial tables actually give the values of the distribution function: suppose X, has a $B\left(20, \frac{3}{10}\right)$-distribution, then $F_X(4.5)$ can be determined with the $B\left(20, \frac{3}{10}\right)$-table:

$$F(4.5) = P(X \leq 4.5) = P(X \leq 4) = \sum_{k=0}^{4} P(X = k) \approx 23.75\%$$

The graphs of the distribution functions in examples 6.2.1 and 6.2.3 show non-decreasing functions $F(x)$, that for small x-values start at the value 0 and for large x-values end up at the maximum 1, possibly in the limit for x approaching infinity. The graph of F is continuous, for the continuous variable and has the shape of a step function for discrete variables. But this function is "continuous from the right" in every value of x. These properties can be proven in general (using Kolmogorov's axioms, since $F(x)$ is a probability), but we just state them:

Property 6.2.4 (Sufficient and necessary properties for a distribution function)
For any distribution function $F(x)$ of a random variable X, we have:

a. F is non-decreasing (if $x_2 > x_1$, then $F(x_2) \geq F(x_1)$).
b. $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$
c. F is continuous from the right ($\lim_{h\to0^+} F(x+h) = F(x)$)

So, any function that satisfies these three properties is a distribution function. The following properties can be derived from them. They are of interest when computing probabilities for a known distribution function.

Property 6.2.5 For a distribution function F of a random variable X we have:

a. $P(a < X \leq b) = F(b) - F(a)$.
b. $P(X > x) = 1 - F(x)$.
c. $P(X < x) = \lim_{h\to0^+} F(x-h)$
d. $P(X = x) = F(x) - P(X < x)$

Proof:
a. $\{X \leq a\}$ and $\{a < X \leq b\}$ constitute a partition of $\{X \leq b\}$.
Then, according Kolmogorov`s axiom (3): $P(X \leq b) = P(X \leq a) + P(a < X \leq b)$
Or $F(b) = F(a) + P(a < X \leq b)$.
b. Follows from the complement rule (chapter 1) where $A = \{X > x\}$.
c. will not be formally proven: $\lim_{h\to0^+}$ means "limit of the function if h **decreases to** 0" ($\lim_{h\downarrow0}$).
d. Consequence of $\{X \leq x\} = \{X < x\} \cup \{X = x\}$. ■

In example 6.2.3 the distribution function of the discrete random variable at hand was a so called "step function": the steps are the probabilities, summing to a total of 1. In example 6.2.1 the distribution function was continuous for all values of X. Then, according to property 6.2.5d, we have: $P(X = x) = 0$ for every $x \in \mathbb{R}$.

Definition 6.2.6 A random variable X is **continuous** if the distribution function F of X is a continuous function.

Of course there are more types of random variables, e.g. a mix of discrete and continuous random variables: a waiting time could be exponentially distributed with probability $\frac{1}{3}$ and Poisson distributed with probability $\frac{2}{3}$. These distributions are outside the scope of this course.

The relation between density and distribution function is easily found from the definition of distribution function:

$$\boldsymbol{F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du}$$

It follows from the fundamental theorem of calculus that $F(x)$ is an anti-derivative of the density function $f(x)$. Note that $F(x)$ is a **specific anti-derivative**: in mathematics we would add a constant to find all antiderivatives: $F(x) + c$. In probability theory we will, however, only use $F(x)$ as defined above, since it should be a probability.
The fundamental theorem of calculus asserts:

$$\boldsymbol{F'(x) = f(x)}$$

Or, considering a (small) interval $\left(x - \frac{1}{2}dx, x + \frac{1}{2}dx\right)$, with length dx (see the graph on page 6-4), and applying the properties of $F(x)$, we find:

$$P\left(x - \tfrac{1}{2}dx < X \leq x + \frac{1}{2}dx\right) = F\left(x + \frac{1}{2}dx\right) - F\left(x - \frac{1}{2}dx\right) \approx F'(x)dx = f(x)dx$$

So, the probability that X lies in an interval with length dx around x, can be approximated by the derivative of the distribution function times dx (assuming that the derivative exists).
From this we can conclude that the distribution of the continuous random variable X (all possible probabilities w.r.t. X) is given by either the density function f or the distribution function F. One follows from the other.

Example 6.2.7 In example 6.2.1 we used the following distribution function of the duration X of a telephone call:

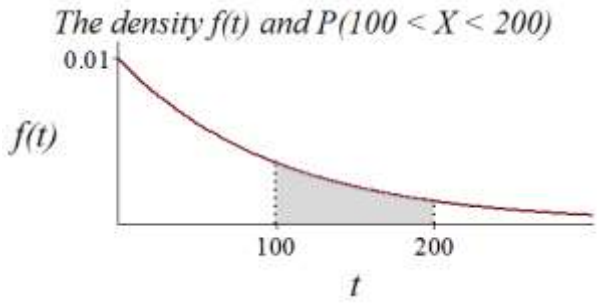
$$F(t) = \begin{cases} 1 - e^{-0.01t} & \text{voor } t \geq 0 \\ 0 & \text{voor } t < 0 \end{cases}$$

We computed the probability:

$$P(100 < X \leq 200) = F(200) - F(100) = (1 - e^{-2}) - (1 - e^{-1}) \approx 23.3\%$$

The density function of X is:

$$f(t) = \frac{d}{dt}F(t) = \begin{cases} 0.01e^{-0.01t} & \text{voor } t \geq 0 \\ 0 & \text{voor } t < 0 \end{cases}$$



This function is called the **exponential density function** with **parameter 0.01**.

The probability $P(100 < X \leq 200)$ could also be computed using $f(x)$:

$$P(100 < X \leq 200) = \int_{100}^{200} 0.01e^{-0.01t}dt = -e^{-0.01t}|_{t=100}^{t=200} = -e^{-2} + e^{-1} \approx 23.3\% \quad \blacksquare$$

For discrete X the distribution can be given by either the probability function $P(X = x)$ or the distribution function $F(x) = P(X \leq x)$, but usually we will prefer to use the probability function.

Property 6.2.8 (Properties of continuous distributions)

For a continuous random variable X with density function f and (cumulative) distribution function F we have:

- **a.** $P(X = x) = 0$, for $x \in \mathbb{R}$.
- **b.** $P(X \in [a, b]) = \int_a^b f(x)dx = F(b) - F(a)$
  The closed interval $[a, b]$ can be replaced by an open interval (a, b).
- **c.** $F(x) = \int_{-\infty}^{x} f(u)du$
- **d.** $f(x) = \frac{d}{dx}F(x)$
- **e.** If the density function $f(x)$ of X is symmetric about $x = c$, then $E(X) = c$. (provided that $E(X)$ exists).

6.3 The uniform, exponential and standard normal distributions

The uniform distribution on the interval $[a, b]$

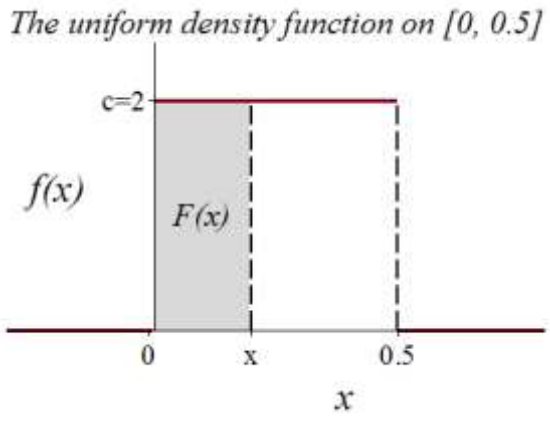
Example 6.3.1 Consider the situation of measurement of a quantity (time in seconds, weight in grams) in integer units. The real value 387.84 is observed as 388 (measurement error 0.16) and 238.435 as 238 (measurement error 0.435). We assume that observation of the value means "rounding to the nearest integer".
If the measurement error X is the (absolute) difference of the real and the observed value, then $S_X = [0, 0.5\,]$. For the density function f of X we should define $f(x) = 0$ for $x \notin [0, 0.5\,]$
Within the interval any value is "equally likely" and because X is a continuous random variable, $f(x)dx$, for fixed interval length dx, should be the same, everywhere within $[0, 0.5]$.
Conclusion: f is constant on $[0, 0.5\,]$, $f(x) = c$. The total area has to be 1 (property 6.1.3):

$$\int_{-\infty}^{\infty} f(x)dx = 1 = \text{the area of the rectangle}$$

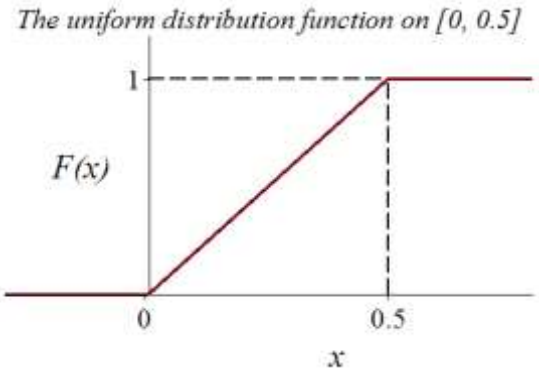
$$= 0.5 \cdot c, \text{so } c = 2$$

$$f(x) = \begin{cases} 2, & \text{if } x \in [0, 0.5] \\ 0, & \text{if } x \notin [0, 0.5] \end{cases}$$

X is said to have a **uniform density function** on the **interval** $\left[\mathbf{0}, \frac{\mathbf{1}}{\mathbf{2}}\right]$.



The distribution function $F(x) = P(X \leq x)$ can easily be derived from the graph.
If $x \in \left[0, \frac{1}{2}\right]$, this probability is the area in the graph: $F(x) = x \cdot 2 = 2x$, etc.



So

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 2x, & \text{if } 0 \leq x \leq 0.5 \\ 1, & \text{if } x > 0.5 \end{cases}$$

Differentiation leads indeed to the density function: $f(x) = \frac{d}{dx}F(x) = 2$ if $0 \leq x \leq 0.5$.

The "mean" measurement error is intuitively equal to $\frac{1}{4}$ (using the symmetry of f), the midpoint of the interval $[0, 0.5\,]$. This can be verified, using the definition of $E(X)$:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_{0}^{\frac{1}{2}} x \cdot 2dx = [x^2]_{x=0}^{x=1/2} = \frac{1}{4}$$

Similarly, we have

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx = \int_{0}^{1/2} x^2 \cdot 2dx = \left[\tfrac{2}{3}x^3\right]_{x=0}^{x=1/2} = \frac{1}{12}$$

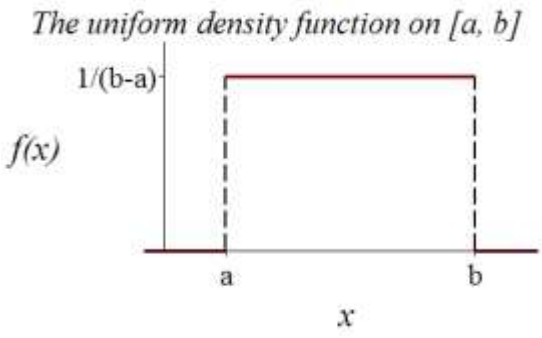
So: $var(X) = E(X^2) - (EX)^2 = \frac{1}{12} - \left(\frac{1}{4}\right)^2 = \frac{1}{48}$ ■

Apart from these measurement errors the uniform distribution is widely used as a model of random numbers taken from a given interval. In general an interval of shape $[a, b]$:

Definition 6.3.2 The random variable X has a **uniform distribution on the interval $[\boldsymbol{a}, \boldsymbol{b}]$,** if

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{for } x \in [a, b] \\ 0, & \text{for } x \notin [a, b] \end{cases}$$

Short notation: $\boldsymbol{X \sim U(a, b)}$



Sometimes an open interval (a, b) is chosen instead of a closed one.
It is easily seen that this $f(x)$ in this definition 6.3.2 is a density function: the area under the (non-negative density function) is a rectangle with area $\frac{1}{b-a} \cdot (b-a) = 1$.
The expected value ("the approximate average value of many random numbers from the interval $[a, b]$") is, using the line of symmetry: $\frac{a+b}{2}$, the midpoint of the interval.

Property 6.3.3 The expectation and variance of the uniform distribution on $[a, b]$ are:

a. $\boldsymbol{E(X) = \frac{a+b}{2}}$

b. $\boldsymbol{var(X) = \frac{(b-a)^2}{12}}$

Proof:

a. According to property 6.2.8.e:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_a^b x \cdot \frac{1}{b-a}dx = \left[\frac{1}{2} \cdot \frac{x^2}{b-a}\right]_{x=a}^{x=b} = \frac{1}{2} \cdot \frac{b^2-a^2}{b-a} = \frac{b+a}{2}$$

b. $E(X^2) = \int_a^b x^2 \cdot \frac{1}{b-a}dx = \left[\frac{1}{3} \cdot \frac{x^3}{b-a}\right]_{x=a}^{x=b} = \frac{1}{3} \cdot \frac{b^3-a^3}{b-a} = \frac{b^2+ab+a^2}{3}$, so:

$$var(X) = E(X^2) - (EX)^2 = \frac{b^2+ab+a^2}{3} - \left(\frac{b+a}{2}\right)^2 = \frac{4b^2+4ab+4a^2-(3b^2+6ab+3a^2)}{12} = \frac{(b-a)^2}{12}$$

■

The uniform distribution on [0,1], or the $U(0,1)$-distribution, has many applications in technical sciences: it offers a simple model of random numbers between 0 and 1. They are easily generated, e.g. with random number button ($ran\#$) on your calculator or similar formulas in Excel.

These "pseudo random generators" use special functions for generating numbers which are quite unpredictable. Once we have such a random number, we can easily generate "arbitrary observations" from a distribution we would like to simulate. E.g. normally distributed yearly returns of a company's market share or exponentially distributed service times. In the latter case we can use the simulated service times to see how a system, that handles the requested services, is performing, before implementing the system in reality.

The exponential distribution with parameter λ

In examples 6.2.1 and 6.2.7 we encountered situations where the exponential distribution gives a valid model of the real life situations. For the duration X of a telephone call we found:

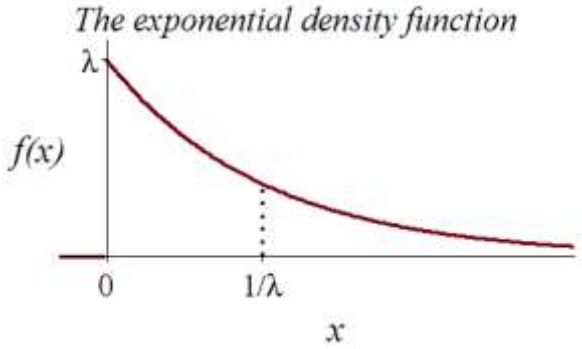
- $f(t) = 0.01e^{-0.01t}$ (als $t \geq 0$) and
- $P(X > t) = e^{-0.01t}$ is the **survival rate**: the probability that the duration of the call exceeds t seconds decreases exponentially for increasing t.
- $F(t) = P(X \leq t) = 1 - e^{-0.01t}$, for $t \geq 0$ (a duration is always non-negative).

Such a distribution is called an **exponential distribution, with parameter λ** (= 0.01 here).

Definition 6.3.4 The random variable X has an **exponential distribution with parameter λ** (> 0) if

$$\boldsymbol{f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{voor } x \geq 0 \\ 0 & \text{voor } x < 0 \end{cases}}$$

Brief notation: $\boldsymbol{X \sim Exp(\lambda)}$



Property 6.3.5 If X is an exponentially distributed variable, with parameter λ, then:

a. $\boldsymbol{P(X > x) = e^{-\lambda x}}$, for $x \geq 0$

b. $\boldsymbol{F(x)} = \begin{cases} \boldsymbol{1 - e^{-\lambda x}} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$

c. $\boldsymbol{E(X) = \frac{1}{\lambda}}$

d. $\boldsymbol{var(X) = \frac{1}{\lambda^2}}$

Proof:

a. $P(X > x) = \int_x^{\infty} \lambda e^{-\lambda t} dt = -e^{-\lambda t}\Big|_{t=x}^{t\to\infty} = 0 - \left(-e^{-\lambda x}\right) = e^{-\lambda x}$, if $x \geq 0$
if $x < 0$, then $P(X > x) = 1$

b. The cdf follows directly from a.: $F(x) = P(X \leq x) = 1 - P(X > x)$

c. To determine the expectation we will use the mathematical technique of "integration by parts" (see appendix):

$$E(X) = \int_0^{\infty} x \cdot \lambda e^{-\lambda x} dx = x \cdot -e^{-\lambda x}\Big|_{x=0}^{x\to\infty} - \int_0^{\infty} 1 \cdot -e^{-\lambda x} dx = 0 + -\frac{1}{\lambda} e^{-\lambda x}\Big|_{x=0}^{x\to\infty} = \frac{1}{\lambda}$$

d. Similar to c. we find: $E(X^2) = \int_0^{\infty} x^2 \cdot \lambda e^{-\lambda x} dx = \ldots = \frac{2}{\lambda^2}$

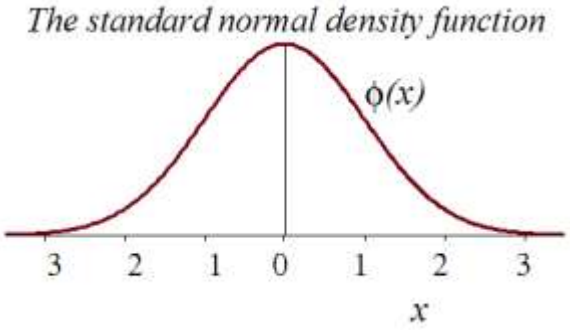
So $var(X) = E(X^2) - (EX)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$ ■

The exponential distribution can often be applied as a real life probability model for waiting times, service times and interarrival times (of clients). Furthermore the exponential distribution is sometimes an adequate model of lifetimes, where "dying" is not caused by wearing out (aging), but by coincidental external causes.

The standard normal distribution

The standard normal distribution is a special case of the (general) normal distribution in section 5 of this chapter. Its central position in probability theory and statistics justifies a special notation of the variable and its distribution:

- A standard normal random variable is indicated by $\boldsymbol{Z}$, unless we explicitly use another notation.
- The standard normal density function will be denoted as $\boldsymbol{\varphi(z)}$, so $\varphi = f_Z$.
- The standard normal distribution function is $\boldsymbol{\Phi(z)}$, so $\Phi(z) = P(Z \leq z)$ or $F_Z(z)$. (φ and Φ, pronounced as "phi", are the Greek versions of f and F.)



Definition 6.3.6 The continuous random variable Z has a **standard normal distribution** if

$$\boldsymbol{\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}}, \quad \text{where } \boldsymbol{z} \in \mathbb{R}$$

The factor $\frac{1}{\sqrt{2\pi}}$ is necessary to make φ a density function (total area 1).

To show that $\varphi(z)$ is a density function, that is $\int_{-\infty}^{\infty} \varphi(z)dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 1$, one could transfer to polar coordinates, but this proof is not part of this course.
Because of the symmetry of the graph of φ one is inclined to state that $E(Z) = 0$, if $E(Z)$ exists. Actual computation confirms this:

$$E(Z) = \int_{-\infty}^{\infty} z \cdot \varphi(z)dz = \int_{-\infty}^{\infty} z \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \Big|_{z\to-\infty}^{z\to\infty} = 0$$

If we want to use the formula $var(Z) = E(Z^2) - (EZ)^2$ to compute the variance we need the value of $E(Z^2)$, the weighted average of the values of z^2, weighed with "probability" $\varphi(z)dz$. Applying integration by parts we find:

$$E(Z^2) = \int_{-\infty}^{\infty} z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = z \cdot -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \Big|_{z\to-\infty}^{z\to\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 0 + 1$$

Note that the last integral is the total area of the standard normal density function φ:

$$\int_{-\infty}^{\infty} \varphi(z)dz = 1$$

So $var(Z) = E(Z^2) - (EZ)^2 = 1 - 0^2 = 1$
Since $\mu = 0$ and $\sigma^2 = 1$, the following short notation of the standard normal distribution is used:

$$\boldsymbol{Z \sim N(0, 1)}$$

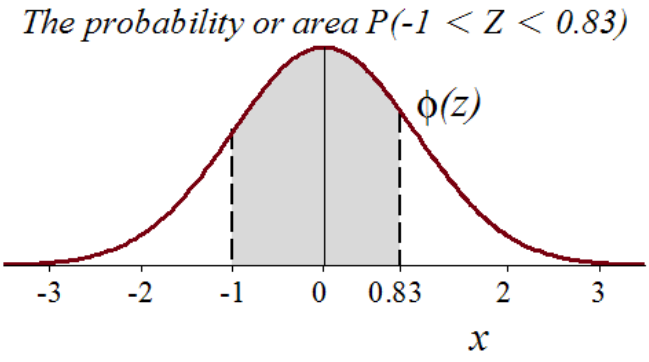
Example 6.3.7 Probabilities such as $P(-1 \le Z \le 0.83)$ could be given by an integral:

$$P(-1 \le Z \le 0.83) = \int_{-1}^{0.83} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \text{???}$$

The anti-derivative of $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ does not exist! Therefore we will have to determine a numerical approximations of these integrals (probabilities): a Riemann sum on the interval.
To avoid these time consuming operation over and over again the values of the standard normal distribution function $\Phi(z) = P(Z \le z)$ can be found in the **standard normal table** of the *cdf* (see the Tab-6 page at the end of the reader), for **positive *z*, given in two decimals.**

Without integrals or numerical approximation, we can find, simply using the table:

$$\begin{aligned} &P(-1 \le Z \le 0.83) \\ &\quad = P(Z \le 0.83) - P(Z \le -1.00) \\ &\quad = \Phi(0.83) - \Phi(-1.00) \\ &\quad = 0.7967 - 0.1587 \\ &\quad = 63.80\% \end{aligned}$$

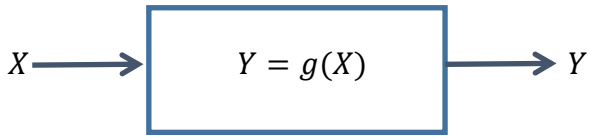


The second probability $P(Z \leq -1.00) = \Phi(-1.00)$ is determined by using the symmetry of the graph of φ about the line $x = 0$:

$$P(Z \leq -1.00) = P(Z \geq 1.00) = 1 - \Phi(1.00) = 1 - 0.8413 = 0.1587 \qquad \blacksquare$$

6.4 Functions of a continuous random variable

Suppose X is a random **input signal** of a device**,** e.g. X could be a voltage. The value of X on a particular moment is unknown, but it can be observed as a realization of a specified distribution**.**
The device transforms the signal to an output signal Y, according a transformation function g:

Sketch: 

In the example of a voltage the device could be an "amplifier", such that $Y = g(X) = a \cdot X$ $(a > 0)$ or a "rectifier" $Y = g(X) = |X|$.
There does not seem to be a direct method to express the density function of Y in the (known) density of X. E.g., $f_Y(y) = a \cdot f_X(y)$, $a \neq 1$, is not a density function and $f_Y(y) = f_X\left(\frac{y}{a}\right)$ neither. But it is possible to **express probabilities of events with respect to Y in probabilities of events with respect to X**.

Example 6.4.1 For the rectifier $Y = |X|$ we have:

$$P(Y \leq 3) = P(|X| \leq 3) = P(-3 \leq X \leq 3) = F_X(3) - F_X(-3),$$

where F_X is the known distribution function of X and $P(Y \leq 3)$ is the same as $F_Y(3)$.
Generalizing this idea we will find for the unknown distribution function $F_Y(y) = P(Y \leq y)$:

- If $y < 0$, then the event $\{Y \leq y\} = \{|X| \leq y\} = \emptyset$, so: $F_Y(y) = 0$ if $y < 0$.
- If $y \geq 0$, then: $F_Y(y) = P(|X| \leq y) = P(-y \leq X \leq y) = F_X(y) - F_X(-y)$ (applying property 6.2.8b.)

The density function of Y can now be found by differentiating the distribution function F_Y (do not forget to apply the chain rule for $f_X(-y)$):

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ f_X(y) + f_X(-y) & \text{if } y \geq 0 \end{cases}$$

Therefore we can use the known density function of X to determine the density function of Y: we determined the distribution of Y and if we want we can compute $E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy$.■

In the example above we found an approach of the problem that "worked": if we know the distribution of X and we want to determine the distribution (density function) of $Y = g(X)$, we can first express the distribution function $F_Y(y)$ in the distribution function $F_X(x)$ of X. Consequently we can use the derivative of this equality to express the density function of Y in the distribution of X.

Example 6.4.2. Z has a standard normal distribution and $g(Z) = Z^2$ the transformation function: a. What is the density function of the "output signal" $Y = Z^2$?

b. Compute $EY = E(Z^2)$.

In this case we know: $f_Z(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$, where $z \in \mathbb{R}$

Let us start with the b. part, since we have encountered this expectation in the introduction of the standard normal distribution in section 6.3:

$$E(Y) = E(Z^2) = \int_{-\infty}^{\infty} z^2 \varphi(z) dz = \cdots = 1$$

To answer the a. part we have to determine the distribution of $Y = Z^2$:

1. First express $F_Y(y)$ in F_Z $(= \Phi$, in this case):

$$F_Y(y) = P(Y \le y) = P(Z^2 \le y) = \begin{cases} 0 & \text{if } y \le 0 \\ P\left(-\sqrt{y} \le Z \le \sqrt{y}\right) & \text{if } y > 0 \end{cases}$$

So, if $y > 0$: $F_Y(y) = \Phi\left(\sqrt{y}\right) - \Phi\left(-\sqrt{y}\right)$

2. Differentiate $F_Y(y)$ to express $f_Y(y)$ in the known f_Z $(= \varphi)$.

If $y \le 0$, then $f_Y(y) = \frac{d}{dy} F_Y(y) = 0$

If $y > 0$, then $f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}\left[\Phi\left(\sqrt{y}\right) - \Phi\left(-\sqrt{y}\right)\right]$

$$= \frac{1}{2\sqrt{y}}\left[\varphi\left(\sqrt{y}\right) + \varphi\left(-\sqrt{y}\right)\right]$$

3. Use the known density function of Z to find the formula for $f_Y(y)$.

$f_Y(y) = 0$, for $y \le 0$ and

$$f_Y(y) = \frac{1}{2\sqrt{y}}\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sqrt{y})^2} + \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sqrt{y})^2}\right] = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}, \qquad \text{for } y > 0.$$

(*note that the last expression does not exist if $y = 0$: that is why we (arbitrarily) defined for $y = 0$ that $f_Y(y) = 0$. However, we could have left it undefined.*)

We performed the computation of $E(Y) = E(Z^2)$ in the b. part by computing $E(Z^2) = 1$, but now we can compute $E(Y)$ "directly", using the density of Y:

$$E(Y) = \int_{-\infty}^{\infty} y \cdot f_Y(y) dz = \int_0^{\infty} y \cdot \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y} dy = \int_0^{\infty} \frac{\sqrt{y}}{\sqrt{2\pi}} e^{-\frac{1}{2}y} dy$$

This is not an elementary integral, we need to apply a substitution: try $y = z^2$, so $dy = 2zdz$

$$E(Y) = \int_0^{\infty} \frac{z}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \cdot 2z dz = \int_0^{\infty} 2z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

We recognize the integral $2\int_0^{\infty} z^2 \cdot \varphi(z) dz$: because of the symmetry of the (even) function $z^2 \cdot \varphi(z)$ equals $\int_{-\infty}^{\infty} z^2 \cdot \varphi(z) dz = E(Z^2)$, so indeed $E(Y) = E(Z^2) = 1$ ■

In example 6.4.2 we determined the distribution of Z^2, the square of a standard normal random variable: $f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}$ $(y > 0)$.

This distribution is the **Chi-square distribution with one degree of freedom**.
Chi-square distributions play an important role in statistics**.**

Furthermore we can generalize the applied method in this example:
if $Y = g(X)$, where X has a specified distribution (we know the density function), then the distribution of Y can be derived from the distribution of X in a 3 steps approach:

1. First express $F_Y(y)$ in F_X. $(F_Y(y) = P(Y \le y) = P(g(X) \le y) = \cdots)$
2. Then compute the derivative to express $f_Y(y)$ in f_X. $\left(f_Y(y) = \frac{d}{dy}F_Y(y)\right)$
3. Finally, use the specified density function f_X to determine the formula for $f_Y(y)$.

This approach enables us to determine $E(Y)$ in two ways:

- Using the distribution of Y: $E(Y) = \int_{-\infty}^{\infty} y \cdot f_Y(y)dy$ or
- Using the distribution of X: $E(Y) = Eg(X) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx$

Of course we will choose the method which is the simplest analytically the simplest.

We will illustrate the approach above for a linear relation, $Y = aX + b$, where a and b are real valued constants. We proved before that the expected value is $E(Y) = E(aX + b) = aE(X) + b$, but what is the distribution of Y if we know the distribution of X?
Evidently, if $a = 0$, then $Y = aX + b$ has a degenerate distribution of Y: then $P(Y = b) = 1$ and $E(Y) = aE(X) + b = b$.

Property 6.4.3 If the continuous random variable X has a density function f_X, then for $Y = aX + b$, with $a \neq 0$:

$$\boldsymbol{f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)}$$

Proof: we will prove the formula $a > 0$ (for $a < 0$ see a numerical example in exercise 6.5a) with the approach above:

1. $F_Y(y) = P(Y \le y) = P(aX + b \le y) = P\left(X \le \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$
2. Differentiate $F_Y(y)$ w.r.t. y, applying the chain rule:

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{1}{a} \cdot f_X\left(\frac{y-b}{a}\right)$$ ■

Example 6.4.4 A random number generator produces random numbers X between 0 and 1, so X has a uniform distribution: $f_X(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{elsewhere} \end{cases}$
How could we generate random numbers in other intervals, such as $[4, 7]$?
A common sense solution is to compute $Y = 3X + 4$: if X lies between 0 and 1, then Y lies between 4 and 7.
Is $Y = 3X + 4$ really $U(4, 7)$-distributed?
According property 6.4.3 we have: $f_Y(y) = \frac{1}{3} \cdot f_X\left(\frac{y-4}{3}\right)$
$f_X\left(\frac{y-4}{3}\right) = 1$, if $0 \le \frac{y-4}{3} \le 1$, which is equivalent to $0 \le y - 4 \le 3$ and $4 \le y \le 7$, so:

$$f_Y(y) = \begin{cases} \frac{1}{3} & \text{if } 4 \le y \le 7 \\ 0 & \text{elsewhere} \end{cases}, \qquad \text{so } Y \sim U(4, 7)$$ ■

Example 6.4.4 is easily generalized:

If $X \sim U(0,1)$, then $Y = (b - a)X + a \sim U(a, b)$.

The bounds of the range of Y can easily be checked: if $X = 0$, then $Y = a$ and if $X = 1$, then $Y = b$.

Another application of random numbers between 0 and 1 is generating observations taken from a specific probability distribution. They can be used for simulation purposes, e.g. when testing the performance of complicated systems with services and interarrival times, often modeled as exponentially distributed variables.

Property 6.4.5 If X has a uniform distribution on $(0,1)$,
then $\boldsymbol{Y = -\frac{\ln(X)}{\lambda}}$ has an **exponential** distribution with parameter $\lambda\ (>0)$.

Proof: we will apply the 3 steps approach to derive the density function f_Y:

1. $F_Y(y) = P(Y \le y) = P\left(-\frac{\ln(X)}{\lambda} \le y\right) = P(\ln(X) \ge -\lambda y) = P\left(X \ge e^{-\lambda y}\right)$
$= 1 - F_X\left(e^{-\lambda y}\right)$
2. Differentiating: $f_Y(y) = \frac{d}{dy}F_Y(y) = \lambda e^{-\lambda y} f_X\left(e^{-\lambda y}\right)$
3. $X \sim U(0,1)$, so $f_X\left(e^{-\lambda y}\right) = 1$ if $0 \le e^{-\lambda y} \le 1$, so for $y \ge 0$

Conclusion: $f_Y(y) = \begin{cases} \lambda e^{-\lambda y} & \text{if } y \ge 0 \\ 0 & \text{elsewhere} \end{cases}$, so $Y \sim Exp(\lambda)$ ■

6.5 The normal distribution $N(\mu, \sigma^2)$

If we choose an arbitrary 20 years old Dutchman and measure his height (in cm), his weight (in kg) and his lung content (in liters), we can consider these quantities to be continuous random variables. Usually these types of "natural" variables have symmetric distributions: they vary symmetrically around an average according to a mound shaped density function.
If X is the length of a 20 years old Dutchman and the "mean length" is 183 cm, so $E(X) = 183$, then many of these Dutchmen will have a length close to 183.
The larger the distance to 183 cm, the fewer 20 years old Dutchmen have this length.
The symmetry means, for example, that the proportion of men with a length less than 173 cm equals the proportion of men with a length larger than 193 cm.
This variable and many other variables in biology, economy, engineering, etc. show such a symmetric, **mound shaped** (or **bell shaped**) density function, which is known as the normal distribution or the Gauss distribution.
Because of its many applications the normal distribution plays a central role in both probability theory and statistics.

Definition 6.5.1 The random variable X has a **normal distribution with parameters μ and σ^2** if the density function of X is defined by

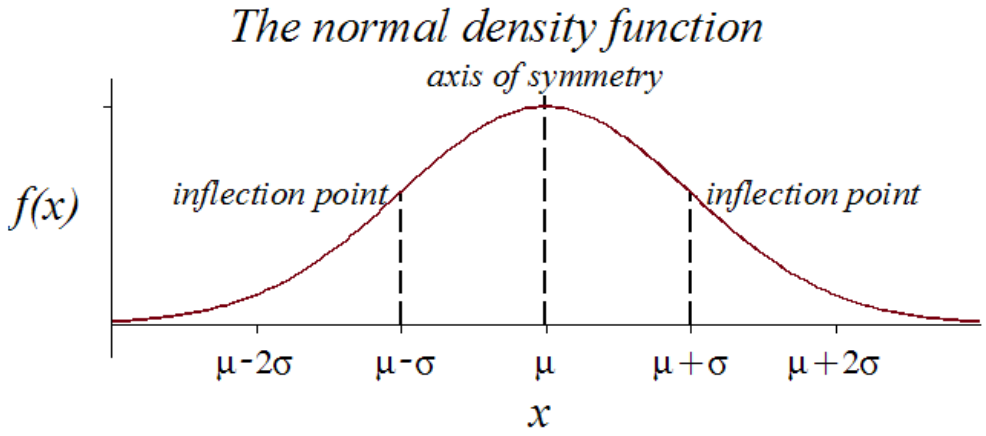
$$\boldsymbol{f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}, \text{ with } x \in \mathbb{R}$$

Short notation: $\boldsymbol{X \sim N(\mu, \sigma^2)}$ or: X is $N(\mu, \sigma^2)$-distributed.

In general there are no restrictions to the value μ, but σ^2 should be positive: $\sigma^2 > 0$.

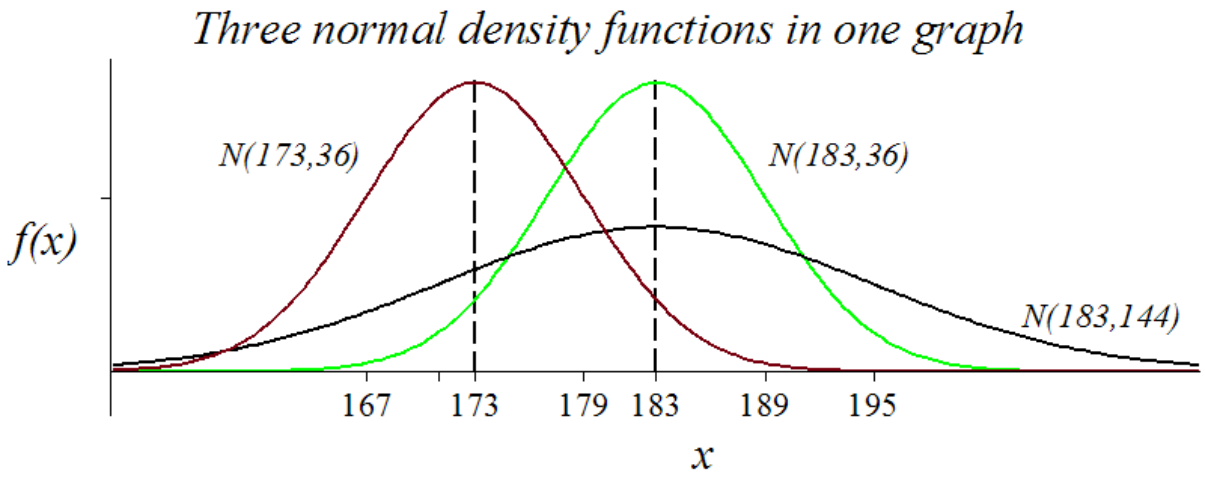
In section 6.3 we encountered the standard normal or $N(0,1)$-distribution: it has parameters $\mu = 0$ and $\sigma^2 = 1$. Emphasizing the importance of the $N(0,1)$-distribution we will introduce a special notation φ for the density function instead of the usual notation f.

For graphing the $N(\mu,\sigma^2)$-distribution, we note that $f(x)$ is **symmetric about the line $x = \mu$** because of the square in the exponent of the e-power and the maximum is $f(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}$.



Furthermore the graph of f has **two inflection points, for $x = \mu - \sigma$** and **$x = \mu + \sigma$** we have $f``(x) = 0$, and the X-axis is a horizontal asymptote. It will be shown in property 6.5.2 that the symbols μ and σ^2 are really the expectation and variance of the normal distribution.

The standard deviation $\sigma = \sqrt{\sigma^2}$ is a so called scale parameter: the larger σ, the smaller the probability of values close to μ, since $f(\mu)$ decreases and the larger the probability of large deviations from μ. This is illustrated by the graphs of the density functions of height X in a population having a $N(183,144)$-, a $N(183,36)$- or a $N(173,36)$- distribution.
Note that the inflection points lie at a distance σ = 12 and σ = 6 cm from the mean height μ = 183 cm and σ = 6 from μ = 173, respectively.



How to compute probabilities for the normal distribution, e.g.: if the height of an arbitrarily chosen man is $N(183,36)$-distributed, what is the probability that his height is less than 190 cm?

Just like in the case of a standard normal distribution, these probabilities cannot be computed by integration of the density function: we will have to apply the technique of numerical approximation.
But all normal distributions have identical shapes: apart from the scale they are all look-a-likes of the standard normal distribution. By rescaling any $N(\mu, \sigma^2)$-distribution to the $N(0,1)$-distribution, we can use the table of approximated standard normal probabilities $\Phi(z) = P(Z \leq z)$, where $\Phi(z)$ is the special notation of the standard normal distribution function. The validity of the described approach is confirmed by the following property.

Property 6.5.2 If $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0,1)$, then:

a. $\boldsymbol{\sigma Z + \mu \sim N(\mu, \sigma^2)}$
b. The **z-score** $\boldsymbol{\frac{X-\mu}{\sigma} \sim N(0,1)}$
c. $\boldsymbol{E(X) = \mu}$ and $\boldsymbol{var(X) = \sigma^2}$

Proof:

a. If $Y = \sigma Z + \mu$, then, applying $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$ for $Y = aX + b$ (property 6.4.3):
$f_Y(y) = \frac{1}{\sigma} \cdot \varphi\left(\frac{y-\mu}{\sigma}\right)$, where $f_Z(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$.
So $f_Y(y) = \frac{1}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$, the density function of the $N(\mu, \sigma^2)$-distribution.

b. We have $Y = \frac{X-\mu}{\sigma} = \frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma}$, so if we apply property 6.4.3 on $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, we can show that Y has a $N(0,1)$-density function.

c. According the b-part it follows: if X is $N(\mu, \sigma^2)$-distributed, then $Z = \frac{X-\mu}{\sigma}$ is $N(0,1)$-distributed. For a standard normally distributed Z we have: $E(Z) = 0$ and $var(Z) = 1$, so:

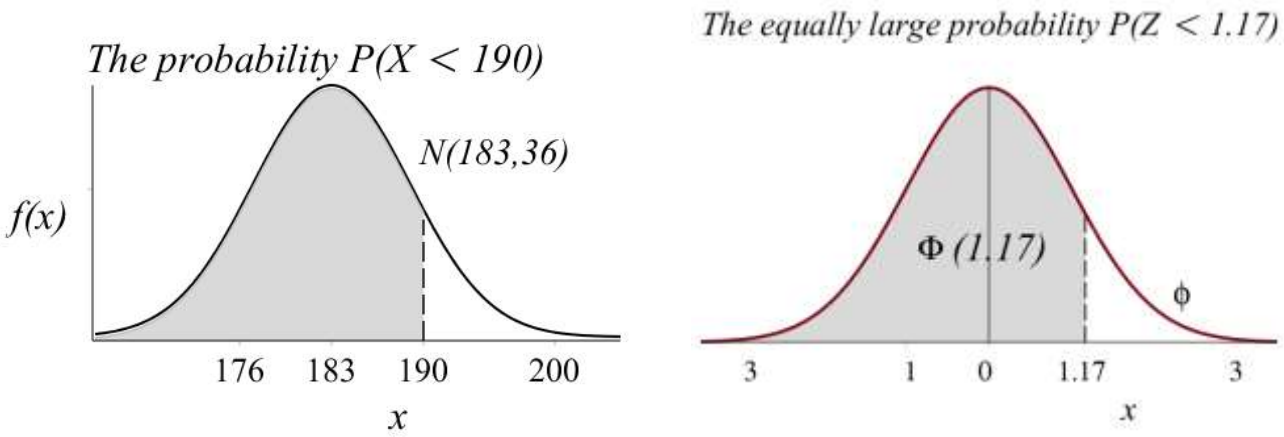
$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = E\left(\frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma} \cdot E(X) - \frac{\mu}{\sigma} = 0 \iff E(X) = \mu.$$

$$\text{Similarly: } var(Z) = var\left(\frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma}\right) = \left(\frac{1}{\sigma}\right)^2 \cdot var(X) = 1 \iff var(X) = \sigma^2 \quad \blacksquare$$

Property 6.5.2b will be applied often: whenever we want to compute probabilities with respect to a normally distributed variable.

Example 6.5.3 X is the height (in cm) of a man, arbitrarily chosen from a population of $N(183, 36)$-distributed heights.

- The standard deviation of the heights in the population is $\sigma = \sqrt{36} = 6$.
- $Z = \frac{X-183}{6}$ has a standard normal distribution.
- $P(X \leq 190) = P\left(\frac{X-183}{6} \leq \frac{190-183}{6}\right) \approx P(Z \leq 1.17) = \Phi(1.17) = 0.8790$
  We will call $\frac{190-183}{6} \approx 1.17$ the **z-score** of the height 190 cm: usually we will round this value in two decimals, since this is the precision of the standard normal table.



- $P(X > 200) = P\left(Z > \frac{200-183}{6}\right) \approx 1 - \Phi(2.83) = 1 - 0.9977 = 0.23\%$
  A height larger than 2 meter does not occur often (in this population).
- $P(X \leq 176) = P\left(\frac{X-183}{6} \leq \frac{176-183}{6}\right) \approx \Phi(-1.17) = 1 - \Phi(1.17) = 1 - 0.8790$
  $= 12.10\%$
  Compare $P(X \geq 190)$ and $P(X \leq 176)$ in the graph above:
  $P(X \leq 176) = 1 - P(X \leq 190)$ (symmetry about $x = 183$)
- $P(183 \leq X \leq 190) = P\left(\frac{183-183}{6} \leq Z \leq \frac{190-183}{6}\right) \approx \Phi(1.17) - \Phi(0)$
  $= 0.8790 - 0.5 = 37.10\%$

  in the last computation we used:
  $183 \leq X \leq 190 \quad \Leftrightarrow \quad 0 \leq X - 183 \leq 7 \quad \Leftrightarrow \quad 0 \leq \frac{X-183}{6} \leq \frac{7}{6}$ ■

The empirical rule, z-scores en percentiles

When introducing the variance σ^2 and the standard deviation σ we gave an interpretation of these measures of variation, using the **empirical rule for mound shaped distributions**: the probabilities of X lying in intervals having shape $(\mu - k\sigma, \mu + k\sigma)$ are approximately 68% ($k = 1$), 95% ($k = 2$) and 99.7% ($k = 3$), respectively.
These percentages are based on the normal distribution. This fact is easily checked:

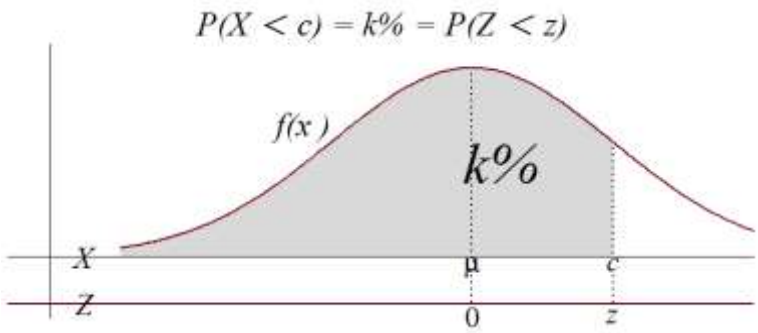
If $k = 2$ we have: $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = P\left(-2 \leq \frac{X-\mu}{\sigma} \leq 2\right)$
$= \Phi(2) - \Phi(-2)$
$= \Phi(2) - \big(1 - \Phi(2)\big)$
$= 2 \cdot 0.9772 - 1 = 0.9544 \approx 95\%$

Probabilities w.r.t. a $N(\mu, \sigma^2)$-distributed X are obtained by "standardization":

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

where the real value $\frac{x-\mu}{\sigma}$ is called $\boldsymbol{z}$ **−score** (z-value) of the bound x.
Note that $\frac{X-\mu}{\sigma}$ is a variable.

The $\boldsymbol{k^{th}}$ **percentile** is such a value c that $\boldsymbol{P(X \leq c) = k\%}$.
If we search the z-score $\boldsymbol{z}$ in the $N(0, 1)$-table, such that $\Phi(z) = k\%$, it follows from $P(X \leq c) = \Phi\left(\frac{c-\mu}{\sigma}\right) = k\%$, that $\frac{c-\mu}{\sigma} = z$, or: $c = \mu + z \cdot \sigma$ is the k^{th} percentile.
The computation of the k^{th} percentile is illustrated in the following graph.



Example 6.5.4 (Process control)
Production processes are regularly checked: is the production still in control or should we reset the production parameters? If the precise dimensioning of products is an issue (e.g. length, weight, content), we usually choose to aim at a desired level μ. The precision (allowed error) is expressed by the standard deviation σ. The process control can in that case be carried out by the so called 3σ-rule: if the products' measurements are (often) outside the interval $(\mu - 3\sigma, \mu + 3\sigma)$, we have an indication that the production is out of control.
Many of these measurements (lengths, weights, contents) can be modelled with the normal distribution. If production parameters are set well the measurements should be coming from the $N(\mu, \sigma^2)$-distribution. Then, applying the empirical rule, the probability of finding measurements outside the "tolerance bounds" $\mu \pm 3\sigma$ is small: only 0.3%.
By the way, if we use large random samples to check the process control, e.g. n = 1000, the probability to find at least one value outside the interval is large: $1 - 0.997^{1000} \approx 95\%$
And the expected number measurements outside the interval is 3 out of 1000. ■

For **linear transformations $\boldsymbol{Y = aX + b}$** of the $N(\mu, \sigma^2)$-distributed variable X we can show that Y is normally distributed as well (the proof is similar to the proof given in property 6.5.2a). The parameters can simply be determined by using the rules for expectation and variance:

- $E(Y) = E(aX + b) = aE(X) + b = a\mu + b$ and
- $var(Y) = var(aX + b) = a^2 var(X) = a^2\sigma^2$

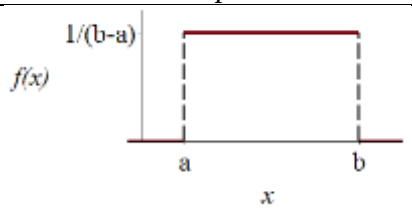
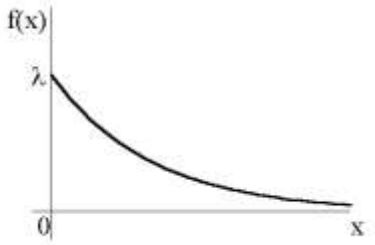
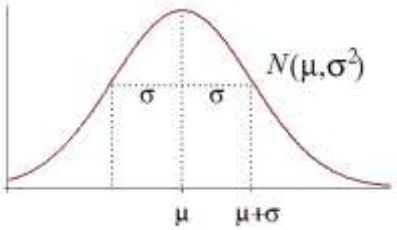
So:

Property 6.5.5 For a $N(\mu, \sigma^2)$-distributed random variable X we have:
$\boldsymbol{Y = aX + b}$ is $\boldsymbol{N(a\mu + b, a^2\sigma^2)}$-distributed (for all $a \neq 0$ and $b \in \mathbb{R}$).

In the next chapter we will see that property 6.5.5 can be extended to linear combinations of two normally distributed variables X and Y, or more than two of these variables.

6.6 Overview of frequently used continuous distributions

In the course of this chapter we encountered several types of continuous distributions. The distributions that we will apply most frequently are shown in the table below, along with the formulas for expectation and variance.
In situations where we have **random numbers**, taken from a given interval, we will use the uniform distribution on the interval as a model. For real life situations, where **waiting times**, service times or interarrival times play a role the exponential distribution can often be applied as a proper model. And, last but not least, the normal distribution provides a model for measurable quantities in nature: in biology, engineering, economy, business, etc.
In the table the density function is only given for values x, where $f(x)$ is *not* equal to 0.

Distribution	*Density function*	$E(X)$	$var(X)$	*Graph*
Uniform $U(a,b)$	$f(x)=\frac{1}{b-a}$, $a \le x \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
Exponential $Exp(\lambda)$	$f(x)=\lambda e^{-\lambda x}$, $x \ge 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	
Normal $N(\mu,\sigma^2)$	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, $x \in \mathbb{R}$	μ	σ^2	

The following relations between distributions can be given:

- Standardisation: if $X \sim N(\mu,\sigma^2)$, then $Z=\frac{X-\mu}{\sigma} \sim N(0,1)$
- Link between the standard normal and normal distribution:
  if $Z \sim N(0,1)$, then $X=\sigma Z+\mu \sim N(\mu,\sigma^2)$
- An exponential distribution can be simulated with random numbers between 0 and 1:
  if $X \sim U(0,1)$, then $Y=-\frac{\ln(X)}{\lambda} \sim Exp(\lambda)$
- Relation between $U(0,1)$ and $U(a,b)$: if $X \sim U(0,1)$, then $Y=(b-a)X+a \sim U(a,b)$

6.7 Exercises

1. The density function of X is given by: $f(x) = \begin{cases} 1 - \frac{1}{2}x, & \text{if } 0 \le x \le 2 \\ 0, & \text{elsewhere} \end{cases}$

- **a.** Sketch the graph of this density function and the probability $P(X > 1)$. Compute $P(X > 1)$.
- **b.** Determine $E(X)$, $E(X^2)$ and $var(X)$.
- **c.** Find the distribution function $F(x)$ (pay attention to the values of x, for which the formula is given!) and use F to compute $P(X > 1)$ a second time.

2. X is exponentially distributed with parameter λ.

- **a.** Give the density function f, sketch the graph of f and show that $\int_{-\infty}^{\infty} f(x)\,dx = 1$
- **b.** Check that $E(X) = \frac{1}{\lambda}$. Is $P(X > EX) = \frac{1}{2}$?
- **c.** Compute the median M, the value such that $P(X > M) = 0.50$.
- **d.** Determine the mode of X, that is the value in S_X where f attains its maximum value. Mark the position of the expectation μ, the median M and the mode m in the graph.

3. X is uniformly distributed on the interval [0, 4]

- **a.** Answer all questions of exercise 2 for this distribution.
- **b.** Determine the distribution function of X (a piecewise defined function) and sketch its graph.

4. The density function of X is given by: $f(x) = \frac{c}{x^3}$, if $x > 1$, and $f(x) = 0$ elsewhere.

- **a.** Determine c, sketch the probability $P(X > 2)$ in the graph of f and compute $P(X > 2)$.
- **b.** Determine both $E(X)$ and the median M (such that $P(X \le M) = P(X \ge M) = \frac{1}{2}$).
- **c.** Determine the distribution function F_X of X.

5. **a**. Use the 3 steps in section 6.4 (deriving the density function of $Y = g(X)$ for given distribution of X) to show that $Y = 5 - 2X \sim U(3,5)$ if $X \sim U(0,1)$.

b. If a random number X between 0 and 1 is available (so $X \sim U(0,1)$), how can you use X to "generate" a random number Y of a given interval (a, b)?

c. If $X \sim Exp(\lambda = 3)$, determine the density functions of $Y = 2X$ and $Z = X^2$.

6. X is a random number between 0 and 1, so X is uniformly distributed on (0, 1). And we will compute Y as follows: $Y = \frac{1}{X}$

- **a.** Determine the density function of Y.
- **b.** Compute $P(Y > 2)$ in two ways, using the distribution of X and of Y, respectively.
- **c.** Determine $E(Y)$ in two ways (if possible), using the distribution of X and Y.

7. X is exponentially distributed with parameter $\lambda = 1$ and $Y = \sqrt{|X|}$.

a. Determine the distribution function of X and use it to derive the distribution function of Y.
b. Determine the density function of Y and $E(Y)$.

8. X has a $N(1,4)$-distribution. (*use the N(0, 1)-table to solve this exercise*)

a. Sketch the graph of X and determine $P(X > 2)$, $P(|X| > 2)$ and $P(|X - 1| < 2)$.
b. Determine the **90th percentile** c of the distribution of X: c is such that $P(X \leq c)$ = 90%
c. Find the 10th percentile as well.

9. (**The empirical rule**). X is $N(\mu, \sigma^2)$-distributed.
Show that the probability $P(|X - \mu| < k \cdot \sigma)$ does not depend on the values of μ and σ and that the probabilities for k = 1, 2 and 3 are 68.3%, 95.4% and 99.7%, respectively.

10. At a farm the chickens produce eggs, of which the weight can be modelled with a normal distribution with "mean" 50 gram and a standard deviation $\sigma = 5$ gram. (Sketch this distribution.)
The farmer wants to sell the eggs in 5 weight classes, which should be equally large. How should the farmer choose the boundaries of the five classes?

11. $E(X - \mu)^3$ is used to define a **measure of skewness** of a distribution: $E(X - \mu)^3 = 0$ for symmetrical distributions; if the distribution is skewed to the right (the density function shows a "tail to the right", like the exponential distribution has), $E(X - \mu)^3$ is positive; and $E(X - \mu)^3$ is negative, if the distribution is skewed to the left (tail to the left).

a. Express $E(X - \mu)^3$ in the first, second and third moment: $E(X)$, $E(X^2)$ and $E(X^3)$.
b. Determine the first 3 moments and $E(X - \mu)^3$, if $X \sim U(0,1)$
c. Determine the first 3 moments and $E(X - \mu)^3$, if $X \sim Exp(\lambda = 1)$
d. Is the value of $E(X - \mu)^3$ in b. and c. indeed 0 (b.) positive (c.)?

12. (Extra exercise with respect to increasing functions of a variable: $Y = g(X)$.)

a. If $Z \sim N(0,1)$, show that $Y = e^Z$ has the following density function:

$$f_Y(y) = \frac{1}{y\sqrt{2\pi}} e^{-\frac{1}{2}(\ln(y))^2}$$

(This density function has application in economic sciences.)
b. (Generalization) If $Y = g(X)$, where X has a known density f_X and g een monotonously increasing function with inverse $u = g^{-1}$, show that:
$f_Y(y) = f_X\big(u(y)\big) \cdot u`(y)$ (this formula will be applied in module 4 - TBK)

Some hints for the exercises of chapter 6:

1. First write down the formulas for $E(X)$, $E(X^2)$, $var(X)$ and $F(x)$.
2. Expectation, Median and mode are different "measures of centre".
3. idem
4. See exercise 1.
5. Compare the questions with the solutions given in 6.4.4. an 6.4.5
6. Idem
7. Idem
8. Compare this problem to the examples given on pages 6.20-21.
9. Idem
10. Idem
11. **a.** Expand $(X-\mu)^3$ and compute the expected value for each term.
b./c. Apply $E(X^k) = \int_{-\infty}^{\infty} x^k \cdot f(x)dx$.

Chapter 7 Two or more continuous variables

7.1 Independence

If we have two **dependent discrete** random variables X and Y, the joint distribution determines the level of dependence: the correlation coefficient ρ is a measure of linear relation of the joint distribution. In case of **independence** we can easily compute the joint probabilities of two variables X and Y by using the marginal distributions of X and Y:

$$P(X = x \text{ and } Y = y) = P(X = x) \cdot P(Y = y), \qquad \text{for each pair } (x, y)$$

For **continuous** distributions a similar approach is possible, but not fully covered by this reader. We will only give an indication as how joint continuous distributions are defined and we will quickly turn to the case of two (or more) **independent continuous** variables.
We are especially interested in the distribution of the sum and the mean of two (or more) continuous and independent variables.

Therefore we start by giving the **general definition of independence of random variables**:

Definition 7.1.1 The random variables X and Y are **independent** if for each pair of sets $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$ we have:

$$\boldsymbol{P(X \in A \text{ and } Y \in B) = P(X \in A) \cdot P(Y \in B)}$$

Applying this definition to **independent discrete** variables we could choose $A = \{x\}$ and $B = \{y\}$ in the definition above, finding: $P(X \in \{x\} \text{ and } Y \in \{y\}) = P(X \in \{x\}) \cdot P(Y \in \{y\})$, which is the same as our previous definition in chapter 5:

$$P(X = x \text{ and } Y = y) = P(X = x) \cdot P(Y = y).$$

Reversely, it can be shown that, from the last equality for each pair (x, y), the equality in definition 7.1.1 follows: for discrete variables the equalities are equivalent.

For **independent continuous** variables X and Y we can immediately compute joint probabilities if we know the distributions of both X and Y, e.g.:

$$P(X \leq 4 \text{ and } Y > 3) \overset{\text{ind.}}{=} P(X \leq 4) \cdot P(Y > 3)$$

This approach enables us to find the distribution of the **maximum** and the **minimum** of two independent continuous variables.

Example 7.1.2 A helpdesk has two employees. If both are occupied serving customers, new customers have to wait in line. Suppose that a customer enters when both the employees are busy. He will be served as soon as one of the employees is available.
If we want to give information about the waiting time of the newly arrived customer, we need a probability model for this situation. Let us assume that the two service times of the customers in service (measured from the moment of entrance of the third customer) are independent and exponentially distributed variables X and Y, with the same expectation $E(X) = E(Y) = 4$ *minutes* (so the parameter $\lambda = \frac{1}{4}$).
The waiting time W of the third customer (or the first in line) is the smallest of the two service times: $W = \min(X, Y)$.
What distribution does W have? And what is the expected waiting time, $E(W)$?

Since W is a function of X and Y we could first try to derive the distribution function of W from the known distributions of X and Y: remember that for the exponential distribution we have:

$$P(X > x) = P(Y > x) = e^{-\lambda x}, \text{ if } x \geq 0$$

So: $F_W(w) = P(\min(X, Y) \leq w)$
$= 1 - P(\min(X, Y) > w)$ because of the complement rule
$= 1 - P(X > w \text{ and } Y > w)$
$\overset{\text{ind.}}{=} 1 - P(X > w) \cdot P(Y > w)$ using the independence of X and Y
$= 1 - e^{-\lambda w} \cdot e^{-\lambda w}$ since $P(X > x) = e^{-\lambda x}$ ($\lambda = \frac{1}{4}$)
$= 1 - e^{-2\lambda w}$, for $w \geq 0$

And $F_W(w) = 0$, for $w < 0$

So $f_W(w) = \frac{d}{dw} F_W(w) = 2\lambda e^{-2\lambda w}$, for $w \geq 0$
In this formula we recognize the exponential density function with parameter $2\lambda = 2 \cdot \frac{1}{4} = \frac{1}{2}$.
So the expected waiting time is $E(W) = \frac{1}{2\lambda} = 2$. ■

Similarly we can derive the distribution of the maximum of X and Y (exercise 2) and both computations can be extended to n independent variables $X_1, X_2, \ldots, X_n$, where we can apply the assumpton of independence as follows:

$$P(\max(X_1, \ldots, X_n) \leq x) = P(X_1 \leq x_1 \text{ and } \ldots \text{ and } X_n \leq x_n)$$
$$= P(X_1 \leq x_1) \cdot \ldots \cdot P(X_n \leq x_n)$$

The approach above does not "work" for functions as $X \cdot Y$ and $X + Y$.
E.g., the event $\{X + Y \leq w\}$ cannot be expressed in the shape $\{X \in A \text{ and } Y \in B\}$. For this kind of problems we would need to introduce **joint continuous distributions, which are not part of this course**. Nevertheless, in section 7.2 we will discuss a result of this approach: the density function of the sum of two independent continuous variables.

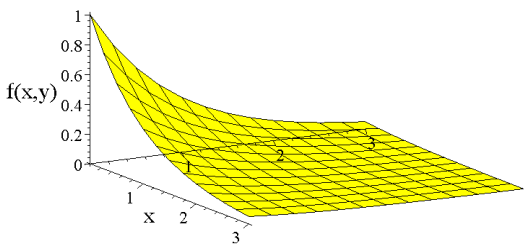
For interested students we give a rough picture for the concept of joint continuous distributions.

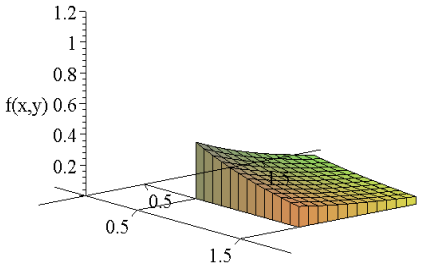
Intermezzo: the continuous joint distribution of X and Y in an example.
Similar to one continuous variable a joint density function of one continuous variable the **joint density function $f(x,y)$** can be given and graphed in 3 dimensions: $z = f(x,y)$.
Then a probability is a volume above a desired area in the XY-plane under the graph of f, e.g.:

$$f(x,y) = \begin{cases} e^{-x-y} & , \quad \text{if } x \geq 0 \text{ and } y \geq 0 \\ 0 & \quad \text{elsewhere} \end{cases}$$

Below a sketch of the graph of f and the "volume" $P(0.5 \leq X \leq 1.5 \text{ and } 0.5 \leq Y \leq 1.5)$ is shown:

The joint probability $P(0.5 < X < 1.5 \text{ and } 0.5 < Y < 1.5)$





The actual computation of the probability is a "double" (repeated) integral:

$$\int_{0.5}^{1.5}\int_{0.5}^{1.5} e^{-x-y}\,dxdy \quad \text{(the technique is described in calculus books)}$$

In a similar way the **joint distribution function $F(x,y)$** of the pair (X,Y) is defined and determined:

$$F(x,y) = P(X \leq x \text{ and } Y \leq y)$$
$$= 1 - e^{-x} - e^{-y} + e^{-x-y}, \text{for } x \geq 0 \text{ and } y \geq 0$$

Since, e.g., $P(X \leq x) = \lim_{y\to\infty} P(X \leq x \text{ and } Y \leq y)$, we can find the **marginal distribution of X**:

$$F_X(x) = \lim_{y\to\infty} F(x,y) = 1 - e^{-x}, \text{ for } x \geq 0$$

And:

$$f_X(x) = \frac{d}{dx}F_X(x) = e^{-x}, \text{ for } x \geq 0$$

This is the exponential density function with $\lambda = 1$, so $E(X) = var(X) = 1$.
Y has the same distribution as X.

The conditional distribution can be defined an analogue way as in the discrete case:
The **conditional density function of X given $Y = y > 0$**, applied to the example is:

$$f_X(x|Y = y) = \frac{f(x,y)}{f_y(y)} = \frac{e^{-x-y}}{e^{-y}} = e^{-x}, \quad \text{for } x \geq 0$$

The **conditional expectation** $E(X|Y = y) = \int_{-\infty}^{\infty} x \cdot f_X(x|Y = y)dx = \int_{-\infty}^{\infty} x\, f_X(x)dx = 1$
In this example we saw that $f_X(x|Y = y) = f_X(x)$, so information about Y $(Y = y)$ does not affect the density function of X: X and Y are independent, in this example.
The following equalities can be derived from the general definition 7.1.1 of **independence**, for $x \geq 0$ and $y \geq 0$:

$$F(x,y) = 1 - e^{-x} - e^{-y} + e^{-x-y} = (1 - e^{-x})(1 - e^{-y}) = F_X(x) \cdot F_Y(y)$$
$$\text{And} \quad f(x,y) = e^{-x-y} = e^{-x} \cdot e^{-y} = f_X(x) \cdot f_Y(y)$$

In case of independence the joint distribution function and joint density function are **products of the marginal distribution functions and density functions**, respectively. ■

7.2 The convolution integral

The approach described in the intermezzo enables us to derive the density function of $X + Y$ for two independent continuous random variables: first the distribution function $F_{X+Y}(z) = P(X + Y \leq z)$, a double integral of $f_X(x) \cdot f_Y(y)$ on the half plane $x + y \leq z$, is determined. Then the derivative of $F_{X+Y}(z)$ results in the expression for $f_{X+Y}(z)$ below.

Property 7.2.1 (the convolution integral)

If X and Y are independent continuous variables, we have

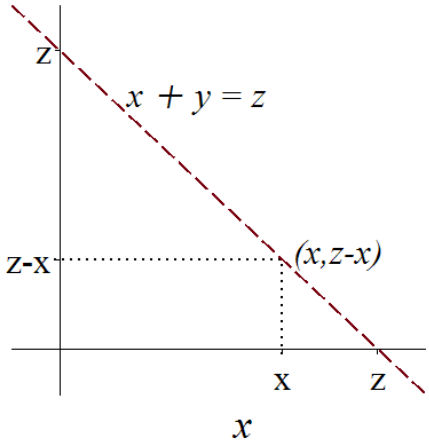
$$\boldsymbol{f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx}$$

The analogy with the convolution sum is not coincidental:

- For two independent discrete variables X and Y we determined the probability, that $X + Y = z$ occurs, by adding the probabilities that (X, Y) attains the grid points (x, y) on the line $x + y = z$ (see the graph below).
- For continuous X and Y we integrate (Riemann sum!) the densities $f_X(x) \cdot f_Y(y)$ over the line $x + y = z$.

$$P(X + Y = z) = \sum_x P(X = x)P(Y = z - x)$$

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_x(x) f_Y(z-x)\,dx$$



Instead of the sum (addition) of two independent variables this is referred to as the convolution of the variables

Example 7.2.2 (convolution of independent exponentially distributed variables)
"You are the third in line."
What does that announcement tell us about the waiting time and the expected waiting time?
If we consider the situation that all customers are served by one employee, then "third in line" means I have to wait for two services to be completed. A possible probability model is:

Model: the service times X and Y are independent and exponentially distributed with parameter λ.

This model implies that $f_x(x) = f_Y(x) = \lambda e^{-\lambda x}$ $(x \geq 0)$ and that the convolution integral can be applied to find the density function of the total waiting time $X + Y$:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx\,, \qquad \text{where } f_X(x) = 0 \text{ if } x < 0 \text{ and } f_Y(z-x) = 0 \text{ if } x > z$$

$$= \int_0^z \lambda e^{-\lambda x} \cdot \lambda e^{-\lambda(z-x)} dx$$

$$= \int_0^z \lambda^2 e^{-\lambda z} dx = \lambda^2 e^{-\lambda z} \int_0^z 1 \cdot dx \qquad \text{(note that } x \text{ is the variable and } z \text{ is fixed)}$$

$$= \lambda^2 e^{-\lambda z} [x]_{x=0}^{x=z} = \lambda^2 z e^{-\lambda z}, \qquad \text{for } z \geq 0$$

And $f_{X+Y}(z) = 0$ if $z < 0$.
This distribution is known as the **Erlang distribution** with parameters $n = 2$ and λ,
The sum $X + Y + W$ of independent, $Exp(\lambda)$-distributed variables has an Erlang distribution with parameters $n = 3$ and λ: this distribution can be derived with the convolution integral applied on $X + Y$ (above we determined its distribution) and W ($W \sim Exp(\lambda)$).
The Erlang distribution will be discussed in more detail in chapter 8.

The density function of $f_{X+Y}(z)$ can be used to compute $E(X + Y)$ and $var(X + Y)$, but we will prefer to apply the rules of expectation and variance in this case:

$$E(X + Y) = E(X) + E(Y) = \frac{1}{\lambda} + \frac{1}{\lambda} = \frac{2}{\lambda}$$

$$var(X + Y) \overset{\text{ind.}}{=} var(X) + var(Y) = \frac{1}{\lambda^2} + \frac{1}{\lambda^2} = \frac{2}{\lambda^2}$$ ■

7.3 The sum of independent and normally distributed variables

If X and Y are independent and both normally distributed, is the sum $X + Y$ normal as well? Yes!
To prove this claim we can apply the convolution integral, as shown below in the "simplest" example (7.3.2) of two standard normal variables, with some analytic effort.
The answer to the question what the parameters of the normal distribution of $X + Y$ are, is a relatively easy question to answer:

$$\mu = E(X + Y) = E(X) + E(Y) \qquad \text{or} \qquad \mu_{X+Y} = \mu_X + \mu_Y$$

And $$\sigma^2 = var(X + Y) \overset{\text{ind.}}{=} var(X) + var(Y) \qquad \text{or} \qquad \sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

Of course this can be extended to more than 2 independent normally distributed variables.

Example 7.3.1 For a new petrol station the company uses a normally distributed daily demand (in liters): the expected daily demand of normal petrol is 600 *l* and the standard deviation is 100 *l*. The daily demands on the days of a week are assumed to be independent and all have the given distribution. How large should the capacity of the petrol tank at least be such that the probability, that the petrol in stock is insufficient for one week, is at most 5%?

Probability model:
$X_1, X_2, \ldots, X_7$ are the daily demands of normal petrol during 7 consecutive days: they are independent and $N(600, 100^2)$-distributed.

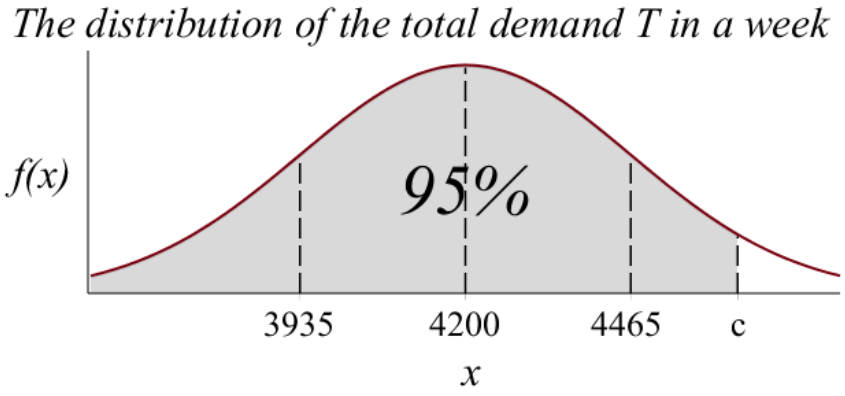
The total demand during the week, $T = X_1 + X_2 + \cdots + X_7$, is normally distributed as well with

$$\mu_T = E(X_1 + \cdots + X_7) = E(X_1) + \cdots + E(X_7) = 7 \cdot 600 = 4200\ l \text{ and}$$

$$\sigma_T^2 = var(X_1 + \cdots + X_7) = var(X_1) + \cdots + var(X_7) = 7 \cdot 100^2 = 70000$$

So $\sigma_T = \sqrt{70000} \approx 265\ l$

In the graph below the distribution of T is sketched and the minimum capacity c of the tank is indicated, such that the condition is fulfilled.



c must be such that $P(T > c) \leq 5\%$ or (rescale to the $N(0,1)$-distribution):
$P(T \leq c) = P\left(Z \leq \frac{c-4200}{265}\right) \geq 95\%$
From the $N(0,1)$ -table it follows that $\frac{c-4200}{265} \geq 1.645$, so: $c \geq 4200 + 1.645 \cdot 265 \approx 4636\ l$
Note: within the field of "Inventory Control" (*Supply Chain Management*) $4636\ l$ is called the *safety stock* at a *(cycle) service level* of 95%. ■

Besides the sum of independent and (all) $N(\mu, \sigma^2)$-distributed variables X_1, …,X_n, the **sample mean** plays a central role in statistics, if we have a random sample from the $N(\mu, \sigma^2)$-distribution:

$$\overline{\boldsymbol{X}}_{\boldsymbol{n}} = \frac{\mathbf{1}}{\boldsymbol{n}}(\boldsymbol{X_1} + \cdots + \boldsymbol{X_n}) = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{X_i}$$

Since $X_1 + \cdots + X_n$ is normally distributed, so is $\overline{X}_n = \frac{1}{n} \cdot [X_1 + \cdots + X_n]$ (property 6.5.5 assures that if X is normally distributed, then $Y = aX + b$ is as well), with parameters:

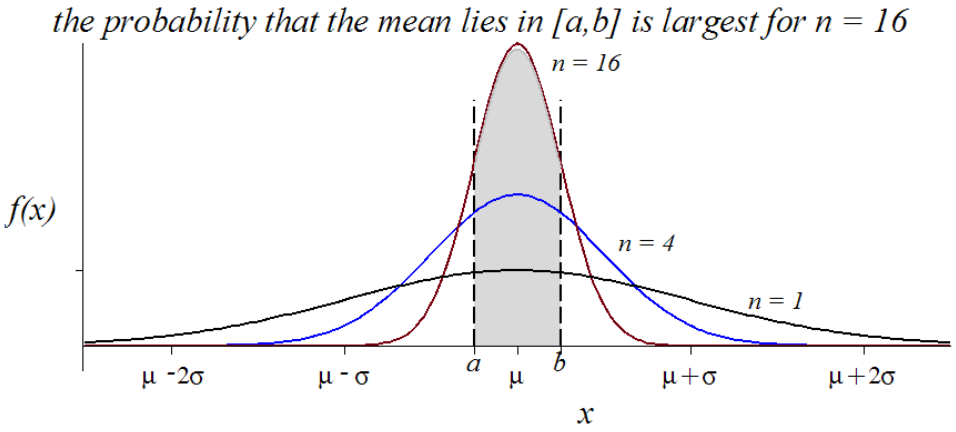
$$\mu_{\overline{X}_n} = E\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n}E(X_1 + \cdots + X_n) = \frac{1}{n} \cdot n\mu = \mu \qquad \text{and}$$

$$\sigma^2_{\overline{X}_n} = var\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n^2}var(X_1 + \cdots + X_n) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n},$$

so $\sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}$

Since $var(\overline{X}_n) = \frac{\sigma^2}{n}$ decreases, if n, the number of measurements or sample size, increases, we know that the probability that $\overline{X}_n$ attains a value close to μ increases at the same time. In the graph below the distribution of the sample mean is sketched for $n = 1$, $n = 4$ and $n = 16$.

For $n = 16$ the probability (area) that the mean is in a small interval around μ, is the largest.



If the variables are normally distributed and independent, which is the case for "random samples from a normal population", then sums $X+Y$ and $X_1+\cdots+X_n$ and the associated means $\frac{X+Y}{2}$ and $\overline{X}_n$ are all normally distributed.
As stated before the formal proof is, in general, analytically difficult. That is why we restrict ourselves to an example of which the analysis will not be part of tests and exams.

Example 7.3.2 If X and Y are independent and both $N(0,1)$-distributed, then $X+Y$ is according the convolution integral (property 7.2.1) normally distributed as well:

$$f_{X+Y}(z)=\int_{-\infty}^{\infty}f_X(x)f_Y(z-x)dx=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-x)^2}dx$$

Combining the e-powers, we can "split off a square" as follows:

$$-\tfrac{1}{2}x^2-\tfrac{1}{2}(z-x)^2=-\left[x^2-zx+-\tfrac{1}{2}z^2\right]=-\left(x-\tfrac{1}{2}z\right)^2-\tfrac{1}{4}z^2$$

$$\text{So } f_{X+Y}(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{4}z^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\left(x-\frac{1}{2}z\right)^2}dx=\frac{1}{\sqrt{2\pi\cdot 2}}e^{-\frac{1}{2}\cdot\frac{z^2}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\cdot\frac{1}{2}}}e^{-\frac{1}{2}\cdot\frac{\left(x-\frac{1}{2}z\right)^2}{\frac{1}{2}}}dx$$

Observing the last integral closely we recognize a normal density function with $\mu=\frac{1}{2}z$ and $\sigma^2=\frac{1}{2}$: the integral is the total area 1, so we found:

$$f_{X+Y}(z)=\frac{1}{\sqrt{2\pi\cdot 2}}e^{-\frac{1}{2}\cdot\frac{z^2}{2}},\qquad \text{the N}(0,2)-\text{density function.}$$

In conclusion: if X and Y are independent and both $N(0,1)$, then $X+Y\sim N(0+0{,}1+1)$
Applying property 6.4.3 we find that the mean $\frac{X+Y}{2}$ is normally distributed according a $N\left(0,\frac{1}{2}\right)$-distribution: $var\left(\frac{X+Y}{2}\right)=\frac{1}{4}var(X+Y)=\frac{2}{4}=\frac{1}{2}$. ■

As stated we can generalize this proof to two independent normal (not standard normal) variables X and Y: both the sum $X+Y$ and the difference $X-Y$ are normally distributed. Taking this information for granted, we only have to find the parameters (μ and σ^2) by applying the rules for expectation and variance.

$$E(X+Y)=E(X)+E(Y)\quad\text{and}\quad E(X-Y)=E(X)-E(Y)$$

$$var(X+Y)\overset{\text{ind.}}{=}var(X)+var(Y)\quad\text{and}\quad var(X-Y)\overset{\text{ind.}}{=}var(X)+var(Y)$$

Property 7.3.3 If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent, then we have:

a. $\boldsymbol{X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)}$
b. $\boldsymbol{X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)}$

Example 7.3.4
Stock market experts developed the following model for the yearly return of two funds (the yearly return X on agricultural products and Y on oil products):

- $X \sim N(8, 36)$ and $Y \sim N(12, 64)$
- X and Y are independent

The expected return on oil products is 50% larger than the return on agricultural products, but how large is the probability that, nevertheless, the return on agricultural products is larger?
Solution: we have to compute the probability $P(X > Y)$: standardizing each of the variables X and Y separately will not bring a solution but rewriting the event does:

$$P(X > Y) = P(X - Y > 0)$$

According property 7.3.3b we have: $X - Y \sim N(8 - 12, 36 + 64)$, so:

$$P(X > Y) = P(X - Y > 0) = P\left(Z > \frac{0 - (-4)}{\sqrt{100}}\right) = 1 - \Phi(0.4) = 1 - 0.6554 = 34.46\% \blacksquare$$

Generalizing the properties above for a sum of n variables:

Property 7.3.5
If $X_i \sim N(\mu_i, \sigma_i^2)$, for $i = 1, \ldots, n$, and $X_1, X_2, \ldots, X_n$ are independent and $S_n = \sum_{i=1}^n X_i$, then:

a. $\boldsymbol{E(S_n) = \sum_{i=1}^n \mu_i}$ and $\boldsymbol{var(S_n) = \sum_{i=1}^n \sigma_i^2}$
b. $\boldsymbol{S_n \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)}$

Property 7.3.5a. applies to not normal X_i`s with expectation μ_i and variance σ_i^2 as well.
Note that $\sigma_{S_n} \neq \sigma_1 + \cdots + \sigma_n$, but $\sigma_{S_n} = \sqrt{(\sigma_1)^2 + \cdots + (\sigma_n)^2}$.
In words we will remember this property by stating that "the sum of independent normally distributed variables is normally distributed as well, where the sum of the expectations and the sum of the variances are the parameters".
Property 7.3.5 is directly applicable to random samples taken from a normally distributed population:

Property 7.3.6 If $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$ and $X_1, \ldots, X_n$ are independent, then we have:

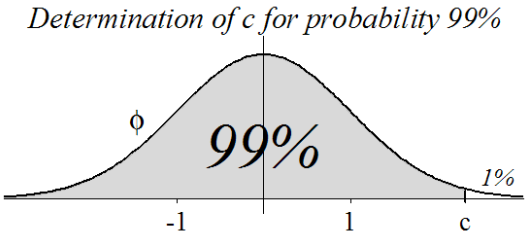
a. if $S_n = \sum_{i=1}^n X_i$, then $\boldsymbol{S_n \sim N(n\mu, n\sigma^2)}$ and
b. if $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then $\boldsymbol{\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$

Example 7.3.7
An elevator shows a sign that it can carry up to a maximum of 1000 kg.
What is the maximum number of persons that should be added to the sign if we want to make sure that the probability of overload with this number of persons is at most 1%, if we assume that the weights of the elevator users are independent and all $N(75,100)$-distributed?

We will model this situation as follows: if n is the maximum number of persons allowed, then their weights $X_1, X_2, \ldots, X_n$ are independent and $X_i \sim N(75,100)$.
According property 7.3.6 the total weight $\sum_{i=1}^{n} X_i$ is $N(n \cdot 75, n \cdot 100)$-distributed.
We will choose n such that the probability of overload $P(\sum_{i=1}^{n} X_i > 1000)$ is at most 1%, or such that $P(\sum_{i=1}^{n} X_i \leq 1000) \geq 0.99$.
After standardization we can use the standard normal distribution:

$$P\left(\sum_{i=1}^{n} X_i \leq 1000\right) = P\left(\frac{\sum_{i=1}^{n} X_i - 75n}{\sqrt{100n}} \leq \frac{\sum_{i=1}^{n} X_i - 75n}{\sqrt{100n}}\right) = \Phi\left(\frac{1000 - 75n}{\sqrt{100n}}\right) \geq 0.99$$

Using the $N(0,1)$-tabel we find a value c such that $\Phi(c) = 0.99$: $c = 2.33$.



Since $\Phi\left(\frac{1000-75n}{\sqrt{100n}}\right) \geq 0.99$, we have: $\frac{1000-75n}{\sqrt{100n}} \geq 2.33$.
The solution (the largest possible integer n) can be found by squaring, but a simpler approach is trying suitable integer values of n since we know that $n \leq \frac{1000}{75} \approx 13.3$.
For $n = 13$ we have $\frac{1000-75n}{\sqrt{100n}} \approx 0.69$, and for $n = 12$ we have $\frac{1000-75n}{\sqrt{100n}} \approx 2.89 > 2.33$
So $\boldsymbol{n = 12}$.
If 12 persons enter the elevator the expected weight is $12 \cdot 75 = 900$ and the probability of overload is:

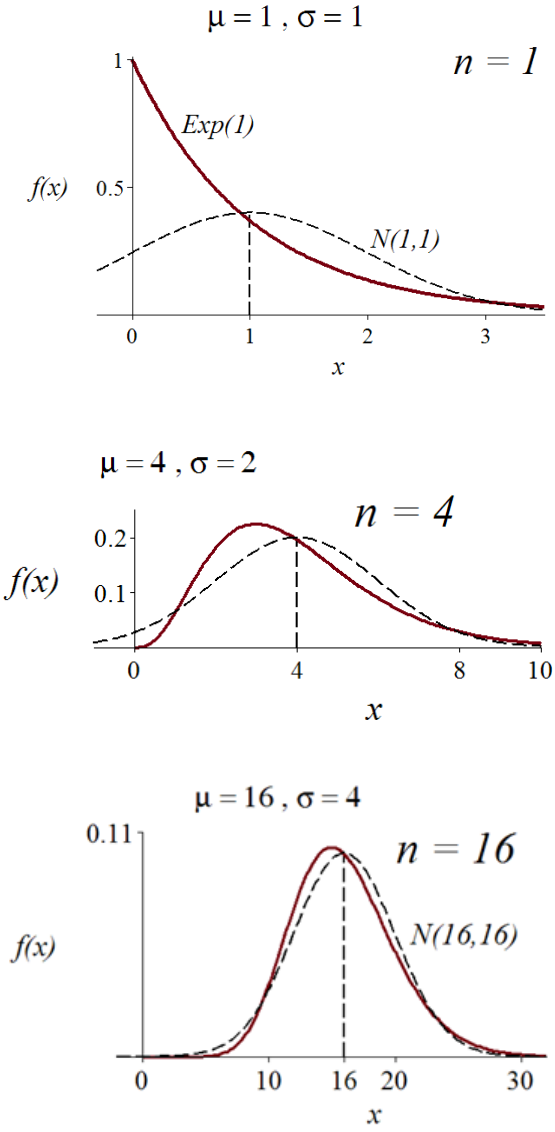
$$1 - \Phi\left(\frac{1000 - 900}{\sqrt{1200}}\right) \approx 1 - \Phi(2.89) = 0.19\% \quad (< 1\%)$$ ■

7.4 The Central Limit Theorem

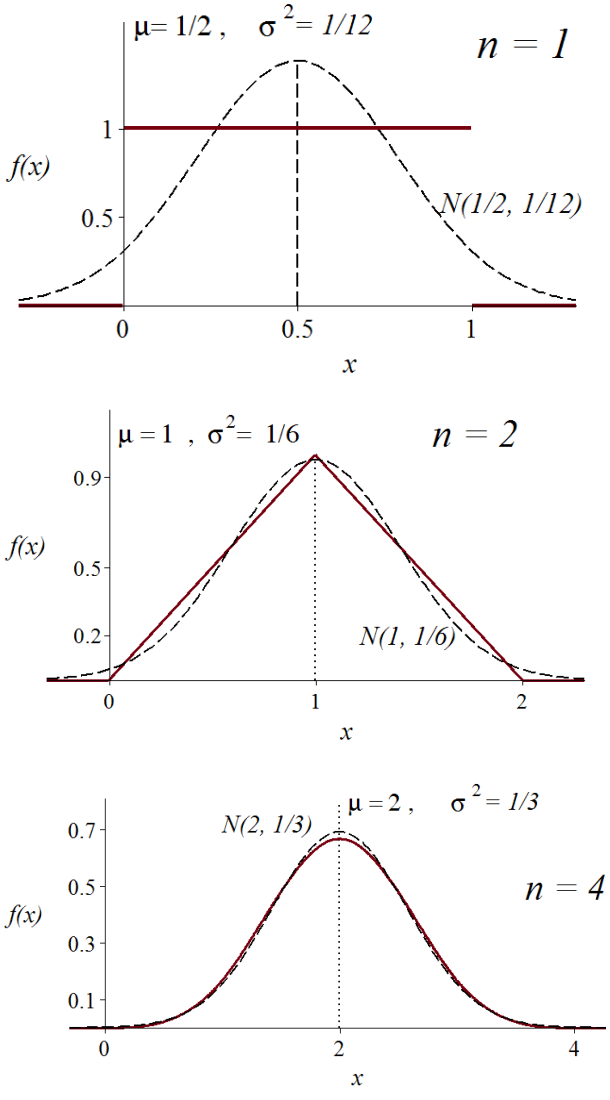
In the previous section we have seen that the sum of n independent $N(\mu, \sigma^2)$-distributed variables is normally distributed as well with expectation $n\mu$ and variance $n\sigma^2$.
If $X_1, X_2, \ldots, X_n$ are **not normally** distributed, but nevertheless are **independent and all have the same distribution with expectation μ and variance σ^2**, then for $S_n = \sum_{i=1}^{n} X_i$ we can state that $E(S_n) = n\mu$ and $var(S_n) = n\sigma^2$, but the normal distribution of S_n does not apply.
The z-score $\frac{S_n - n\mu}{\sqrt{n\sigma^2}}$ in this case is **not standard normal**.
Probabilities with respect to S_n could be computed if we first determine the distribution of S_n.
If, e.g., n customers are being served by a service desk employee and the service times of the

customers $X_1, X_2, \ldots, X_n$ are independent and exponentially distributed random variables all with parameter $\lambda = 1$ ($E(X) = var(X) = 1$), then the density function of the total service time $S_n = \sum_{i=1}^{n} X_i$ can be determined by repeated application of the convolution integral. This will be shown in the next chapter.
The obtained density functions are shown below in the left column of graphs and compared to the "corresponding" normal distributions, that is the normal distributions with the same $\mu = E(S_n) = n$ and $\sigma^2 = var(S_n) = n$.
In the right column of graphs the same is done for independent $U(0,1)$-distributed numbers $X_1, X_2, \ldots, X_n$.

X_i's have an Exponential distribution (λ=1)



X_i's have a Uniform distribution on [0,1]



The consecutive graphs reveal the issue that the Central Limit Theorem addresses: the distribution of the sum $\sum_{i=1}^{n} X_i$ tends to the corresponding normal distribution as n increases. This convergence to the normal distribution is "slower" for the exponential distribution than for the uniform distribution. This phenomenon occurs for any distribution of the X_i`s:
the sum $S_n = \sum_{i=1}^{n} X_i$ is for "large n" approximately normally distributed, or, to be precise, the **standardized** $\frac{S_n - n\mu}{\sqrt{n\sigma^2}}$ is **approximately $N(0,1)$-distributed**.

This informal statement is a consequence of the mathematical limit in the property that will not be formally proven.

Property 7.4.1 The Central Limit Theorem (CLT)

If $X_1, X_2, \ldots$ is a sequence of independent, identically distributed variables, with expectation μ and variance $\sigma^2 > 0$, then for $S_n = \sum_{i=1}^{n} X_i$ we have:

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq z\right) = \Phi(z),$$

where **Φ is the standard normal distribution function.**

Consequence: if n is "sufficiently large", then:

- $\frac{S_n - n\mu}{\sqrt{n\sigma^2}}$ is approximately $N(0,1)$-distributed
- S_n is approximately $N(n\mu, n\sigma^2)$-distributed
- $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately $N\left(\mu, \frac{\sigma^2}{n}\right)$-distributed

When n is "sufficiently large" (for which values), depends on the desired precision for approximating probabilities and on the type of distribution of the X_i's: from the graphs we concluded that f_{S_n} converges more rapidly to the normal distribution for the uniform distribution (symmetric) than for the exponential distribution ("skewed to the right").
For practical application of these approximations we will use one rule:
Rule of thumb for normal approximation with the CLT: $n \geq 25$

Example 7.4.2
What is the probability that the sum of 50 random numbers between 0 and 1 is less than 24?
The **probability model** for these 50 random numbers can be given by defining variables $X_1, X_2, \ldots, X_{50}$, that are independent and all uniformly distributed on [0,1].
Since $n = 50$ is sufficiently large to apply the CLT, we can use an approximated normal distribution of the sum $S_{50} = \sum_{i=1}^{50} X_i$: de parameters are $n\mu = 50 \cdot \frac{1}{2} = 25$ and $n\sigma^2 = 50 \cdot \frac{1}{12} = \frac{25}{6}$, where $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{12}$ for the $U(0,1)$-distribution.
So, the **normal approximation of the desired probability $P(S_{50} < 24)$** is:

$$P(S_{50} < 24) = P\left(\frac{S_{50} - 25}{\sqrt{25/6}} \leq \frac{24 - 25}{\sqrt{25/6}}\right) \overset{\text{CLT}}{\approx} \Phi(-0.49) = 1 - \Phi(0.49) = 31.21\%$$

By the way, the same question can also be phrased in terms of the mean: "What is the probability that the mean of 50 random numbers between 0 and 1 is less than 0.48 (= $\frac{24}{50}$)?
The CLT can also be applied to $\overline{X}_{50} = \frac{1}{50} S_{50}$ (you might even formulate $\overline{X}_{50}$ as a sum $\sum_{i=1}^{n}\left(\frac{X_i}{50}\right)$ of identically distributed variables):
$\overline{X}_{50}$ is approximately $N\left(\mu, \frac{\sigma^2}{n}\right)$-distributed, so $N\left(\frac{1}{2}, \frac{\frac{1}{12}}{50}\right)$-distributed. Therefore:

$$P(S_{50} < 24) = P\left(\overline{X}_{50} < 0.48\right) \overset{\text{CLT}}{\approx} \mathrm{P}\left(Z \leq \frac{0.48 - 0.50}{\sqrt{1/600}}\right) \approx \Phi(-0.49) = 31.21\% \quad \blacksquare$$

The CLT is another reason why the normal distribution is widely used as a model in nature on one hand, but also in physics and statistics on the other hand: in statistics we often determine the mean of many measurements and in physics measurement errors are often assumed to be normally distributed. A measurement error can often be seen as a sum of many (independent) elementary errors, where the CLT applies (or the generalized version of the CLT, that was proven by Bessel in 1838 for not identically distributed elementary errors).

The CLT can also be applied to discrete random variables.
If X has a binomial distribution with parameters n and p, we can conceive X as a sum of n independent, $B(1,p)$-distributed variables X_i ("alternatives"): $X = \sum_{i=1}^{n} X_i$.
If n is sufficiently large, X is according to the CLT approximately normally distributed with the binomial expectation $E(X) = n \cdot p$ and variance $var(X) = n \cdot p(1-p)$.
Since $X = \sum_{i=1}^{n} X_i$, in these formulas p and $p(1-p)$ are the expectation and the variance of the $B(1,p)$-distributions of the X_i's.

Property 7.4.3
(consequence of the CLT: normal approximation of the binomial distribution)

If $\boldsymbol{X \sim B(n,p)}$, then, **for sufficiently large $\boldsymbol{n}$, $\boldsymbol{X}$** is approximately $\boldsymbol{N(np, np(1-p))}$

For many values of $n \leq 25$ and several values of p, tables are available and should be preferred: the table values are based on exact computations and not on approximations. If for a value of $n < 25$ and p tables are not available, we will use the binomial probability function for exact computation. In practice we will prefer mostly to compute exact probabilities with Excel (or any other aid), even for large n.
In this course for "large n", that is $n \geq 25$, we can either use a Poisson approximation (if p is close to 0 or 1) or the normal approximation (p not close to 0 or 1):

Rules of thumb for approximation of binomial probabilities:

- $\boldsymbol{n \geq 25}$
- Use the **Poisson-approximation** with $\mu = np$, if $\boldsymbol{np \leq 10}$ **or** $\boldsymbol{n(1-p) \leq 10}$
- Use the **normal approximation** according to the CLT with $\mu = np$ and $\sigma^2 = np(1-p)$ if $\boldsymbol{np > 5}$ **àn**d $\boldsymbol{n(1-p) > 5}$

According to these rules sometimes both approximations are allowed, e.g. if $5 < np \leq 10$.
Remember that for $n < 25$ exact computation is mandatory.
A "good" normal approximation of binomial probabilities should be conducted with so called **continuity correction**. This technical correction is introduced in the following example where $n = 25$, the lowest value for which the normal approximation can be applied.

Example 7.4.4 Drivers in The Netherlands get their driver`s license after passing both the practical test and the theoretical exam. The theoretical exam consists of 70 yes/no-questions: a pass requires at least 60 correct answers.
Suppose a participant knows the answer to 45 out of 70 items and he decides to answer the 25 remaining questions by flipping a coin. We will assume that the 45 answers to questions he knew are all correct.
What is the probability that he will pass for his theoretical exam, or that he has at least 15 out of 25 are correct?

Model: $X =$ "# correct answers to the 25 questions he did not know": $X \sim B(25, 0.5)$.
We can apply the normal approximation since $n = 25 \geq 25$ and $np = n(1-p) = 12.5 > 5$.
So, according to the CLT, is X approximately $N(12.5, 6.25)$, where $\sigma^2 = np(1-p) = 6.25$.
He passes if 15 out of 25 coin tossed answers correct, so

$$P(X \geq 15) \overset{\text{CLT}}{\approx} P\left(Z \geq \frac{15 - 12.5}{\sqrt{6.25}}\right) = 1 - \Phi(1) = 0.1587$$

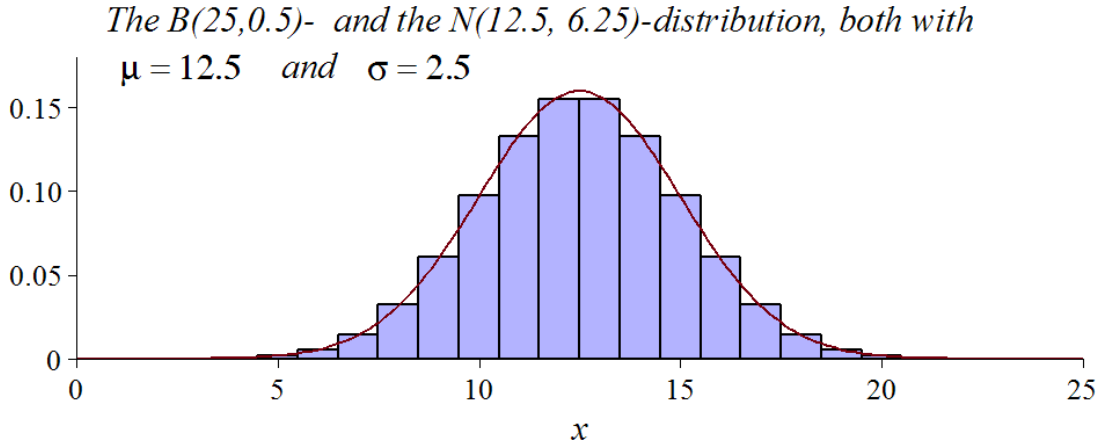
An alternative computation can be found by restating the probability to "he passes if he has more than 14 correct answers" ($X \geq 15$ is equivalent to $X > 14$):

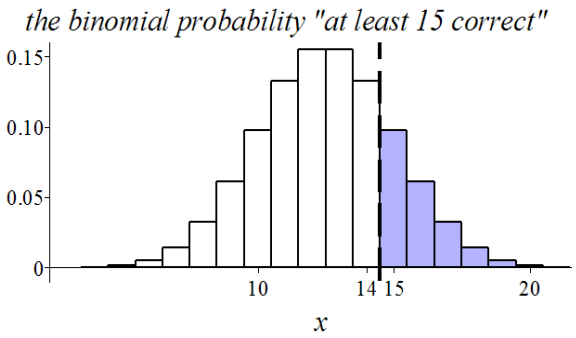
$$P(X > 14) \overset{\text{CLT}}{\approx} P\left(Z \geq \frac{14 - 12.5}{\sqrt{6.25}}\right) = 1 - \Phi(0.60) = 0.2743$$

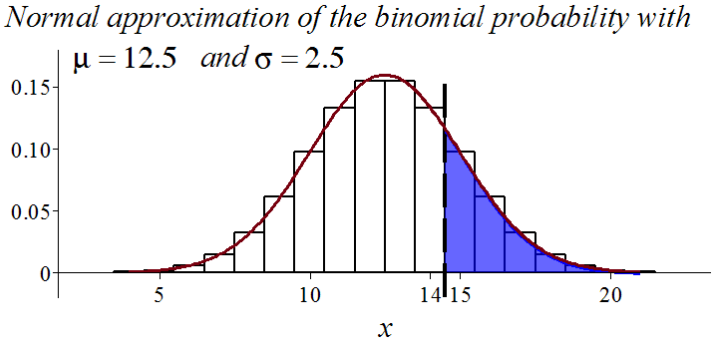
Both probabilities are approximations of the same binomial probability, but the difference is 11.5%! The different computations and results are based on the fact that X is an integer valued variable: $P(X \geq 15) = P(X > 14)$, but the associated z-scores of 15 and 14 are 1 and 0.6, which leads to large differences in the standard normal table.
Applying exact computation with the $B(25, 0.5)$-table we find:

$$P(X \geq 15) = 1 - P(X \leq 14) = 1 - 0.7878 = 0.2121$$

If we take the graphs below into account, we are inclined not to determine the z-score of 14 or 15, but 14.5.







$$P(X \geq 15) \overset{\text{cont. corr.}}{=} P(X \geq 14.5) \overset{\text{CLS}}{\approx} P\left(Z \geq \frac{14.5 - 12.5}{\sqrt{6.25}}\right) = 1 - \Phi(0.80) = 0.2119$$

Indeed: continuity correction gives (by far) the best approximation of the real probability. ■

Above we have seen why and how continuity correction (c.c.) is applied:
we transfer from a discrete (binomial) to a continuous (normal) distribution:
approximating the probability of $X \geq 15$, we included the interval (14.5, 15.5) around the boundary 15, introducing a new boundary, so: $P(X \geq 15) \overset{\text{c.c.}}{=} P(X \geq 14.5)$.
The event $X > 14$ implies that 15 is the lowest value, since X can only attain integer values: $P(X > 14) \overset{\text{c.c.}}{=} P(X \geq 14.5)$, the same probability.

Normal approximation of the binomial distribution with continuity correction:
If

- X is $B(n,p)$-distributed for sufficiently large $n \geq 25$ with $np > 5$ and $n(1-p) > 5$
- Y has a $N(np, np(1-p))$-distribution,

then we can apply normal approximations of the binomial probabilities **with continuity correction** as follows:

- $P(X \leq k) \overset{\text{c.c.}}{=} P\left(X \leq k + \frac{1}{2}\right) \overset{\text{CLT}}{\approx} P\left(Y \leq k + \frac{1}{2}\right) = \Phi\left(\frac{k+\frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$
- $P(X < k) \overset{\text{c.c.}}{=} P\left(X \leq k - \frac{1}{2}\right) \overset{\text{CLT}}{\approx} P\left(Y \leq k - \frac{1}{2}\right) = \Phi\left(\frac{k-\frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$
- $P(X = k) \overset{\text{c.c.}}{=} P\left(X \leq k + \frac{1}{2}\right) - P\left(X \leq k - \frac{1}{2}\right) \overset{\text{CLT}}{\approx} \Phi\left(\frac{k+\frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-\frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$

The effect of continuity correction is smaller as n increases.

Example 7.4.5 In example 5.6.1 we showed, using the inequality of Chebyshev, that

$$P\left(\left|\frac{X}{n} - p\right| < 0.01\right) \geq 0.90, \qquad \text{if } n \geq 25000$$

Interpretation of this probability: the sample size n should be at least 25000 to make sure that the probability that the sample proportion $\frac{X}{n}$ deviates less than 1% from the real proportion p, is at least 90%.
The probability above can be rewritten as a probability with respect to X:

$$P(\, n(p-0.01) \leq X \leq n(p+0.01)\,) \geq 0.90$$

We can try to determine the value of n, satisfying this condition, but now we will use the approximation normal distribution of X. In this case we will not apply continuity correction since the boundaries $n(p \pm 0.01)$ are not necessarily integer:

$$\begin{aligned} P(\, n(p-0.01) \leq X \leq n(p+0.01)\,) &\overset{\text{CLT}}{\approx} \Phi\left(\frac{n(p+0.01) - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{n(p-0.01) - np}{\sqrt{np(1-p)}}\right) \\ &= \Phi\left(\frac{0.01n}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{-0.01n}{\sqrt{np(1-p)}}\right) \\ &= 2\Phi\left(\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 \geq 0.90 \end{aligned}$$

From this it follows that $\Phi\left(\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right) \geq 0.95$, so $\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}} \geq 1.645$ and $n \geq \left(\frac{1.645}{0.01}\right)^2 p(1-p)$.

n depends on the (unknown) p, but we can use that $p(1-p) \leq \frac{1}{4}$, for all $p \in [0,1]$.
So $n \geq \left(\frac{1.645}{0.01}\right)^2 \cdot \frac{1}{4} \approx 6765$ for all possible values of the real proportion p.
The normal approximation leads to a much smaller value of n than Chebyshev`s rule, which is not surprising: Chebyshev`s rule is a general property, whilst normal approximation uses specific properties of the binomial distribution. ■

The normal approximation in this example is performed on the integer number X, but we could also directly approximate the probability $P\left(\left|\frac{X}{n} - p\right| < 0.01\right)$, by using the approximate normal distribution of $\frac{X}{n}$, with the same result.
Since the **number $\boldsymbol{X}$** is approximately $N(np, np(1-p))$, the **sample proportion** $\boldsymbol{\frac{X}{n}} = \frac{1}{n} \cdot X$ is approximately normal as well: a $\boldsymbol{N\left(p, \frac{p(1-p)}{n}\right)}$-distribution, since:

$$\mu = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n} \cdot np = p \text{ and}$$
$$\sigma^2 = var\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right)^2 var(X) = \left(\frac{1}{n}\right)^2 \cdot np(1-p) = \frac{p(1-p)}{n}$$

In statistics we will denote the sample proportion $\frac{X}{n}$ as $\boldsymbol{\hat{p}}$, the usual estimate of the (unknown) population proportion p.

Closing the discussion of the Central limit Theorem we will give one more application: the **normal approximation of the Poisson distribution**.
We know that the Poisson distribution approximates the $B(n,p)$ distribution for "large n and small p", but if $\mu = np > 10$ we will use a normal approximation.
If the Poisson distribution is the suitable model from the beginning and the parameter $\mu > 10$, there are no probability tables available, We can apply the CLT in this case, but why?
Well, we know that a Poisson variable is the number of events in a specified period and/or area. By splitting the period or area into n equally large parts and defining variables X_i (the number of events in part i), we are creating a model with independent X_i`s, such that $X = \sum_{i=1}^{n} X_i$, where the CLT applies for sufficiently large n.
For example, if X is the Poisson distributed number of customers who call a service desk in an hour with a mean of $\mu = 90$ customers per hour, then the number of calling customers in a minute is Poisson as well, with mean $\mu = \frac{90}{60} = 1.5$.
For each minute in an hour we define an X_i and $X = \sum_{i=1}^{60} X_i$.
Assuming that the X_i's are independent, the CLT asserts that X is approximately normally distributed with expectation $60 \cdot 1.5 = 90$ and variance $60 \cdot 1.5 = 90$.

In general we can state that for large $\boldsymbol{\mu}$ **(> 10) the Poisson distribution can be approximated by the $\boldsymbol{N(\mu, \mu)}$-distribution**. Usually we will apply continuity correction to get the best possible approximation.

7.5 Exercises

1. Two men are arbitrarily chosen from a population of 20 year old men whose weights are distributed according a $N(80, 100)$ distribution: Their weights are X and Y.

- **a.** Compute $P(X > 90 \text{ and } Y > 90)$.
- **b.** Compute $P(X + Y > 180)$
- **c.** Why is the probability in b. larger than the one in a.?

2. In the examples 6.1.4 and 6.1.6 we used the (given) density function of the maximum M of 3 random numbers between 0 and 1: $f_M(m) = \begin{cases} 3m^2 & \text{if } 0 \le m \le 1 \\ 0 & \text{elsewhere} \end{cases}$
In this exercise we will derive this density from the $U(0,1)$-distribution of 3 random numbers X_1, X_2 and X_3. So $M = \max(X_1, X_2, X_3)$.

- **a.** First give the density function $f(x)$ and distribution function $F(x)$ of the $U(0,1)$-distribution.
- **b.** Then express $F_M(m) = P(max(X_1, X_2, X_3) \le m) = \cdots$ in F.
  After applying the distribution function $F(x)$ in a., you can find $f_M(m)$ via the derivative of $F_M(m)$.

3. X and Y are independent and exponentially distributed waiting times with (different) parameters λ_1 and λ_2.

- **a.** Use the convolution integral to find the density function of $X + Y$ if $\lambda_1 = \lambda_2 = 1$.
- **b.** Determine the density function of $X + Y$ if $\lambda_1 = 1$ and $\lambda_2 = 2$.
- **c.** Compute $P(X > 1 \text{ and } Y < 1)$ if $\lambda_1 = 1$ and $\lambda_2 = 2$.

4. If Z_1 and Z_2 are independent and standard normal $\big(N(0,1)\big)$ and $X = Z_1^2$ and $Y = Z_2^2$, then $X + Y = Z_1^2 + Z_2^2$ has (by definition) a **Chi-square distribution with 2 degrees of freedom.** In example 6.4.2 we showed that $X = Z_1^2$ (and $Y = Z_2^2$) has a Chi-square distribution with 1 degree of freedom: $f(x) = \frac{1}{\sqrt{2\pi x}} e^{-\frac{1}{2}x}$, for $x > 0$.
Apply the convolution integral to derive the density function of $X + Y$. In the analysis you will have to use the fact that $\int_0^z \frac{1}{\sqrt{x(z-x)}} dx = \pi$ (see your calculus book).
Which distribution is apparently the same as this Chi-square distribution with 2 degrees of freedom? (*In statistics we will use this Chi-square distribution to find confidence intervals and conduct hypotheses testing*.)

5. A survey is conducted on the financial situation of parents who are both working: the salary X of the husband and the salary Y of his wife are both determined.
Suppose that these variables have expectations μ_X and μ_Y and variances σ_X^2 and σ_Y^2.

- **a.** Is it reasonable to state that $\mu_X + \mu_Y$ is the expectation of the total salary $X + Y$? Explain your answer.
- **b.** Is it reasonable to state that $\sigma_X^2 + \sigma_Y^2$ is the variance of the total salary $X + Y$? Explain your answer.

6. Breakable bottles are transported in a lorry. During the transport force exercised on the bottles (e.g. due to shocks), which are assumed to have a $N(50, 100)$-distribution. The breaking strength of the bottles is $N(60, 36)$-distributed.
Give in a probability model the assumptions needed to determine the probability that a bottle breaks, and compute this probability.

7. X and Y are independent: X has a $N(4, 1)$- and Y has a $N(2, 4)$-distribution.
- **a.** Compute $P(3X \leq 2Y - 1)$.
- **b.** Compute $\rho(X, 3X - 2Y)$.

8. (former exam exercise) Compute, or approximate, the following probabilities by expressing them in the standard normal distribution function Φ and state explicitly which properties you are using.

- **a.** The probability that an arbitrary boy and an arbitrary girl weigh more than 150 kg if the weight of boys is $N(75, 250)$ and the weight of girls is $N(65, 150)$.
- **b.** $P(\sum_{i=1}^{100} X_i \leq 58)$, if $X_1, X_2, \ldots, X_{100}$ are independent and all exponentially distributed, with parameter $\lambda = 2$.

9. (former exam exercise) A machine processes jobs. From past experience we know that the process time is, on average, 95 units of time, and the standard deviation of the process time is 20 units of time. Suppose 100 jobs are offered to be processed (one at a time) and we want to compute the probability that the mean process time will be at least 100 units of time.

- **a.** Which assumptions are necessary (in a probability model) to compute this probability using the CLT?
- **b.** Compute (approximate) the requested probability.

10. A researcher in the Netherlands conducted a survey among 250 arbitrarily chosen adult citizens, as to determine whether a party (A) still has the same proportion of voters as during last elections: 25%.
Let X be the number of party A voters among the chosen 250 adults.

- **a.** What kind of distribution would you assume for X and which approximating distribution could you use to determine probabilities if the population proportion of party A voters is still really 25%?
- **b.** Approximate, with continuity correction, the probability that the researcher finds that party A looses at least 3% of the voters (the proportion in the sample is 22% or less), where in reality the true population proportion is still 25%.
- **c.** The researcher also recorded the number of party B voters in the sample. Approximate the probability to find more than 5 party B voters in the sample if in reality the proportion of party B voters in the population is 1%.

11. To get an impression how many bicycles in The Netherlands have malfunctioning lights, the lights of 100 arbitrarily chosen bicycles are checked by the police.
X is defined as the number of bicycles with malfunctioning lights in the sample.
The probability that an arbitrary bicycle has malfunctioning lights is denoted as p.

- **a.** Give (a reasonable choice of) the distribution of X: express $E(X)$ and $var(X)$ in p.

b. Give a lower bound of the probability $P\left(\left|\frac{X}{100} - p\right| \leq 0.05\right)$, based on a normal approximation of the sample proportion $\frac{X}{100}$.
To find the lower bound you may use the inequality $p(1-p) \leq \frac{1}{4}$ for all p.

12. The number of iPhones a dealer sells a day is assumed to have a Poisson distribution with expectation 6. Furthermore it is assumed that the sales on consecutive days are independent.
Once a week the shop is supplied with new iPhones to sell. The shop is open during 6 days each week.

a. What is the probability that a stock of 40 iPhones is sufficient for one week`s sales? (This probability is called the "service level".)
b. How large should the stock of iPhones be such that it is enough with a probability of at least 99%? (Or: "at what safety stock is the service level 99%?"

13. (Former test exercise) The quality control of the mass production of nails is organized by measuring the nails in a relatively small sample of n nails. The company guarantees that at most 1% of the nails have a size outside prescribed tolerance bounds. For answering the following questions assume that exactly 1% of the nails are substandard (outside tolerance bounds). X is the number of substandard nails in a random sample of n nails.

a. Compute $P(X \geq 1)$, the probability that at least one of the nails is substandard, in a random sample of $n = 15$ nails.
b. Compute $P(X \leq 3)$ for a random sample of $n = 200$ nails.
c. If the sample size $n = 4000$ nails, compute $E(X)$, $var(X)$ and $P(X \geq 50)$.

Some hints for solving chapter 7 exercises:

1. **b**. What distribution does $X + Y$ have? Why?
2. Compare this problem with the approach in example 7.1.2.
3. First write down the general formula of the convolution integral. The integration line is $x + y = z$.
4. See 3.
5. Common sense: would you state that there is a relation between the salary of a man and that of his wife?
6. Rewrite the event $X > Y$ to $X - Y > 0$ and determine the distribution of $X - Y$.
7. **a.** Use the same approach as in 6.
b. Recall the computational rules for $cov(X,Y)$, $var(X)$ and $\rho(X,Y)$.
8. Reason for both parts whether you can use an exact normal distribution or an approximating one (applying the CLT).
9. Is normality van de job service times given? Is such an assumption for each job reasonable and necessary to answer the question?
10. 22% of 250 = … voters.
11. $\frac{X}{n}$ is approximately normal: for the parameters use the binomial distribution of X to determine $\mu = E\left(\frac{X}{n}\right)$ and $\sigma^2 = var\left(\frac{X}{n}\right)$.
12. $n = 6$ is too small to apply the CLT, but the expectation $\mu = 6 \cdot 6 = 36$ for a week is sufficiently large (> 10) to apply a normal approximation of the Poisson distribution! See the last remark in the chapter.

Chapter 8 Waiting times

8.1 Waiting time distributions and the lack of memory property

Waiting times, service times, lifetimes and interarrival times play an important role in information technology and technical business applications. Stochastic waiting times occur in post offices, at cash registers in supermarkets, access roads to roundabouts and all kind of electronic devices, helpdesks, telecommunication networks, websites and computer systems. Often we are interested in the performance of the system and specific aspects, such as the expected length of the waiting line, the mean waiting time or service time of customers, the maximum capacity of the system, etc.
In this chapter we will discuss the basic waiting time models and their properties.

Example 8.1.1 The mean service time of a customer at a counter of a post office is two minutes. Let us assume that the service time X (in minutes) of an arbitrary customer can be modelled by an exponential distribution.
Since the mean service time can be interpreted as the expectation $E(X) = \frac{1}{\lambda} = 2$, the parameter $\lambda = \frac{1}{2}$. As $E(X) = 2$ minutes is the mean service time, $\lambda = \frac{1}{2}$ is the "mean number of customers served in one minute". For this reason λ is called the **intensity of the service process**. (In ten minutes we would expect to serve about $10 \cdot \lambda = 5$ customers.)
Suppose that two counters in the post office are open for service. When entering the post office, customer 1 is served at the first counter and another customer (2) walks to the second counter to be served. Will it be my turn quicker if I choose to stand in line at the first counter, where the customer was already being served?
Intuitively you might think this is the case, but analysis shows there is no difference.
If the service times X_1 and X_2 of customers 1 and 2 both have the given exponential distribution, then the probability that the service time of customer 2 is more than t minutes is given by:

$$P(X_2 > t) = 1 - F_{X_2}(t) = e^{-\lambda t}$$

(We will use the general notation with λ: keep in mind that in this case $\lambda = \frac{1}{2}$.)
And if customer 1 was being served s minutes at my entrance, then the probability that the **remaining service time** is more than t minutes, after entering, is a conditional probability: the probability of a total service time more than $s + t$, given that the service time is larger than s:

$$P(X_1 > t + s | X_1 > s) = \frac{P(X_1 > t + s)}{P(X_1 > s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X_2 > t)$$

In words: irrespectively the time the service has been conducted, the **remaining** service time has the same (exponential) distribution as if measured from the start of the service.
This is why the exponential distribution is said the have the **lack of memory property**. ■

Situations as described in the example above induced the formulation of a **waiting time paradox**.
If, for example, the lifetime X of a new type of light bulb is supposed to have an exponential distribution with "mean" lifetime $E(X) = 10000$, then at any point of its life the remaining life time has still the same exponential distribution: if the bulb "survived" 5000 hours, the remaining expected lifetime is 10000 hours. So at that point of time the total expected life time is 5000 + 10000 = 15000 hours, 5000 hours more than the average of all lifetimes. This alleged contradiction can be intuitively explained by the fact that, if we consider the proportion of bulbs that survive 5000 hours, all the bulbs with a shorter life time are left out of consideration. Expressed as a conditional expectation we found:

$$E(X|X > 5000) = 5000 + E(X) = 5000 + 10000$$

As stated before the exponential distribution only applies to lifetimes of items that do not wear out or age. For all current light designs this is not the case: even modern led lights will not live eternally. In this sense the example we used above is futuristic. Nevertheless for some technical devices the (by far) main reason of malfunction are "reasons outside the device itself". One could think of a solar panel in space where damage is caused by, e.g. space grit.

Example 8.1.2 The exponentially distributed service time X in example 8.1.1 can be transferred to a discrete variable, as follows: instead of a the exact service time X we could only record in which minute Y the service ends. If the service is completed within one minute, $\{Y = 1\}$ is observed. And $\{Y = n\} = \{n - 1 \leq X < n\}$. Then for $n = 1, 2, 3, \ldots$ we have:

$$\begin{aligned} P(Y = n) &= P(n - 1 \leq X < n) \\ &= P(X \geq n - 1) - P(X \geq n) \\ &= e^{-(n-1)\lambda} - e^{-n\lambda} \\ &= \left(e^{-\lambda}\right)^{n-1}(1 - e^{-\lambda}). \end{aligned}$$

In this probability function we could recognize a geometric probability function $P(Y = n) = (1 - p)^{n-1}p$, where the success probability is apparently $p = 1 - e^{-\lambda} = P(X < 1)$: the probability that the service is completed within the first minute.
This probability function can also be derived from the lack of memory property of X.
We know that $P(X > 1) = e^{-\lambda}$. If the service has taken already n minutes (n is an arbitrary integer), then the probability that the service will not be completed in the next minute is $e^{-\lambda}$.
In formula:

$$P(X > n + 1|X > n) = P(X > 1) = e^{-\lambda}$$

And, applying the complement rule: $P(X \leq n + 1|X > n) = 1 - P(X > 1) = 1 - e^{-\lambda}$

Since $\{X > 2\} = \{X > 2 \text{ and } X > 1\}$ we have, applying the product rule $P(A \cap B) = P(A|B)P(B)$:

- $P(X > 2) = P(X > 2 \text{ and } X > 1) = P(X > 2|X > 1) \cdot P(X > 1) = P(X > 1)^2 = \left(e^{-\lambda}\right)^2$
- $P(X > 3) = P(X > 3 \text{ and } X > 2) = P(X > 3|X > 2) \cdot P(X > 2) = P(X > 1)^3 = \left(e^{-\lambda}\right)^3$
- Or, in general (using induction):
  $P(X > n) = P(X > n|X > n - 1) \cdot P(X > n - 1) = P(X > 1)^n = \left(e^{-\lambda}\right)^n$

So $P(Y = n) = P(X > n - 1) - P(X > n) = \left(e^{-\lambda}\right)^{n-1} - \left(e^{-\lambda}\right)^n = \left(e^{-\lambda}\right)^{n-1}(1 - e^{-\lambda})$

Let us compare $E(X)$ en $E(Y)$ for the case $\lambda = \frac{1}{2}$:

X is exponentially distributed with $\lambda = \frac{1}{2}$, so $E(X) = \frac{1}{\lambda} = 2$ and
Y is geometrically distributed with $p = 1 - e^{-\lambda} = 1 - e^{-\frac{1}{2}}$, so $E(Y) = \frac{1}{p} = \frac{1}{1-e^{-0.5}} \approx 2.54$.
We found: $E(Y) > E(X)$, which can be explained easily, since Y is the nearest integer larger than the real service time X. ■

In example 8.1.2 we established a relation between the exponential distribution and the geometric distribution. Moreover, the geometric distribution has the lack of memory property as well. For a geometrically distributed variable Y with $P(Y = n) = (1-p)^{n-1}p$, if $n = 1, 2, 3, ...$, we discussed, in chapter 4, the property $P(Y > n) = (1-p)^n, n = 0, 1, 2, ...$

$$\text{So: } P(Y > m + n | Y > n) = \frac{P(Y > m + n)}{P(Y > n)} = \frac{(1-p)^{m+n}}{(1-p)^n} = (1-p)^m = P(Y > m)$$

Note 8.1.3 Transferring from a continuous time X to a integer number of minutes Y was done by considering the minute in which the service is competed. The reverse transfer can also be made: instead of considering the minute of completion one could consider the second of completion, or the millisecond, etc. By choosing the unit of time smaller and smaller, the probability of completion within that unit will rapidly decrease. It can be shown that the distribution converges to a continuous, geometric distribution. ■

In the text above the memoryless property of the exponential and geometric distribution are explained. The formal definition of this property is given in the following definition.

Definition 8.1.4 The distribution of a random variable X has the **lack of memory property** on its range S_X, if for all $t, s \in S_X$:

$$\boldsymbol{P(X > t + s | X > s) = P(X > t)}.$$

In the examples we demonstrated the meaning of the lack of memory property, and the relation between the geometric and exponential distributions. The lack of memory property (memoryless-ness) is specific for these two distribution: the exponential distribution is the only continuous memoryless distribution and the geometric distribution is the only memoryless discrete distribution.
This is stated in the following summarizing properties (no formal proof is given).

Property 8.1.5 For a continuous random variable X with range $S_X = [0, \infty)$ the following statements are equivalent:

- **a.** X is exponentially distributed parameter λ.
- **b.** $P(X > t) = e^{-\lambda t}$, for $t \geq 0$.
- **c.** The distribution of X has the lack of memory property on S_X and $E(X) = \frac{1}{\lambda}$.

Property 8.1.6 For a discrete random variable X with range $S_X = \{1, 2, \dots\}$ the following statements are equivalent:

- **a.** X is geometrically distributed with parameter p.
- **b.** $P(X > n) = (1-p)^n$, for $n = 0, 1, 2, \dots$.
- **c.** The distribution of X has the lack of memory property on S_X and $p = P(X = 1)$.

8.2 Summation of independent waiting times

In this section we restrict ourselves to continuous waiting times, on which the exponential distribution applies.
Let us imagine a typical waiting time situation: at a counter of a post office a line of 10 persons should be served by an employee. The total service time can be modelled as the sum of 10 independent identically distributed service times $X_1, X_2, \ldots, X_{10}$, for which the exponential distribution applies, so $S = \sum_{i=1}^{10} X_i$.
A similar model can be used for the time needed to send 10 messages through a communication channel, or the first 10 interarrival times of visitors to a website.
Using the assumed exponential distributions and independence of the X_i's, we can apply the Convolution integral repeatedly to find that S has a so called Erlang distribution.

Definition 8.2.1 X has an **Erlang distribution with parameters n and λ,** if

$$f_X(x) = \frac{\lambda(\lambda x)^{n-1} e^{-\lambda x}}{(n-1)!}, \qquad \text{for } x \geq 0 \text{ and } f_X(x) = 0 \text{ for } x < 0$$

Short notation: $X \sim \mathbf{Erlang}(n, \lambda)$.

If $n = 1$, then $f_X(x) = \lambda e^{-\lambda x} (x \geq 0)$: X is exponentially distributed with parameter λ.
In chapter 7 (example 7.2.2) we showed that the sum of two independent $Exp(\lambda)$-distributed random variables has an Erlang distribution with $n = 2$.

In general we have:

Property 8.2.2 If $X_1, X_2, \ldots$ are independent and exponentially distributed with parameter λ, then:

$$S_n = \sum_{i=1}^{n} X_i \sim \mathbf{Erlang}(n, \lambda).$$

Proof (with induction):
The statement is true for $n = 1$.
Induction assumption: suppose S_n has an Erlang distribution with parameters n and λ.
We will have to show that $S_{n+1} = S_n + X_{n+1}$ has an Erlang distribution as well, with parameters $n + 1$ and λ.
Because of the independence of all X_i's $S_n = \sum_{i=1}^{n} X_i$ and X_{n+1} are independent (property 5.4.7) and the Convolution integral applies:
If $s < 0$, then $f_{S_{n+1}}(s) = 0$ (since $S_n \geq 0$ and $X_{n+1} \geq 0$) and if $s \geq 0$, we have

$$f_{S_{n+1}}(s) = \int_{-\infty}^{\infty} f_{S_n}(x)\, f_{X_{n+1}}(s-x)\,dx = \int_{0}^{s} \frac{\lambda(\lambda x)^{n-1} e^{-\lambda x}}{(n-1)!} \cdot \lambda e^{-\lambda(s-x)}\,dx$$

$$= \int_{0}^{s} \frac{\lambda^{n+1} x^{n-1} e^{-\lambda s}}{(n-1)!}\,dx = \frac{\lambda^{n+1} x^{n} e^{-\lambda s}}{n!}\Bigg|_{x=0}^{x=s} = \frac{\lambda^{n+1} s^{n} e^{-\lambda s}}{n!}$$

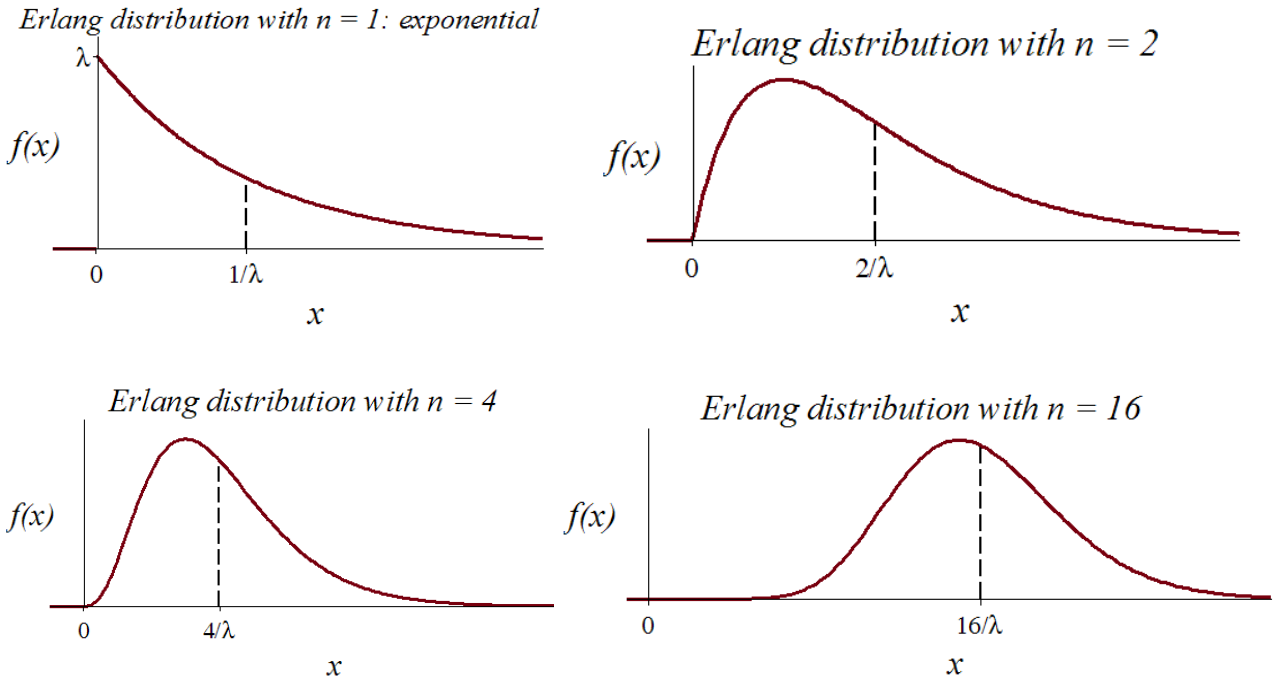
According to the definition of S_{n+1}, this is the Erlang density with parameters $n+1$ and λ. ■

Property 8.2.2 enables us to quickly find the expectation and variance of the Erlang distribution: no need to compute expectation and variance via the definitions:

$$\boldsymbol{E(S_n)} = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = n \cdot \frac{1}{\lambda} = \boldsymbol{\frac{n}{\lambda}} \quad \text{en}$$

$$\boldsymbol{var(S_n)} = var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) = n \cdot \frac{1}{\lambda^2} = \boldsymbol{\frac{n}{\lambda^2}}$$

Some graphs show the shapes of the Erlang distributions if $n = 1$, 2, 4 and 16:



If we want to compute probabilities with respect to the total waiting time (service time) S_n, e.g. the probability that a line of 10 customers will be served in 15 minutes, $P(S_{10} \leq 15)$, then we could find this probability $P(S_{10} \leq 15) = F_{S_{10}}(15)$, using the Erlang density.
In general we can find an expression for the distribution function $F_{S_n}(s)$ of the sum S_n:

if $X_1, X_2, \ldots, X_n$ are independent and $Exp(\lambda)$-distributed and $S_n = \sum_{i=1}^{n} X_i$, then:

$$F_{S_n}(s) = \int_{-\infty}^{s} f_{S_n}(x)dx = \int_{0}^{s} \frac{\lambda(\lambda x)^{n-1}e^{-\lambda x}}{(n-1)!}dx \quad (\textit{now we apply partial integration!})$$

$$= \frac{(\lambda x)^{n-1}}{(n-1)!} \cdot -e^{-\lambda x}\Big|_{x=0}^{x=s} + \int_{0}^{s} \frac{\lambda(\lambda x)^{n-2}e^{-\lambda x}}{(n-2)!}dx$$

$$= -\frac{(\lambda s)^{n-1}e^{-\lambda s}}{(n-1)!} + \int_{0}^{s} \frac{\lambda(\lambda x)^{n-2}e^{-\lambda x}}{(n-2)!}dx = \ldots \quad (\textit{repeat the partial integration})$$

In the end we find: $$F_{S_n}(s) = 1 - \sum_{k=0}^{n-1} \frac{(\lambda s)^k e^{-\lambda s}}{k!}$$

The terms of the summation on the right hand side reflect Poisson probabilities with parameter $\mu = \lambda s$, so:

$$F_{S_n}(s) = 1 - P(Y \leq n-1), \quad \text{where } Y \sim \text{Poisson}(\lambda s)$$

Example 8.2.3 Compute the probability $P(S_{10} \leq 15)$, that the total service time is at most 15 minutes if the service times (in min.) $X_1, X_2, \ldots, X_{10}$ are independent and exponentially distributed with parameter $\lambda = \frac{1}{2}$.

Solution: we have $n = 10$ and (the Poisson variable Y above) $\mu = \lambda s = \frac{1}{2} \cdot 15 = 7.5$, so:

$$P(S_{10} \leq 15) = F_{S_{10}}(15) = 1 - \sum_{k=0}^{9} \frac{7.5^k e^{-7.5}}{k!}$$

We search $P(Y \leq 9)$ in the Poisson table with parameter μ = 7.5 and find:

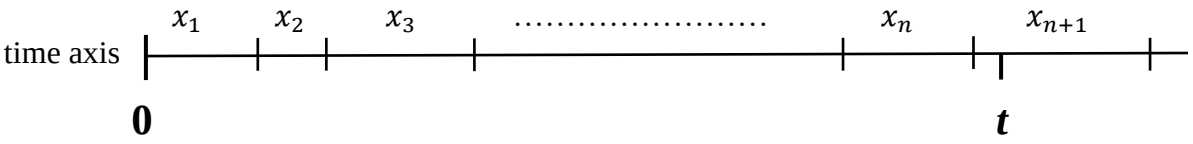
$$P(S_{10} \leq 15) = 1 - 0.776 = 22.4\%$$ ■

The distribution function of S_n and its relation with the Poisson distribution can be used to show that the **number of customers** being served within t minutes is Poisson distributed with parameter $\mu = \lambda t$.

Starting off with the usual assumption of a large series of independent $Exp(\lambda)$-distributed service times X_i`s, we are now interested in the number $N(t)$ of customers who are served during the interval $[0, t]$ of time, especially $P(N(t) = n)$, the probability that after t minutes exactly n customers are served.

This event occurs if the total service time of the first ***n* customers is at most *t* minutes**, but the total service time of the first ***n* + 1 customers is greater than *t* minutes**.

A sketch of this situation is given in the following diagram.



Note that the event $\{N(t) = n\}$ is **not** the same as $\{S_n \leq t\}$: if $S_n \leq t$, then $\{S_{n+1} \leq t\}$ could occur as well (in that case $N(t) > n$).

But the event $\{N(t) \geq n\}$ **is** the same as the event $\{S_n \leq t\}$. Since $P(N(t) \geq n) = P(S_n \leq t)$ and $P(N(t) = n) = P(N(t) \geq n) - P(N(t) \geq n + 1)$, we can state:

$$\begin{aligned} P(N(t) = n) &= P(S_n \leq t) - P(S_{n+1} \leq t) \\ &= 1 - \sum_{k=0}^{n-1} \frac{(\lambda t)^k e^{-\lambda t}}{k!} - \left(1 - \sum_{k=0}^{n} \frac{(\lambda t)^k e^{-\lambda t}}{k!}\right) \\ &= \frac{(\lambda t)^n e^{-\lambda t}}{n!} \end{aligned}$$

In conclusion: $N(t)$ is Poisson distributed with "mean" λt. Intuitively this mean is correct: the expected service time is $\frac{1}{\lambda}$, so in t minutes we expect to serve $\frac{t}{1/\lambda} = \lambda t$.
The mean number of served customers increases as λ increases: λ is the **intensity** of the service. In general, a process where service times or interarrival times are assumed to be independent and exponentially distributed, is referred to as a **Poisson process**, because of the Poisson distribution of the number of customers.

Property 8.2.4 If the interarrival times or service times of customers in a system are independent and exponentially distributed with parameter λ, then the number of arrivals (or served customers) $\boldsymbol{N(t)}$ **in the interval** $[\boldsymbol{0},\ \boldsymbol{t}]$ **Poisson distributed with parameter** $\boldsymbol{\mu = \lambda t}$.

Probabilities with respect to S_n, the sum of service times, interarrival times and lifetimes, should for small n (< 25) be computed via the Erlang distribution or related Poisson distribution (example 8.2.3), but for large n (≥ 25) we will use the CLT (see the graphs earlier in this section):

$$\boldsymbol{S_n \overset{\text{CLT}}{\sim} N\left(\frac{n}{\lambda}, \frac{n}{\lambda^2}\right)}$$

Example 8.2.5 Through a communication channel messages are sent: a message takes, on average, 1 millisecond. To avoid capacity problems in one second (= 1000 milliseconds) a fixed number of messages should be sent such that the probability of overload is less than 0.1%.
To meet this condition the "sending times" we will first assume that the times $X_1, X_2, \ldots, X_n$ (in milliseconds) of the n messages in a second are independent and exponentially distributed random variables with parameter $\lambda = 1$: $E(X_i) = 1$ millisecond and $var(X_i) = 1$ as well. The total sending time is denoted with $S_n = \sum_{i=1}^{n} X_i$: S_n is Erlang distributed with parameters n and λ = 1. According to the CLT (n shoud be close to 1000 and is sufficiently large) S_n is approximately $N(n, n)$-distributed $(\lambda = 1)$.
Suppose that we want to determine n such that:

$$P(S_n > 1000) < \frac{1}{1000}$$

Or:

$$P\left(\frac{S_n - n}{\sqrt{n}} > \frac{1000 - n}{\sqrt{n}}\right) < \frac{1}{1000}$$

In the standard normal table we find $\Phi(3.09) = 0.999$, so the inequality holds if:

$$\frac{1000 - n}{\sqrt{n}} > 3.09$$

After some computational manipulations (squaring and solving a quadratic equality, or by trying some values of n), we find that the maximum number of messages meeting the condition is n = 906. ■

8.3 Exercises

1. (former exam exercise)
The call rate for mobile phones of a provider is 15 cents for each period of 30 seconds or a part of it: e.g., for a telephone call of 70 seconds 3×15 cents is charged. The provider advertises that telephone costs are "about 30 cents per minute of calling".
We are going to check this statement by assuming that the duration X of a telephone call is exponentially distributed with mean 60 second ($\lambda = \frac{1}{60}$)
We define N to be the integer number of "ticks", the number of times 15 cents are charged:

$$P(N = n) = P[30(n - 1) < X \leq 30n], \quad \text{where } n = 1, 2, \ldots.$$

- **a.** Compute: the probability $P(X \geq 30)$, the mean duration $E(X)$ and the conditional probability $P(X \geq 90 | X \geq 60)$.
- **b.** Show that N is geometrically distributed with parameter $p = 1 - e^{-\frac{1}{2}} \approx 0.39$.
- **c.** Compute the mean charged amount per call $E(15N)$ and compare the result to the mean, advertised by the provider.
- **d.** Compute the variance of the charged amounts per call: $var(15N)$.

2. In the area of Traffic Studies at different points along the road the number of passing cars is counted. Usually the number (X) of cars per unit of time is assumed to be Poisson distributed. The parameter is linearly dependent on the chosen unit of time: if the period is t seconds, the expected number of passing cars equals $\mu = a \cdot t$, where a is a fixed constant (reflecting the number of passing cars per second, so if $t = 1$). Instead, one could also record the durations between two consecutive passing cars.
Given that X has a Poisson distribution with expectation $\mu = a \cdot t$, what can we say about the distribution of the duration Y?
In this exercise we will derive a formula for $P(Y > t)$, given arbitrary value $t > 0$.

- **a.** Argue why the event $\{Y > t\}$ is the same as the event $\{X = 0\}$.
- **b.** Give a formula for $P(X = 0)$ in terms of a and t.
- **c.** Determine $P(Y > t)$ for an arbitrary value t.
  Which distribution does this probability reflect?

3. The random behavior of complicated waiting time systems is sometimes statistically assessed by computer simulations. For that goal waiting times (service times) are generated, using random number generators, that produce random numbers between 0 and 1.
Assume X is such a random number (uniformly distributed on $(0, 1)$), then we can generate a random waiting time, having an exponential distribution with parameter $\lambda = 1$, by computing $Y = \ln\left(\frac{1}{X}\right)$.

- **a.** Show that Y has an exponential distribution with parameter $\lambda = 1$ (property 6.4.5).
- **b.** Verify whether $E(Y) = \ln\left(\frac{1}{E(X)}\right)$,

4. The times between two consecutive clients, logging on to a company`s computer system are considered to be independent and exponentially distributed. The mean time between two consecutive log on`s is 12 seconds (during office hours). We consider the log on`s during one minute: X_1 is the time (in seconds) from the start to the first log on, X_2 is the time between the first and second log on, etc.

- **a.** Compute $E(X_i)$, $var(X_i)$ and $E(\sum_{i=1}^{6} X_i)$
- **b.** Determine $P(X_1 > 12)$ and $P(X_1 > 15 | X_1 > 3)$.
- **c.** Give the distribution of $N =$ "the number of log on`s during a minute (60 seconds)" and compute $P(N \geq 6)$.
- **d.** Give the distribution of $\sum_{i=1}^{6} X_i$ (name and parameters) and compute $P(\sum_{i=1}^{6} X_i \leq 60)$

5. (former exam exercise)
$X_1, X_2, \ldots, X_n$ are independent waiting times: they all are exponentially distributed with parameter $\lambda = \frac{1}{4}$.
The sum of waiting times is $S_n = \sum_{i=1}^{n} X_i$ and the mean waiting time is $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

- **a.** Give (without proof) the distribution of S_n, $E(S_n)$ and $var(S_n)$.
- **b.** Derive the density function of S_2 from the density functions of X_1 and X_2. (Apply the convolution integral).
- **c.** Compute (for $n = 2$): $P(\overline{X}_2 > 5)$.
- **d.** Approximate (for $n = 100$): $P(\overline{X}_{100} > 5)$.

Some hints for the exercises of chapter 8:

1. a. Apply the "lack of memory property" on the conditional probability.

2. Use $P(Y > t) = P(X = 0)$.

3. Compare, if necessary, with the given derivation in 6.4.5.

4. What does $N \geq 6$ mean for the values of $\sum_{i=1}^{6} X_i$? Use this in d. as to avoid integration of the Erlang density function.

5. c. Use $\overline{X}_2 = \frac{S_2}{2}$

d. Choose either to use the approximating distribution of the mean or express the requested probability in S_{100} (and approximate the distribution of S_{100}).

Mathematical Techniques for Probability Theory

(for more details consult your Calculus book)

Series

1. **Newton`s Binomial Theorem:**

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k\, b^{n-k}$$

Application: the summation of binomial probabilities is 1:

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

2. **Geometric series:**

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$$

Application: the summation of geometric probabilities is 1:

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

Derivative of the geometric series (w.r.t. x):

$$\sum_{k=1}^{\infty} k x^{k-1} = \frac{1}{(1-x)^2}$$

Application: the expectation of the geometric distribution is $1/p$:

$$E(X) = \sum_{i=1}^{\infty} i \cdot (1-p)^{i-1}p \overset{k=i}{=} p \cdot \sum_{k=0}^{\infty} k \cdot (1-p)^{k-1} = p \cdot \frac{1}{\big(1-(1-p)\big)^2} = \frac{1}{p}$$

Finite geometric series:

$$\sum_{k=0}^{N} x^k = \frac{1-x^{N+1}}{1-x}$$

Application: finite summation of geometric probabilities:

$$P(X \le 10) = \sum_{i=1}^{10} (1-p)^{i-1}p \overset{k=i-1}{=} p \cdot \sum_{k=0}^{9} (1-p)^k = p \cdot \frac{1-(1-p)^{10}}{1-(1-p)}$$

3. **Taylor-series of a function $f(x)$ at $x = 0$**:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k, \qquad \text{applied to } f(x) = e^x\text{: } \; e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Application: the summation of probabilities and expectation of the Poisson distribution:

$$\sum_{k=0}^{\infty} \frac{\mu^k}{k!} \cdot e^{-\mu} = e^{\mu} \cdot e^{-\mu} = 1 \quad \text{and}$$

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

Differentiation and Integration

4. **Chain rule:**

$$\frac{d}{dx}\big[f\big(g(x)\big)\big] = f'\big(g(x)\big) \cdot g'(x)$$

Product rule:

$$\frac{d}{dx}[f(x)g(x)] = f'(x) \cdot g\,(x) + f(x) \cdot g'(x)$$

5. **Fundamental Theorem of Algebra:**

$$\int_a^b f(x)dx = F(b) - F(a), \quad \text{where } F \text{ is an anti} - \text{derivative of } f \text{, so } F'^{(x)} = f(x)$$

Application: computing probabilities for a continuous variable X with density f_X and distribution function F_X: $P(a < X \leq b) = \int_a^b f_X(x)dx = F_X(b) - F_X(a)$

6. **Integration by parts:**

$$\int_a^b f(x)g'(x)dx = [f(x)g(x)]_{x=a}^{x=b} - \int_a^b f'(x)g(x)dx$$

Application: computation of the expectation of the exponential distribution:

$$E(X) = \int_0^{\infty} x \cdot \lambda e^{-\lambda x}dx \text{, where } f(x) = x \text{ and } g'(x) = \lambda e^{-\lambda x} \text{and } g\,(x) = -e^{-\lambda x},$$

$$\text{So: } E(X) = \int_0^{\infty} x \cdot \lambda e^{-\lambda x}dx = \big[x \cdot -e^{-\lambda x}\big]_{x=0}^{x\to\infty} + \int_0^{\infty} e^{-\lambda x}dx = 0 - \frac{1}{\lambda}e^{-\lambda x}\Big|_{x=0}^{\infty} = \frac{1}{\lambda}$$

List of Probability concepts in Dutch

Names of distributions are in general the same, such as: binomial, geometric and exponential distributions are “binomiale, geometrische en exponentiële verdelingen”, etc.

(Bernoulli) trial	(Bernoulli) poging of experiment
Central Limit Theorem	Centrale Limiet Stelling
conditional distribution	voorwaardelijke verdeling
conditional probability	voorwaardelijke kans
correlation (coefficient)	correlatie(coëfficiënt)
covariance	covariantie
disjoint	disjunct, elkaar uitsluitend
distribution function	verdelingsfunctie
event	gebeurtenis
expectation	verwachting
expected value	verwachtingswaarde
independent	(onderling) onafhankelijk
joint distribution	simultane verdeling
mean	steekproefgemiddelde òf verwachting
marginal distribution	marginale verdeling
mode	modus
mutually exclusive	elkaar uitsluitend
population mean	Verwachting, populatiegemiddelde
population proportion	populatiefractie
(probability) density function	kansdichtheid
(probability) distribution	(kans)verdeling
probability (mass) function	kansfunctie
probability (measure)	kans(maat)
sample	aselecte steekproef
random variable	stochastische variabele
sample mean	steekproefgemiddelde
sample proportion	steekproeffractie
sample size	steekproefuitgebreidheid
sample space	uitkomstenruimte
simple event	elementaire gebeurtenis
standard deviation	Standaardafwijking, standaarddeviatie
variance	variantie
Weak Law of Large Numbers	Zwakke wet van grote aantallen

Answers to exercises

Chapter 1

1. **a.** $AB\overline{C}$ **b.** ABC **c.** $A \cup B \cup C$ **d.** $AB \cup AC \cup BC$
e. $\overline{A}\,\overline{B}\,\overline{C} = \overline{A \cup B \cup C}$ **f.** $A\overline{B}\,\overline{C} \cup \overline{A}B\overline{C} \cup \overline{A}\,\overline{B}C$ **g.** $\overline{ABC} = \overline{A} \cup \overline{B} \cup \overline{C}$ (De Morgan)

2. $\frac{600}{1200}$

3. The probabilities are not equal: the probability of "one Head and one Tail" is e.g. ½.

4. Not correct, the probability is $1 - 0.98^{20} \approx 33.2\%$

5. $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$

6. **a.** 1 **b.** 2 **c.** not possible

7. $A \subset B$ means $B = A \cup (\overline{A}B)$, where A and $\overline{A}B$ are mutually exclusive, so (axiom 3):
$P(B) = P\left(A \cup \overline{A}B\right) = P(A) + P\left(\overline{A}B\right)$.
Since $P\left(\overline{A}B\right) \geq 0$ (axiom 1), we have: $P(B) \geq P(A)$

8. $P(B) = \frac{13}{18}$ and $P\left(\overline{A} \cap \overline{B}\right) = \frac{1}{9}$

Chapter 2

1. **a.** $7! = 5040$ **b.** $10 \cdot 9 \cdot 8 \cdot \ldots \cdot 5 \cdot 4 = \frac{10!}{3!} = 604800$ **c.** $\binom{10}{3} = \frac{10!}{3!7!} = 120$
d. $1 - \frac{\binom{36}{13} + 16 \cdot \binom{36}{12}}{\binom{52}{13}} \approx 1 - 0.0352 = 96.48\%$ **e.** $\binom{30}{6} \cdot \binom{24}{7} \cdot \binom{17}{8} \cdot \binom{9}{9} = \frac{30!}{6!7!8!9!}$

2. $\frac{169}{1000} = \frac{8^3 - 7^3}{10^3}$

3. $\frac{3 \cdot 9 \cdot 8 \cdot 7}{10 \cdot 9 \cdot 8 \cdot 7} = \frac{3}{10}$

4. **a.** $\frac{\binom{16}{5}}{\binom{52}{5}} = \frac{16!/11!}{52!/47!} \approx 0.0017$ **b.** $\frac{4 \cdot 4 \cdot (50 \cdot 49 \cdot 48)}{52!/47!} \approx 0.0060$
c. $\frac{\binom{4}{1}\binom{4}{1}\binom{44}{3}}{\binom{52}{5}} = \frac{5 \cdot 4 \cdot 4 \cdot 4 \cdot (44 \cdot 43 \cdot 42)}{52!/47!} \approx 0.0815$

5. $\frac{\binom{6}{4} + 1}{6!} \approx 0.0222$

6. **a.** $1 - \frac{\binom{90}{10}}{\binom{100}{10}} \approx 0.6695$ **b.** $\frac{\binom{80}{10}}{\binom{100}{10}} \approx 0.0951$

7. **a.** $\frac{\binom{4}{1}\cdot\binom{10}{1}\cdot\binom{86}{3}}{\binom{100}{5}} \approx 0.0544$ **b.** $\frac{\binom{14}{1}\cdot\binom{86}{4}}{\binom{100}{5}} \approx 0.3949$ **c.** $\frac{\binom{86}{5}}{\binom{100}{5}} \approx 0.4626$ **d.** 0.5374

8. $\frac{\binom{N_1}{n_1}\cdot\binom{N_2}{n_2}\cdot\ldots\cdot\binom{N_k}{n_k}}{\binom{N}{n}}$

9. **a.** $2\cdot\frac{\binom{6}{4}\cdot\binom{20}{9}}{\binom{26}{13}} \approx 0.4845$ **b.** $\frac{\binom{6}{3}\cdot\binom{20}{10}}{\binom{26}{13}} \approx 0.3553$

10. **a.** $\binom{13}{6}$ **b.** $\binom{30}{6}\cdot\binom{24}{7}\cdot\binom{17}{8}\cdot\binom{9}{9} = \frac{30!}{6!7!8!9!}$

c. $\binom{n}{n_1}\cdot\binom{n-n_1}{n_2}\cdot\ldots\cdot\binom{n_k}{n_k} = \frac{n!}{n_1!\,n_2!\,\ldots n_k!}$.

11. **a.** $\frac{\binom{7}{2}+\binom{7}{3}}{\binom{11}{5}} \approx 0.1212$ **b.** $\frac{\binom{4}{2}\binom{5}{1}+\binom{5}{2}\binom{4}{1}}{\binom{11}{5}} = \frac{70}{462} \approx 0.1515$

12. $\frac{\binom{10}{4}}{10!/6!} = \frac{1}{4!} = \frac{1}{24}$

Chapter 3

1. **a.** $\frac{0.30}{0.75} = \frac{2}{5}$ (= 40%) **b.** $\frac{3}{10}$

2.

a. $P(B) = 0.97*0.98 + 0.05*0.02 = 0.9516$.

b. $P(\overline{A} \mid \overline{B}) = 39.3\%$.

c. A and B are dependent.

3. $\frac{0.75\cdot0.05}{0.75\cdot0.05+0.02\cdot0.95} \approx 0.6637$

4. $P(\text{Comium}) = 0.1$

5. $\frac{2}{3}$

6. **a.** $\frac{\binom{5}{3}\cdot\binom{7}{2}}{\binom{12}{5}} \approx 0.2652$ **b.** $\sum_{i=3}^{6}\frac{\binom{5}{3}\cdot\binom{7}{i-3}}{\binom{12}{i}}\cdot\frac{1}{6} \approx 0.1385$

7. $\frac{3}{4}$

8. **a.** $P(A\overline{B}) = P(A) - P(AB) = P(A) - P(A)P(B) = P(A)[1 - P(B)] = P(A) \cdot P(\overline{B})$
A and $\overline{B}$ are independent as well, from which it follows that $\overline{A}$ and $\overline{B}$ are independent.
b. $P(A \cap BC) = P(ABC) = P(A)P(B)P(C) = P(A) \cdot P(BC)$.

9. Only possible if $P(A) = 0$ or $P(B) = 0$.

10. $\left(\frac{5}{6}\right)^6 \approx 33.49\%$

11. $\sum_{k=10}^{12}\binom{12}{k}\left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{12-k} \approx 0.054\%$

12. **a.** $P(B) = 0.997 \cdot 0.01 + 0.015 \cdot 0.99 = 0.02482$ **b.** 40.2%

Chapter 4

1. **a**. $P(X > 0) = 0.5$ **b.** $E(X) = -0.2$ **c.** $E(X^2) = 4.6$ **d**. $var(X) = 4.56$ and $\sigma_X \approx 2.14$

2. $P(X = k) = \frac{1}{10}, k = 0, 1, 2, \ldots, 9$, $E(X) = 4.5, E(X^2) = 28.5$ and $var(X) = 8.25$

3. 487

4. $-\frac{17}{216} \approx -0.0787$

5. **a.** $P(X = 7) = \frac{\binom{6}{3}}{\binom{10}{4}}$ **b**. $P(X = k) = \frac{\binom{k-1}{n-1}}{\binom{N}{n}}$, $k = n, n+1, \ldots, N$

6. **a**. $P(X \leq 7) = 0.998$ **b.** $P(X \geq 7) = 0.011$
c. $P(X = 9) = 20.7\%$ **d**. $P(X < 12) = 0.909$

7. **a**. $P(X = 5) = 0.101$ **b**. $P(X < 2) = 0.199.$ **c**. $P(X > 3) = 0.353$

8.
a. Poisson($\mu = 5$), $P(X = 2) \approx 0.0842$. $E(X) = \mu = 5$.
b. Hypergeometric distribution: $P(X = 2) = \frac{\binom{2}{2}\binom{3}{0}}{\binom{5}{2}} = \frac{1}{10}$ and $E(X) = \frac{4}{5}$
c. $X \sim B(100, 0.02)$, $P(X = 2) \approx 27.3\%$ and $E(X) = np = 2$.
d. $X \sim geometric\left(p = \frac{1}{10}\right), P(X = 2) = 0.09$ and $E(X) = 10$
e. X has a homogeneous distribution on {1,2,..,10}, so $P(X = 2) = \frac{1}{10}$ and $E(X) = 5.5$

9. **a**. 0.277 **b**. 0.287

10. **a.** $E(X^k) = \frac{1}{2}c^k$ **b**. $E(X) = \frac{1}{2}c$ en $E(X^2) = \frac{1}{2}c^2$, $var(X) = \frac{1}{4}c^2$

11. a. $c = \frac{2}{3}$ **b.** $Y = X + 1$ is geometric with $p = \frac{2}{3}$ **c**. $E(X) = \frac{1}{2}$ and $var(X) = \frac{3}{4}$

12. a. $B\left(150, \frac{1}{50}\right)$-distribution. **b.** $c = 6$ lines.

13. a. $B(12, 0.15)$-distribution; $P(X > 1.8) \approx 0.5565$
b. Poisson distribution with parameter $\mu = 2$: $P(X > 2) \approx 32.33\%$
c. Hypergeometric distribution with parameters $N = 10, R = 3$ and $n = 4$:
$P(X > 1.2) = \frac{1}{3}$

14. a. $1 - (0.96)^{10} \approx 0.3352$.
b. Approximation with the Poisson distribution($\mu = 4$): 0.567 (exact: 0.5705) **c**. 2.984

15. a. $\binom{9}{2}\left(\frac{5}{6}\right)^7\left(\frac{1}{6}\right)^3 \approx 4.65\%$
b. $S_X = \{m, m+1, \ldots.\}$ and $P(X = k) = \binom{k-1}{m-1}(1-p)^{k-m}p^m$, with $k \in S_X$.

16. a. $M = 2$ **b**. $M = 2$ **c.** $M = 2$ **d**. $M \in [3, 4]$

Chapter 5

1. b. $P(X = 1|Y = 1) = \frac{1}{4}$, $P(X = 2|Y = 1) = \frac{1}{2}$ and $P(X = 3|Y = 1) = \frac{1}{4}$.
So $E(X|Y = 1) = 2$.
c. $P(Y = 1|X = 3) = \frac{1}{2}$

2. b. $P(X = i) = \left(\frac{1}{3}\right)^{i-1}\frac{2}{3}$, if $i = 1, 2, 3, \ldots$, so $X \sim geometric\left(p = \frac{2}{3}\right)$ and $E(X) = \frac{1}{p} = \frac{3}{2}$
c. $P(Y = 0) = \frac{1}{3}$ and $P(Y = 1) = \frac{2}{3}$ ($P(Y = j) = \left(\frac{1}{3}\right)^{1-j}\left(\frac{2}{3}\right)^j$, if $j = 0,1$)
d. Yes.

3.

a. Distribution of X: add row probabilities.
Distribution of Y: add column probabilities.

i	0	1	2	Total
$P(X = i)$	0.2	0.5	0.3	1

j	0	1	2	3	Total
$P(Y = j)$	0.10	0.30	0.45	0.15	1

b. $E(X) = 1.1$ and $var(X) = 0.49$, $E(Y) = 1.65$ and $var(Y) = 0.7275$
c. For $Z = 8Y$ we have $E(Z) = 8 \cdot E(Y) = 8 \cdot 1.65 = 13.20$ and
$var(8Y) = 8^2 var(Y) = 64 \cdot 0.7275 = 46.56$.
d. The values of $T = X + Y$ run from 0 to 5 (add probabilities "diagonally")

t	0	1	2	3	4	5	Total
$P(T = t)$	0.05	0.10	0.20	0.40	0.20	0.05	1

$E(T) = 2.75$ and $var(T) = 1.3875$
e. $var(T) = var(X + Y)$ does not correspond with $var(X) + var(Y) = 0.490 + 0.7275 = 1.2175$, which is caused by the dependence of X and Y.

f. $var(T) = var(X+Y) = 1.3875$ does not correspond with $var(X) + var(Y) = 0.490 + 0.7281 = 1.218$, because of the dependence of X an Y.

4. **a.** $P(N=10) = \left(\frac{1}{2}\right)^{10}$, $P(X=4|N=10) = \binom{10}{4}\left(\frac{1}{2}\right)^{10}$ and

$P(X=4 \text{ and } N=10) = \binom{10}{4}\left(\frac{1}{4}\right)^{10}$

b. $P(N=n) = \left(\frac{1}{2}\right)^{n}$, $n = 1,2,\ldots$

$P(X=x|N=n) = \binom{n}{x}\left(\frac{1}{2}\right)^{n}$, $x = 0,1,2,\ldots,n$

$P(X=x \text{ and } N=n) = \binom{n}{x}\left(\frac{1}{4}\right)^{n}$, $x = 0,1,2,\ldots,n$ en $n = 1,2,\ldots$

c. 5, $\frac{1}{2}n$, $E(X) = E[E(X|N)] = E\left(\frac{1}{2}N\right) = \frac{1}{2}E(N) = \frac{1}{2}\cdot 2 = 1$

d. $P(X=0) = \frac{1}{3}$

e. N is, given $X=0$, geometrically distributed with parameter $p = \frac{3}{4}$,
so $E(N|X=0) = \frac{4}{3}$

5. **a.** If $N=n$, then $S = X_1 + \cdots + X_n$

b. $E(X_i) = \sum_x x \cdot P(X_i = x) = 1000\cdot\frac{1}{10} + 2000\cdot\frac{3}{10} + 3000\cdot\frac{4}{10} + 4000\cdot\frac{2}{10} = 2700$

c. $E(S|N=n) = E(X_1 + \cdots + X_n) = 2700n$

d. $E(S) = E[\,E(S|N)] = E(2700N) = 2700\cdot E(N) = 2700\mu$

6. **a.** $P(X=8 \text{ and } Y=2) = 0.033$.

b. $E(Y|X=8) = 2.4$ and $E(Y|X=x) = 0.3x$.

c. $E(Y) = E[\,E(Y|X)] = E[\,0.3X] = 0.3\cdot E(X) = 3$

7. **a.** $P(X_1 = 10) = 0.9^9\cdot 0.1 = 3.87\%$

b. Use the property $P(X>x) = (1-p)^x$ of the geometric distribution:
$P(20 \le X_1 \le 30) = P(X_1 > 19) - P(X_1 > 30) = 0.9^{19} - 0.9^{30} \approx 9.27\%$

c. $P(X_1 = X_2) \approx 5.26\%$

d. $P(X_1 + X_2 = 20) \approx 2.85\%$

8. **a.** $P(X>i \text{ and } Y>i) \overset{\text{ind}}{=} P(X>i)P(Y>i) = (1-p)^i\cdot(1-p)^i = [(1-p)^2]^i$

b. $P(min(X,Y) > i) = P(X>i \text{ en } Y>i) = [(1-p)^2]^i$, if $i = 0,1,2,\ldots$

c. $P(min(X,Y) = i) = P(min(X,Y) > i-1) - P(min(X,Y) > i)$
$= [(1-p)^2]^{i-1} - [(1-p)^2]^i = [(1-p)^2]^{i-1}[1-(1-p)^2]$
if $i = 1,2,\ldots$

d. So $min(X,Y)$ has a geometric distribution with parameter $1-(1-p)^2 = 2p-p^2$
$\Longrightarrow E[min(X,Y)] = \frac{1}{2p-p^2}$

9. **a.** $E(X+Y) = E(X) + E(Y) = \frac{2}{p}$ en $var(X+Y) \overset{\text{ind.}}{=} var(X) + var(Y) = 2\cdot\frac{1-p}{p^2}$

b. $P(X=i \text{ and } Y=j) \overset{\text{ind.}}{=} P(X=i)\cdot P(Y=j) = (1-p)^{i-1}p\cdot(1-p)^{j-1}p$
$= (1-p)^{i+j-2}p^2 \qquad (i,j = 1,2,\ldots)$

c. $P(X+Y=n) = \sum_{i=1}^{n-1}(1-p)^{i-1}p\cdot(1-p)^{n-i-1}p = \sum_{i=1}^{n-1}(1-p)^{n-2}p^2$
$= (n-1)\cdot(1-p)^{n-2}p^2, \quad n = 2,3,4,\ldots$

10. **a**. X and Y have the same distribution (in the table).

x	0	1	2	Total
$P(X=x)$	0.3	0.4	0.3	1
$x \cdot P(X=x)$	0	0.4	0.6	$1 = E(X)$
$x^2 \cdot P(X=x)$	0	0.4	1.2	$1.6 = E(X^2)$

$E(X) = 1$ and
$var(X) = E(X^2) - \mu^2 = 1.6 - 1 = 0.6$

b. Respectively $+\frac{2}{3}$, 0, 0 and -1.

c. Only in distribution 3 X and Y are independent.
$P(X = i \text{ and } = j) = P(X = i) \cdot P(Y = j)$, for all (i, j)

d. $E(XY) = \sum_i \sum_j \; i \cdot j \cdot P(X = i \text{ and } Y = j) = 1 \cdot 0.2 + 2 \cdot 0.1 + 2 \cdot 0.1 + 4 \cdot 0.2 = 1.4$
$cov(X,Y) = E(XY) - E(X)E(Y) = 1.4 - 1 \cdot 1 = 0.4$
$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{0.4}{\sqrt{0.6}\sqrt{0.6}} = +\frac{2}{3}$

e. $cov(3X, 2 - Y) = 3 \cdot -1 \cdot cov(X,Y) = -3 \cdot 0.4 = -1.2$,
so $\rho(3X, 2 - Y) = -\rho(X,Y) = -\frac{2}{3}$

f. $E(XY) = 1$, so $cov(X,Y) = 0$ and $\rho(X,Y) = 0$.

g. $E(X|Y=0) = \frac{1}{3}$, $E(X|Y=0) = 1$ and $E(X|Y=2) = \frac{5}{3}$.

11. $\rho(X,Y) = \rho(X, -3X+4) = \frac{cov(X,-3X+4)}{\sigma_X \sigma_{-3X+4}}$

Since $var(-3X+4) = (-3)^2 \cdot var(X) = 9 \cdot var(X)$, so $\sigma_{-3X+4} = 3\sigma_X$, and
$cov(X, -3X+4) = -3 \cdot cov(X,X) = -3 \cdot var(X) = -3\sigma_X^2$,
we have $\rho(X,Y) = \frac{-3\sigma_X^2}{\sigma_X \cdot 3\sigma_X} = -1$

12. **a**. $cov(X_1, X_1 + X_2) = cov(X_1, X_1) + cov(X_1, X_2) \overset{\text{ind.}}{=} var(X_1) + 0 = 2$
$\rho(X_1, X_1 + X_2) = \frac{cov(X_1 X_1 +, X_2)}{\sigma_{X_1} \sigma_{X_1 + X_2}} = \frac{var(X_1)}{\sqrt{2} var(X_1)} = \frac{1}{\sqrt{2}}$

b. $\rho(X_1, X_1 + X_2 + \ldots + X_n) = \frac{1}{\sqrt{n}} < \frac{1}{3}$, if $n > 9$

13. **a**. $E(X_1) = 1 \cdot \frac{1}{10} + 0 \cdot \frac{9}{10} = \frac{1}{10}$, $E\left({X_1}^2\right) = \frac{1}{10}$ and $var(X_1) = \frac{1}{10} - \left(\frac{1}{10}\right)^2 = \frac{9}{100}$
$cov(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2) = 1 \cdot 1 \cdot \frac{1}{10} \cdot \frac{1}{9} - \left(\frac{1}{10}\right)^2 = \frac{1}{900}$

b. $E(S) = E\left(\sum_{i=1}^{10} X_i\right) = \sum_{i=1}^{10} E(X_i) = 10 \cdot E(X_1) = 1$
$var(S) = var\left(\sum_{i=1}^{10} X_i\right) = \sum_{i=1}^{10} var(X_i) + \sum \; \sum_{i \neq j} \; cov(X_i, X_j)$
$= 10 \cdot var(X_1) + 90 \cdot \; cov(X_1, X_2) = 10 \cdot \frac{9}{100} + 90 \cdot \frac{1}{900} = 1$

14. **a**.$X + Y \sim Poisson(\mu_1 + \mu_2)$, where $\mu_1 + \mu_2 = 2 + 3$

b. $P(X = k | X + Y = n) = \frac{P(X = x \text{ and } Y = n-k)}{P(X+Y=n)} = \frac{\frac{2^k e^{-2}}{k!} \cdot \frac{3^{n-k} e^{-3}}{(n-k)!}}{\frac{5^n e^{-5}}{n!}} = \frac{n!}{k!(n-k)!} \left(\frac{2}{5}\right)^k \left(\frac{3}{5}\right)^{n-k}$

(where $\mu_1 = 2$ and $\mu_2 = 3$). For $k = 0, 1, 2, \ldots, n$ this is the binomial distribution)

c. $E(X | X + Y = 7) = 7 \cdot \frac{2}{2+3} = 2.8$, since X is, given $X + Y = n$, $B\left(n, \frac{2}{5}\right)$-distributed

Assumptions: X and Y, the numbers of cases of appendicitis and kidney stones resp. are independent and have Poisson-distributions with parameters 2 resp. 3.

Chapter 6

1. **a.** $P(X>1)=\int_1^{\infty} f(x)dx=\int_1^2\left(1-\frac{1}{2}x\right)dx=x-\frac{1}{4}x^2\Big|_{x=1}^{x=2}=1-\frac{3}{4}=\frac{1}{4}$,
(or graphically: determine the area of the triangle: $\frac{1}{2}\cdot 1\cdot\frac{1}{2}=\frac{1}{4}$)

b. $E(X)=\int_0^2 x\cdot\left(1-\frac{1}{2}x\right)dx=\frac{1}{2}x^2-\frac{1}{6}x^3\Big|_{x=0}^{x=2}=\frac{2}{3}$

$E(X^2)=\int_0^2 x^2\cdot\left(1-\frac{1}{2}x\right)dx=\frac{1}{3}x^3-\frac{1}{8}x^4\Big|_{x=0}^{x=2}=\frac{2}{3}$

$var(X)=E(X^2)-(EX)^2=\frac{2}{3}-\left(\frac{2}{3}\right)^2=\frac{2}{9}$

c. $F(x)=P(X\le x)$, so $F(x)=0$ if $x<0$ and $F(x)=1$ if $x>2$.

If $0\le x\le 2$, then: $F(x)=\int_0^x\left(1-\frac{1}{2}u\right)du=u-\frac{1}{4}u^2\Big|_{u=0}^{u=x}=x-\frac{1}{4}x^2$

So $P(X>1)=1-F(1)=1-\frac{3}{4}=\frac{1}{4}$

2. **a**. $f(x)=\lambda e^{-\lambda x}$, if $x\ge 0$ and $f(x)=0$, if $x<0$

$\int_{-\infty}^{\infty}f(x)\,dx=\int_0^{\infty}\lambda e^{-\lambda x}\,dx=-e^{-\lambda x}|_0^{\infty}=-0-(-1)=1$ (graph: see page 6-12).

b. $E(X)=\int_0^{\infty}x\cdot\lambda e^{-\lambda x}\,dx=x\cdot -e^{-\lambda x}|_0^{\infty}+\int_0^{\infty}e^{-\lambda x}\,dx=-\frac{1}{\lambda}e^{-\lambda x}|_0^{\infty}=+\frac{1}{\lambda}$

$P(X>EX)=\int_{1/\lambda}^{\infty}\lambda e^{-\lambda x}\,dx=-e^{-\lambda x}|_{\frac{1}{\lambda}}^{\infty}=e^{-\lambda\cdot\frac{1}{\lambda}}=e^{-1}\approx 36.8\%\ (<\frac{1}{2})$

c. $P(X>M)=\int_M^{\infty}\lambda e^{-\lambda x}\,dx=-e^{-\lambda x}|_M^{\infty}=e^{-\lambda M}=\frac{1}{2}$, so $M=\frac{\ln(2)}{\lambda}$

d. The mode = 0 (see graph).

3. **a.** - $f(x)=\frac{1}{4}$, if $0<x<4$: $\int_{-\infty}^{\infty}f(x)\,dx=4\cdot\frac{1}{4}=1$

- $E(X)=2=$ median (because of f`s symmetry). So $P(X>EX)=P(X\le M)=\frac{1}{2}$

- mode: all values in $[0,4]$

b. $F_X(x)=\begin{cases}0 & \text{if} \quad x\le 0\\ \frac{x}{4} & \text{if } 0<x<4\\ 1 & \text{if} \quad x\ge 4\end{cases}$

4. **a**. $\int_{-\infty}^{\infty}f(x)\,dx=\int_1^{\infty}\frac{c}{x^3}dx=-c\cdot\frac{1}{2}x^{-2}|_1^{\infty}=c\cdot\frac{1}{2}=1$, so $c=2$

$P(X>2)=\int_2^{\infty}\frac{2}{x^3}dx=-2\cdot\frac{1}{2}x^{-2}|_2^{\infty}=\frac{1}{4}$

b. $E(X)=\int_1^{\infty}x\cdot\frac{2}{x^3}dx=-2x^{-1}|_1^{\infty}=2$

$P(X>m)=\int_m^{\infty}\frac{2}{x^3}dx=-2\cdot\frac{1}{2}x^{-2}|_m^{\infty}=\frac{1}{m^2}=\frac{1}{2}$, so $m=\sqrt{2}$

c. $F(x)=\int_1^x\frac{2}{u^3}du=-u^{-2}|_1^x=1-\frac{1}{x^2}$, if $x\ge 1$, and $F(x)=0$ if $x<1$

5. **a**. 1. $F_Y(y)=P(5-2X\le y)=P(-2X\le y-5)=P\left(X\ge\frac{y-5}{-2}\right)=1-F_X\left(\frac{y-5}{-2}\right)$

2. $f_Y(y)=\frac{d}{dy}F_y(y)=\frac{1}{2}f_X\left(\frac{y-5}{-2}\right)$

3. $f_X\left(\frac{y-5}{-2}\right)=1$ if $0<\frac{y-5}{-2}<1$, so $-2<y-5<0$ or: $3<y<5$

$$f_Y(y) = \begin{cases} \frac{1}{2} \cdot 1, & \text{if } 3 < y < 5 \\ 0, & \text{elsewhere} \end{cases}, \text{ so } Y \sim U(3,5)$$

b. Choose $Y = a + (b-a) \cdot X$ (or $Y = b - (b-a)X$ as in a.)

c. $f_Y(y) = \frac{3}{2} e^{-\frac{3}{2}y}$ for $y \geq 0$ and $f_Z(z) = \frac{3}{2\sqrt{z}} e^{-3\sqrt{z}}$ for $z > 0$.

6. **a**. 1. $F_Y(y) = P\left(\frac{1}{X} \leq y\right) = P\left(X \geq \frac{1}{y}\right) = 1 - F_X\left(\frac{1}{y}\right)$, $y > 0$

(and $F_Y(y) = P\left(\frac{1}{X} < y\right) = 0$, if $y < 0$)

2. $f_Y(y) = \frac{d}{dy} F_y(y) = -\frac{1}{y^2} \cdot -f_X\left(\frac{1}{y}\right)$

3. $f_X\left(\frac{1}{y}\right) = 1$ if $\frac{1}{y} > 0$, so if $y > 1 \rightarrow f_Y(y) = \frac{1}{y^2} \cdot 1 = \frac{1}{y^2}$ if $y > 1$

b. $P(Y > 2) = \int_2^\infty f_Y(y)\, dy = \int_2^\infty \frac{1}{y^2} dy = -y^{-1}|_2^\infty = \frac{1}{2}$

$P(Y > 2) = P\left(\frac{1}{X} > 2\right) = P\left(X < \frac{1}{2}\right) = \frac{1}{2}$

c. $E(Y) = \int_{-\infty}^\infty y f_Y(y) dy = \int_1^\infty y \cdot \frac{1}{y^2} dy = \ln(y)\,|_1^\infty = \infty$, so $E(Y)$ does not exist.

And $E(Y) = E\left(\frac{1}{X}\right) = \int_{-\infty}^\infty \frac{1}{x} f_X(x) dx = \int_0^1 \frac{1}{x} dx = \ln(x)\,|_0^1$ exists neither.

7. **a**. if $y > 0$ we have: $F_Y(y) = P\left(\sqrt{|X|} \leq y\right) = P(|X| \leq y^2) = P(-y^2 \leq X \leq y^2)$

$= F_X(y^2) - F_X(-y^2)$

Since $F_X(x) = 1 - e^{-x}$, if $x > 0$ and $F_X(x) = 0$ if $x < 0$, we have

$F_Y(y) = \left(1 - e^{-y^2}\right) - 0 = 1 - e^{-y^2}$, if $y > 0$.

b. $f_Y(y) = \frac{d}{dy} F_y(y) = 2ye^{-y^2}$, if $y > 0$

$E(Y) = \int_{-\infty}^\infty y f_Y(y) dy = \int_0^\infty y \cdot 2ye^{-y^2} dy$

$= \cdots differention\ by\ parts \ldots = \int_0^\infty e^{-y^2} dy$,

This looks like the standard normal density function, for which $\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 1$,

Applying substitution $x = \sqrt{2} \cdot y$, we find: $\int_{-\infty}^\infty e^{-y^2} dy = \sqrt{\pi}$, dus $E(Y) = \frac{1}{2}\sqrt{\pi}$

8. **a.** 0.3085; 0.3753; 0.6826.

e.g. $P(|X-1| < 2) = P(-2 < X - 1 < +2) = P\left(-\frac{2}{2} < \frac{X-1}{2} < \frac{2}{2}\right)$

$= \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 = 2 \cdot 0.8417 - 1 = 68.34\%$.

b. $P(X \leq c) = P\left(\frac{X-1}{2} \leq \frac{c-1}{2}\right) = \Phi\left(\frac{c-1}{2}\right) = 90\%$, so $\frac{c-1}{2} = 1.28$.

$c = 1 + 2 \cdot 1.28 = 3.56$ is the 90th percentile of X.

c. $c = 1 - 2 \cdot 1.28 = -1.56$

9. For instance: $P(-2\sigma < X - \mu < 2 \cdot \sigma) = P\left(-2 < \frac{X-\mu}{\sigma} < 2\right) = \Phi(2) - \Phi(-2)$

$= 2\Phi(2) - 1 = 2 \cdot 0.9772 - 1 = 0.9544 \approx 95.4\%$

10. The interval bounds of the weight classes of eggs are 50 ± 1.27 gr and 50 ± 4.21 gr.

11. **a**. Since $E(X) = \mu$, we have:

$E(X-\mu)^3 = E(X^3 - 3 \cdot X^2 \cdot \mu + 3 \cdot X \cdot \mu^2 - \mu^3)$

$= E(X^3) - 3\mu E(X^2) + 3\mu^2 E(X) - \mu^3 = E(X^3) - 3\mu E(X^2) + 2\mu^3$

b. $E(X) = \frac{1}{2}$, $E(X^2) = \int_0^1 x^2 \cdot dx = \frac{1}{3}x^3|_0^1 = \frac{1}{3}$ and $E(X^3) = \int_0^1 x^3 \cdot dx = \frac{1}{4}x^4|_0^1 = \frac{1}{4}$.
And using a. $E(X-\mu)^3 = \frac{1}{4} - 3 \cdot \frac{1}{2} \cdot \frac{1}{3} + 2 \cdot \left(\frac{1}{2}\right)^3 = 0$
(or directly: $E(X-\mu)^3 = \int_0^1 \left(x - \frac{1}{2}\right)^3 dx = \frac{1}{4}\left(x - \frac{1}{2}\right)^4 |_0^1 = \frac{1}{4}\left(\frac{1}{16} - \frac{1}{16}\right) = 0$).

c. $E(X) = \frac{1}{\lambda} = 1$ and since $var(X) = E(X^2) - (EX)^2 = 1$, is $E(X^2) = 1 + 1 = 2$.
$E(X^3) = \int_0^\infty x^3 \cdot e^{-x} dx = x^3 \cdot -e^{-x}|_0^1 + 3\int_0^\infty x^2 \cdot e^{-x} dx = 3E(X^2) = 6$.
$E(X-\mu)^3 = 6 - 3 \cdot 1 \cdot 2 + 2 \cdot 1 = 2$

d. Correct: the uniform distribution is symmetric: $E(X-\mu)^3 = 0$.
and the exponential distribution is skewed to the right: $E(X-\mu)^3 = 2$.

Chapter 7

1. **a**. $P(X > 90 \text{ en } Y > 90) = P(X > 90) \cdot P(Y > 90) = \left(1 - \Phi\left(\frac{90-80}{10}\right)\right)^2 \approx 2.52\%$

b. $X + Y$ is $N(80+80, 100+100)$-distributed, so
$P(X + Y > 180) = P\left(Z > \frac{180-160}{\sqrt{200}}\right) \approx 1 - \Phi(1.41) = 7.93\%$

c. The event "$X > 90$ and $Y > 90$" is a part of the event "$X + Y > 180$"

2. **a.** $f(x) = 1$ and $F(x) = x$, both if $0 \le x \le 1$

b. $F_M(m) = P(max(X_1, X_2, X_3) \le m) \overset{\text{ind.}}{=} P(X_1 \le m)P(X_2 \le m)P(X_3 \le m) = [F(m)]^3$
$f_M(m) = 3F(m)^2 \cdot f(m) = 3m^2 \cdot 1$, if $0 \le m \le 1$

3. **a**. $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx = \int_0^z e^{-x} e^{-(z-x)} dx = \int_0^z e^{-z} dx = e^{-z} \cdot x|_{x=0}^{x=z}$
$= ze^{-z}$, if $z \ge 0$
(And $f_{X+Y}(z) = 0$, if $z < 0$)

b. $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx = \int_0^z e^{-x} \cdot 2e^{-2(z-x)} dx = \int_0^z 2e^{-2z} e^x dx$
$= 2e^{-2z} \cdot e^x|_{x=0}^{x=z} = 2e^{-z} - 2e^{-2z}$, *if* $z \ge 0$
(and $f_{X+Y}(z) = 0$, *if* $z < 0$)

c. $P(X > 1 \text{ and } Y < 1) = P(X > 1) \cdot P(Y < 1) = e^{-1} \cdot (1 - e^{-2 \cdot 1}) \approx 31.8\%$

4. $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx = \int_0^z \frac{1}{\sqrt{2\pi x}} e^{-\frac{1}{2}x} \cdot \frac{1}{\sqrt{2\pi(z-x)}} e^{-\frac{1}{2}(z-x)} dx$
$= \frac{e^{-\frac{1}{2}z}}{2\pi} \int_0^z \frac{1}{\sqrt{x(z-x)}} dx$. The last integral equals π (given): $f_{X+Y}(z) = \frac{1}{2} e^{-\frac{1}{2}z}$, if $z > 0$.
The Chi square distribution with 2 degrees of freedom is apparently the same as the Exp $\left(\frac{1}{2}\right)$-distribution.

5. **a**. Yes, the property $E(X + Y) = E(X) + E(Y)$ always holds.
b. No, in general we need the assumption of independence for this property, but the salaries of partners are likely to be dependent (partners often share the level of education, the number of working hours are related, etc.)

6. $P(X > Y) = P(X - Y > 0) = 1 - \Phi\left(\frac{0-(-10)}{\sqrt{136}}\right) \approx 0.1965$, so we expect about 20 out of 100 will break.

7. **a**. $\Phi(-1.80) = 0.0359$
b. $\frac{3}{5}$

8. **a**. $X + Y \sim N(75 + 65, 250 + 150)$, so $P(X + Y > 150) = 1 - \Phi\left(\frac{150-140}{\sqrt{400}}\right) = 1 - \Phi\left(\frac{1}{2}\right)$
b. $\sum_{i=1}^{100} X_i \overset{\text{CLT}}{\sim} N\left(100 \cdot \frac{1}{2}, 100 \cdot \frac{1}{4}\right)$, so $P(\sum_{i=1}^{100} X_i \leq 58) = \Phi\left(\frac{58-50}{\sqrt{25}}\right) = \Phi(1.6)$

9. **a**. independence of the 100 service times, all having the same distribution with $\mu = 95$ and σ = 20.
b. $\overline{X} = \frac{1}{100}\sum_{i=1}^{100} X_i \overset{\text{CLT}}{\sim} N\left(\mu, \frac{\sigma^2}{100}\right)$, so $N(95, 4)$ and

$$P\left(\overline{X} > 100\right) \overset{\text{CLT}}{\approx} 1 - \Phi\left(\frac{100-95}{\sqrt{4}}\right) = 1 - \Phi(2.5) = 0.62\%$$

10. **a**. $X \sim B(250, 0.25)$ so X is approximately $N\left(np, np(1-p)\right) = N(62.5, 46.875)$ (rule of thumb $n > 25$, $np > 5$ and $n(1-p) > 5$ is fulfilled)
b. 22% of 250 is 55 voters:

$$P(X \leq 55) \overset{\text{c.c.}}{=} P(X \leq 55.5) \overset{\text{CLT}}{\approx} \Phi\left(\frac{55.5-62.5}{\sqrt{46.875}}\right) \approx \Phi(-1.02) = 0.1539$$

11.
c. Now $\mu = np = 250 \cdot 0.01 = 2.5 < 5$, so use a Poisson approximation with μ =2.5.
$P(X > 5) = 1 - P(X \leq 5) \approx 4.2\%$

12. **a**. $X \sim B(100, p)$, so μ = 100p and $\sigma^2 = 100p(1-p)$
b. $\frac{X}{100} \overset{\text{CLS}}{\sim} N\left(p, \frac{p(1-p)}{100}\right)$, so $P\left(-0.05 \leq \frac{X}{100} - p \leq 0.05\right)$

$$= P\left(-\frac{0.05}{\sqrt{\frac{p(1-p)}{100}}} \leq Z \leq \frac{0.05}{\sqrt{\frac{p(1-p)}{100}}}\right) = \Phi\left(\frac{0.5}{\sqrt{p(1-p)}}\right) - \Phi\left(-\frac{0.5}{\sqrt{p(1-p)}}\right)$$

$p(1-p) \leq \frac{1}{4}$, so this probability is at least $\Phi(1) - \Phi(-1) = 68.21\%$

13. $X =$ "demand of iPads in 6 weekdays" is Poisson($\mu = 6 \cdot 6 = 36$)-distributed.
This approximation is approximated with the $N(36, 36)$-distribution.

a. $P(X \leq 40) \overset{\text{c.c.}}{=} P(X \leq 40.5) = \Phi\left(\frac{40.5-36}{\sqrt{36}}\right) = \Phi(0.75) = 77.34\%$
b. $P(X \leq s) \overset{\text{c.c.}}{=} P(X \leq s + 0.5) = \Phi\left(\frac{s+0.5-36}{\sqrt{36}}\right) \geq 99\%$, then $\frac{s+0.5-36}{\sqrt{36}} = 2.33$, so $s = 35.5 + 2.33 \cdot 6 = 49.48$. The safety stock s should be (at least) 50.

14. **a.** $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.99^{15} = 14.0\%$
b. $P(X \leq 3) = 0.857$ (Poisson –table with $\mu = 2$)
c. $P(X \geq 50) = 1 - P(Z \leq 1.51) = 1 - 0.9345 \approx 6.5\%$

Chapter 8

1. **a**. $E(X) = 60$ and $P(X \geq 30) = P(X \geq 90 | X \geq 60) = e^{-\frac{1}{2}}$.
c. 38.12 $cent$ (> 30 $cent$)
d. 881.48 $cent^2$

2. **a**. $P(Y > t) = P(X = 0)$
b. $P(X = 0) = \frac{(at)^0 e^{-at}}{0!} = e^{-at}$
c. $P(Y > t) = e^{-at}$: this is the same expression as the corresponding probability of an exponential distribution with parameter $\lambda = a$,

3. **b.** $E(Y) = 1$ and $E(X) = \frac{1}{2}$, so $1 = E(Y) \neq \ln\left(\frac{1}{EX}\right) = \ln(2)$

4. **a**. $E(X_i) = 12$, $var(X_i) = 144$ and $E(\sum_{i=1}^{6} X_i) = 12 \cdot 6 = 72$
b. $P(X_1 > 12) = P(X_1 > 15 | X_1 > 3) = e^{-\frac{1}{12}12} = e^{-1}$.
c. $N \sim Poisson(\lambda t)$, with $\lambda t = \frac{1}{12} \cdot 60 = 5$, so $P(N \geq 6) = 1 - P(X \leq 5) = 38.4\%$.
d. $P(\sum_{i=1}^{6} X_i) = P(N \geq 6) = 38.4\%$

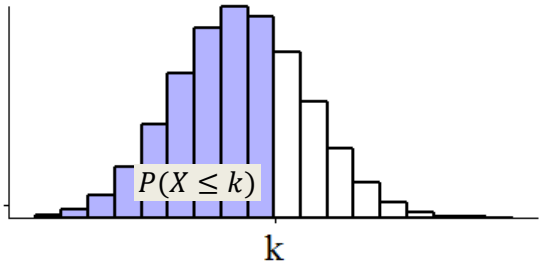
5. **a**. S_n has an Erlang distribution with parameters n and λ = ¼.
$E(S_n) = \frac{n}{\lambda} = 4n$ and $var(S_n) = \frac{n}{\lambda^2} = 16n$
c. $P\left(\overline{X}_2 > 5\right) = \frac{7}{2} e^{-\frac{5}{2}}$
d. $P\left(\overline{X}_{100} > 5\right) \approx 0.62\%$.

Table of binomial probabilities

The tables contain cumulative probabilities $P(X=i)$

$$P(X \le k) = \sum_{i=0}^{k} P(X = i)$$

(rounded in three decimals)



***n* = 5**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.951	0.774	0.590	0.444	0.328	0.237	0.168	0.116	0.078	0.050	0.031	0.402	0.132
1	0.999	0.977	0.919	0.835	0.737	0.633	0.528	0.428	0.337	0.256	0.188	0.804	0.461
2	1.000	0.999	0.991	0.973	0.942	0.896	0.837	0.765	0.683	0.593	0.500	0.965	0.790
3		1.000	1.000	0.998	0.993	0.984	0.969	0.946	0.913	0.869	0.813	0.997	0.955
4				1.000	1.000	0.999	0.998	0.995	0.990	0.982	0.969	1.000	0.996

***n* = 6**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.941	0.735	0.531	0.377	0.262	0.178	0.118	0.075	0.047	0.028	0.016	0.335	0.088
1	0.999	0.967	0.886	0.776	0.655	0.534	0.420	0.319	0.233	0.164	0.109	0.737	0.351
2	1.000	0.998	0.984	0.953	0.901	0.831	0.744	0.647	0.544	0.442	0.344	0.938	0.680
3		1.000	0.999	0.994	0.983	0.962	0.930	0.883	0.821	0.745	0.656	0.991	0.900
4			1.000	1.000	0.998	0.995	0.989	0.978	0.959	0.931	0.891	0.999	0.982
5					1.000	1.000	0.999	0.998	0.996	0.992	0.984	1.000	0.999

***n* = 7**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.932	0.698	0.478	0.321	0.210	0.133	0.082	0.049	0.028	0.015	0.008	0.279	0.059
1	0.998	0.956	0.850	0.717	0.577	0.445	0.329	0.234	0.159	0.102	0.063	0.670	0.263
2	1.000	0.996	0.974	0.926	0.852	0.756	0.647	0.532	0.420	0.316	0.227	0.904	0.571
3		1.000	0.997	0.988	0.967	0.929	0.874	0.800	0.710	0.608	0.500	0.982	0.827
4			1.000	0.999	0.995	0.987	0.971	0.944	0.904	0.847	0.773	0.998	0.955
5				1.000	1.000	0.999	0.996	0.991	0.981	0.964	0.938	1.000	0.993
6						1.000	1.000	0.999	0.998	0.996	0.992		1.000

***n* = 8**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.923	0.663	0.430	0.272	0.168	0.100	0.058	0.032	0.017	0.008	0.004	0.233	0.039
1	0.997	0.943	0.813	0.657	0.503	0.367	0.255	0.169	0.106	0.063	0.035	0.605	0.195
2	1.000	0.994	0.962	0.895	0.797	0.679	0.552	0.428	0.315	0.220	0.145	0.865	0.468
3		1.000	0.995	0.979	0.944	0.886	0.806	0.706	0.594	0.477	0.363	0.969	0.741
4			1.000	0.997	0.990	0.973	0.942	0.894	0.826	0.740	0.637	0.995	0.912
5				1.000	0.999	0.996	0.989	0.975	0.950	0.912	0.855	1.000	0.980
6					1.000	1.000	0.999	0.996	0.991	0.982	0.965		0.997
7							1.000	1.000	0.999	0.998	0.996		1.000

Tab-2

***n* = 9**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.914	0.630	0.387	0.232	0.134	0.075	0.040	0.021	0.010	0.005	0.002	0.194	0.026
1	0.997	0.929	0.775	0.599	0.436	0.300	0.196	0.121	0.071	0.039	0.020	0.543	0.143
2	1.000	0.992	0.947	0.859	0.738	0.601	0.463	0.337	0.232	0.150	0.090	0.822	0.377
3		0.999	0.992	0.966	0.914	0.834	0.730	0.609	0.483	0.361	0.254	0.952	0.650
4		1.000	0.999	0.994	0.980	0.951	0.901	0.828	0.733	0.621	0.500	0.991	0.855
5			1.000	0.999	0.997	0.990	0.975	0.946	0.901	0.834	0.746	0.999	0.958
6				1.000	1.000	0.999	0.996	0.989	0.975	0.950	0.910	1.000	0.992
7						1.000	1.000	0.999	0.996	0.991	0.980		0.999
8								1.000	1.000	0.999	0.998		1.000

***n* = 10**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.904	0.599	0.349	0.197	0.107	0.056	0.028	0.013	0.006	0.003	0.001	0.162	0.017
1	0.996	0.914	0.736	0.544	0.376	0.244	0.149	0.086	0.046	0.023	0.011	0.485	0.104
2	1.000	0.988	0.930	0.820	0.678	0.526	0.383	0.262	0.167	0.100	0.055	0.775	0.299
3		0.999	0.987	0.950	0.879	0.776	0.650	0.514	0.382	0.266	0.172	0.930	0.559
4		1.000	0.998	0.990	0.967	0.922	0.850	0.751	0.633	0.504	0.377	0.985	0.787
5			1.000	0.999	0.994	0.980	0.953	0.905	0.834	0.738	0.623	0.998	0.923
6				1.000	0.999	0.996	0.989	0.974	0.945	0.898	0.828	1.000	0.980
7					1.000	1.000	0.998	0.995	0.988	0.973	0.945		0.997
8							1.000	0.999	0.998	0.995	0.989		1.000
9								1.000	1.000	1.000	0.999		

***n* = 15**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.860	0.463	0.206	0.087	0.035	0.013	0.005	0.002	0.000	0.000	0.000	0.065	0.002
1	0.990	0.829	0.549	0.319	0.167	0.080	0.035	0.014	0.005	0.002	0.000	0.260	0.019
2	1.000	0.964	0.816	0.604	0.398	0.236	0.127	0.062	0.027	0.011	0.004	0.532	0.079
3		0.995	0.944	0.823	0.648	0.461	0.297	0.173	0.091	0.042	0.018	0.768	0.209
4		0.999	0.987	0.938	0.836	0.686	0.515	0.352	0.217	0.120	0.059	0.910	0.404
5		1.000	0.998	0.983	0.939	0.852	0.722	0.564	0.403	0.261	0.151	0.973	0.618
6			1.000	0.996	0.982	0.943	0.869	0.755	0.610	0.452	0.304	0.993	0.797
7				0.999	0.996	0.983	0.950	0.887	0.787	0.654	0.500	0.999	0.912
8				1.000	0.999	0.996	0.985	0.958	0.905	0.818	0.696	1.000	0.969
9					1.000	0.999	0.996	0.988	0.966	0.923	0.849		0.991
10						1.000	0.999	0.997	0.991	0.975	0.941		0.998
11							1.000	1.000	0.998	0.994	0.982		1.000
12									1.000	0.999	0.996		
13										1.000	1.000		

Tab-3

***n* = 20**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.818	0.358	0.122	0.039	0.012	0.003	0.001	0.000	0.000	0.000	0.000	0.026	0.000
1	0.983	0.736	0.392	0.176	0.069	0.024	0.008	0.002	0.001	0.000	0.000	0.130	0.003
2	0.999	0.925	0.677	0.405	0.206	0.091	0.035	0.012	0.004	0.001	0.000	0.329	0.018
3	1.000	0.984	0.867	0.648	0.411	0.225	0.107	0.044	0.016	0.005	0.001	0.567	0.060
4		0.997	0.957	0.830	0.630	0.415	0.238	0.118	0.051	0.019	0.006	0.769	0.152
5		1.000	0.989	0.933	0.804	0.617	0.416	0.245	0.126	0.055	0.021	0.898	0.297
6			0.998	0.978	0.913	0.786	0.608	0.417	0.250	0.130	0.058	0.963	0.479
7			1.000	0.994	0.968	0.898	0.772	0.601	0.416	0.252	0.132	0.989	0.661
8				0.999	0.990	0.959	0.887	0.762	0.596	0.414	0.252	0.997	0.809
9				1.000	0.997	0.986	0.952	0.878	0.755	0.591	0.412	0.999	0.908
10					0.999	0.996	0.983	0.947	0.872	0.751	0.588	1.000	0.962
11					1.000	0.999	0.995	0.980	0.943	0.869	0.748		0.987
12						1.000	0.999	0.994	0.979	0.942	0.868		0.996
13							1.000	0.998	0.994	0.979	0.942		0.999
14								1.000	0.998	0.994	0.979		1.000
15									1.000	0.998	0.994		
16										0.998	0.994		
17										1.000	1.000		

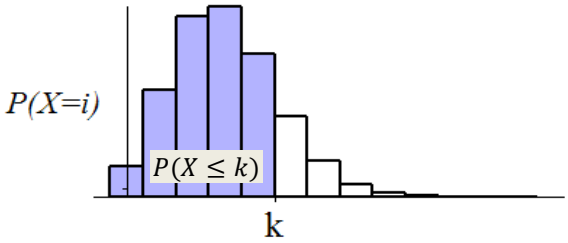
***n* = 25**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.778	0.277	0.072	0.017	0.004	0.001	0.000	0.000	0.000	0.000	0.000	0.010	0.000
1	0.974	0.642	0.271	0.093	0.027	0.007	0.002	0.000	0.000	0.000	0.000	0.063	0.001
2	0.998	0.873	0.537	0.254	0.098	0.032	0.009	0.002	0.000	0.000	0.000	0.189	0.004
3	1.000	0.966	0.764	0.471	0.234	0.096	0.033	0.010	0.002	0.000	0.000	0.382	0.015
4		0.993	0.902	0.682	0.421	0.214	0.090	0.032	0.009	0.002	0.000	0.594	0.046
5		0.999	0.967	0.838	0.617	0.378	0.193	0.083	0.029	0.009	0.002	0.772	0.112
6		1.000	0.991	0.930	0.780	0.561	0.341	0.173	0.074	0.026	0.007	0.891	0.222
7			0.998	0.975	0.891	0.727	0.512	0.306	0.154	0.064	0.022	0.955	0.370
8			1.000	0.992	0.953	0.851	0.677	0.467	0.274	0.134	0.054	0.984	0.538
9				0.998	0.983	0.929	0.811	0.630	0.425	0.242	0.115	0.995	0.696
10				1.000	0.994	0.970	0.902	0.771	0.586	0.384	0.212	0.999	0.822
11					0.998	0.989	0.956	0.875	0.732	0.543	0.345	1.000	0.908
12					1.000	0.997	0.983	0.940	0.846	0.694	0.500		0.958
13						0.999	0.994	0.975	0.922	0.817	0.655		0.984
14						1.000	0.998	0.991	0.966	0.904	0.788		0.994
15							1.000	0.997	0.987	0.956	0.885		0.998
16								0.999	0.996	0.983	0.946		1.000
17								1.000	0.999	0.994	0.978		
18									1.000	0.998	0.993		
19										1.000	0.998		
20											1.000		

Table of Poisson probabilities

The tables contain cumulative probabilities

$$P(X \le k) = \sum_{i=0}^{k} \frac{\mu^i e^{-\mu}}{i!}$$



(Rounded in three decimales)

μ \ k	0	1	2	3	4	5	6	7	8	9	10
0.02	0.980	1.000									
0.04	0.961	0.999	1.000								
0.06	0.942	0.998	1.000								
0.08	0.923	0.997	1.000								
0.10	0.905	0.995	1.000								
0.15	0.861	0.990	0.999	1.000							
0.20	0.819	0.982	0.999	1.000							
0.25	0.779	0.974	0.998	1.000							
0.30	0.741	0.963	0.996	1.000							
0.35	0.705	0.951	0.994	1.000							
0.40	0.670	0.938	0.992	0.999	1.000						
0.45	0.638	0.925	0.989	0.999	1.000						
0.50	0.607	0.910	0.986	0.998	1.000						
0.55	0.577	0.894	0.982	0.998	1.000						
0.60	0.549	0.878	0.977	0.997	1.000						
0.65	0.522	0.861	0.972	0.996	0.999	1.000					
0.70	0.497	0.844	0.966	0.994	0.999	1.000					
0.75	0.472	0.827	0.959	0.993	0.999	1.000					
0.80	0.449	0.809	0.953	0.991	0.999	1.000					
0.85	0.427	0.791	0.945	0.989	0.998	1.000					
0.90	0.407	0.772	0.937	0.987	0.998	1.000					
0.95	0.387	0.754	0.929	0.984	0.997	1.000					
1.00	0.368	0.736	0.920	0.981	0.996	0.999	1.000				
1.1	0.333	0.699	0.900	0.974	0.995	0.999	1.000				
1.2	0.301	0.663	0.879	0.966	0.992	0.998	1.000				
1.3	0.273	0.627	0.857	0.957	0.989	0.998	1.000				
1.4	0.247	0.592	0.833	0.946	0.986	0.997	0.999	1.000			
1.5	0.223	0.558	0.809	0.934	0.981	0.996	0.999	1.000			
1.6	0.202	0.525	0.783	0.921	0.976	0.994	0.999	1.000			
1.7	0.183	0.493	0.757	0.907	0.970	0.992	0.998	1.000			
1.8	0.165	0.463	0.731	0.891	0.964	0.990	0.997	0.999	1.000		
1.9	0.150	0.434	0.704	0.875	0.956	0.987	0.997	0.999	1.000		
2.0	0.135	0.406	0.677	0.857	0.947	0.983	0.995	0.999	1.000		
2.2	0.111	0.355	0.623	0.819	0.928	0.975	0.993	0.998	1.000		
2.4	0.091	0.308	0.570	0.779	0.904	0.964	0.988	0.997	0.999	1.000	
2.6	0.074	0.267	0.518	0.736	0.877	0.951	0.983	0.995	0.999	1.000	
2.8	0.061	0.231	0.469	0.692	0.848	0.935	0.976	0.992	0.998	0.999	1.000

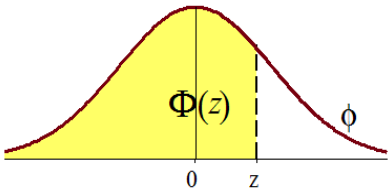
Tab-5

Poisson probabilities (continuation)

μ \ k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3.0	0.050	0.199	0.423	0.647	0.815	0.916	0.966	0.988	0.996	0.999	1.000					
3.2	0.041	0.171	0.380	0.603	0.781	0.895	0.955	0.983	0.994	0.998	1.000					
3.4	0.033	0.147	0.340	0.558	0.744	0.871	0.942	0.977	0.992	0.997	0.999	1.000				
3.6	0.027	0.126	0.303	0.515	0.706	0.844	0.927	0.969	0.988	0.996	0.999	1.000				
3.8	0.022	0.107	0.269	0.473	0.668	0.816	0.909	0.960	0.984	0.994	0.998	0.999	1.000			
4.0	0.018	0.092	0.238	0.433	0.629	0.785	0.889	0.949	0.979	0.992	0.997	0.999	1.000			
4.2	0.015	0.078	0.210	0.395	0.590	0.753	0.867	0.936	0.972	0.989	0.996	0.999	1.000			
4.4	0.012	0.066	0.185	0.359	0.551	0.720	0.844	0.921	0.964	0.985	0.994	0.998	0.999	1.000		
4.6	0.010	0.056	0.163	0.326	0.513	0.686	0.818	0.905	0.955	0.980	0.992	0.997	0.999	1.000		
4.8	0.008	0.048	0.143	0.294	0.476	0.651	0.791	0.887	0.944	0.975	0.990	0.996	0.999	1.000		
5.0	0.007	0.040	0.125	0.265	0.440	0.616	0.762	0.867	0.932	0.968	0.986	0.995	0.998	0.999	1.000	
5.2	0.006	0.034	0.109	0.238	0.406	0.581	0.732	0.845	0.918	0.960	0.982	0.993	0.997	0.999	1.000	
5.4	0.005	0.029	0.095	0.213	0.373	0.546	0.702	0.822	0.903	0.951	0.977	0.990	0.996	0.999	1.000	
5.6	0.004	0.024	0.082	0.191	0.342	0.512	0.670	0.797	0.886	0.941	0.972	0.988	0.995	0.998	0.999	1.000
5.8	0.003	0.021	0.072	0.170	0.313	0.478	0.638	0.771	0.867	0.929	0.965	0.984	0.993	0.997	0.999	1.000
6.0	0.002	0.017	0.062	0.151	0.285	0.446	0.606	0.744	0.847	0.916	0.957	0.980	0.991	0.996	0.999	0.999
6.2	0.002	0.015	0.054	0.134	0.259	0.414	0.574	0.716	0.826	0.902	0.949	0.975	0.989	0.995	0.998	0.999
6.4	0.002	0.012	0.046	0.119	0.235	0.384	0.542	0.687	0.803	0.886	0.939	0.969	0.986	0.994	0.997	0.999
6.6	0.001	0.010	0.040	0.105	0.213	0.355	0.511	0.658	0.780	0.869	0.927	0.963	0.982	0.992	0.997	0.999
6.8	0.001	0.009	0.034	0.093	0.192	0.327	0.480	0.628	0.755	0.850	0.915	0.955	0.978	0.990	0.996	0.998
7.0	0.001	0.007	0.030	0.082	0.173	0.301	0.450	0.599	0.729	0.830	0.901	0.947	0.973	0.987	0.994	0.998
7.2	0.001	0.006	0.025	0.072	0.156	0.276	0.420	0.569	0.703	0.810	0.887	0.937	0.967	0.984	0.993	0.997
7.4	0.001	0.005	0.022	0.063	0.140	0.253	0.392	0.539	0.676	0.788	0.871	0.926	0.961	0.980	0.991	0.996
7.6	0.001	0.004	0.019	0.055	0.125	0.231	0.365	0.510	0.648	0.765	0.854	0.915	0.954	0.976	0.989	0.995
7.8	0.000	0.004	0.016	0.048	0.112	0.210	0.338	0.481	0.620	0.741	0.835	0.902	0.945	0.971	0.986	0.993
8.0	0.000	0.003	0.014	0.042	0.100	0.191	0.313	0.453	0.593	0.717	0.816	0.888	0.936	0.966	0.983	0.992
8.5	0.000	0.002	0.009	0.030	0.074	0.150	0.256	0.386	0.523	0.653	0.763	0.849	0.909	0.949	0.973	0.986
9.0	0.000	0.001	0.006	0.021	0.055	0.116	0.207	0.324	0.456	0.587	0.706	0.803	0.876	0.926	0.959	0.978
9.5	0.000	0.001	0.004	0.015	0.040	0.089	0.165	0.269	0.392	0.522	0.645	0.752	0.836	0.898	0.940	0.967
10.0	0.000	0.000	0.003	0.010	0.029	0.067	0.130	0.220	0.333	0.458	0.583	0.697	0.792	0.864	0.917	0.951

μ \ k	16	17	18	19	20	21	22
6.0	1.000						
5.2	1.000						
6.4	1.000						
6.6	0.999	1.000					
6.8	0.999	1.000					
7.0	0.999	1.000					
7.2	0.999	1.000					
7.4	0.998	0.999	1.000				
7.6	0.998	0.999	1.000				
7.8	0.997	0.999	1.000				
8.0	0.996	0.998	0.999	1.000			
8.5	0.993	0.997	0.999	0.999	1.000		
9.0	0.989	0.995	0.998	0.999	1.000		
9.5	0.982	0.991	0.996	0.998	0.999	1.000	
10.0	0.973	0.986	0.993	0.997	0.998	0.999	1.000

Standard normal probabilities



The table gives the distribution function Φ for a N(0,1)-variable Z

$$\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}}\, dx$$

Last column: N(0,1)-density function (z in 1 dec.): $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

	Second decimal of **z**										
z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**φ(z)**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	0.3989
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.3970
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.3910
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	0.3814
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	0.3683
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	0.3521
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	0.3332
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	0.3123
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	0.2897
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	0.2661
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	0.2420
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	0.2179
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	0.1942
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	0.1714
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	0.1497
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	0.1295
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	0.1109
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	0.0940
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	0.0790
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767	0.0656
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	0.0540
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857	0.0440
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890	0.0355
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	0.0283
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	0.0224
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	0.0175
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964	0.0136
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974	0.0104
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981	0.0079
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	0.0060
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990	0.0044
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993	0.0033
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	0.0024
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997	0.0017
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	0.0012
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.0009
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.0006

Index