Course: Probability Theory

Jürgen Pilz

Institut für Statistik, Universität Klagenfurt
Universitätsstr. 65-67, 9020 Klagenfurt, Austria
juergen.pilz@aau.at

www.jpilz.net

Master Study Program in Applied Data Science
CUAS / Villach, Austria

0. Counting, Combinatorics

Start our journey through probability with various approaches to do counting

Reason to do so: Likelihood of the occurrence of a given event E should be equal to

$$Lik(E) = \frac{\text{number of scenarios that are constituents of } E}{\text{total number of possible scenarios}}$$

assuming all scenarios are equally likely to occur.

Example

Drawing a card from a standard card deck.
Let E be the event of getting a spade. We have 13 spades in a card deck of 52 cards, and thus $Lik(E) = 13/52 = 1/4$.

Permutation

We need to figure out the numbers in the numerator and denominator of $Lik(E)$ above in more complex settings.
The science of counting is captured by a branch of mathematics called **combinatorics**.

0.1 Permutations

Assume that there are n **distinguishable** items.

Definition: An arrangement of all items is called a **permutation**.

Example

Q: How many ways can the set of letters $\{a, b, c, d\}$ be lined up?

A:

abcd	*abdc*	*acbd*	*acdb*	*adbc*	*adcb*
bacd	*badc*	*bcad*	*bcda*	*bdac*	*bdca*
cabd	*cadb*	*cbad*	*cbda*	*cdab*	*cdba*
dabc	*dacb*	*dbac*	*dbca*	*dcab*	*dcba*

Thus, there are $24 = 4 \cdot 3 \cdot 2 \cdot 1$ arrangements.

Permutation

In general, the number of permutations of n distinguishable elements reads

$$P_n = n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$$

Example

Q: How many ways can 10 books be lined up on a shelf?

A: $P_{10} = 10! = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 3628800.$

In the programming language *R* this reads

```
> factorial(10)
[1] 3628800
```

Alternatively, you can write in *R*:

```
> prod(1:10)
[1] 3628800
```

Permutation

We now consider two variants of permutations with n distinguishable elements.

0.1.1 k- splitting

We perform k experiments sequentially. In the first experiment, we have n_1 distinct items and line them up. In the second experiment, we line up n_2 distinct elements, and repeat the process till we reach the kth experiment in which we line up the remaining n_k elements, $n_1 + n_2 + \ldots + n_k = n$. Then the total number of possible permutations is given by

$$P_k^{seq} := n_1! \cdot n_2! \cdot \ldots \cdot n_k!; \quad n_1 + n_2 + \ldots + n_k = n$$

Example

Suppose there are 3 distinct books in algebra, 4 distinct books in calculus and 6 distinct books in probability theory & statistics (PTS). Books are to be lined up in the order of algebra, calculus and PTS.

Example (cont'd)

Q: How many arrangements are possible?

A: $n = 13; n_1 = 3, n_2 = 4, n_3 = 6$. Thus,

$$P_k^{seq} = 3! \cdot 4! \cdot 6! = 6 \cdot 24 \cdot 720 = 103680.$$

In the programming language *R* this reads

```
> factorial(3)*factorial(4)*factorial(6)
[1] 103680
```

Alternatively, you can write in *R*:

```
> a=c(factorial(3),factorial(4),factorial(6))
> prod(a)
[1] 103680
```

0.1.2 Groups of indistinguishable elements

Consider the situation where we have k distinguishable groups among the n elements to be lined up, but within the groups the items are indistinguishable.
Since items in the same group are not distinguishable, there are $n_1! \cdot n_2! \cdot \ldots \cdot n_k!$ duplicate arrangements. Actually, we thus have only

$$P_n^k := \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_k!}$$

distinguishable arrangements.

Example

Q1: How many different card games are there when playing skat?

A1: $n = 32$ cards, $k = 4$ groups (3 players, 2 cards left uncovered)
Therefore, $n_1 = n_2 = n_3 = 10, n_4 = 2$, and thus,
$P_{32}^4 := \frac{32!}{10! \cdot 10! \cdot 10! \cdot 2!} = 2753294408504640$

Grouping in Permutations

Example (cont'd)

```
> library(gmp)
> a=div.bigz(16*prod(23:31),factorial(10))
> b=div.bigz(prod(11:22),factorial(10))
> mul.bigz(a,b)
[1] 2753294408504640
```

Note: Suppose you need 10 minutes for playing a single card game. Then it would take about 52.35 billion years to play all the games.

Example

Q2: How many different 16-digit code signals can be formed using three $+$'s, four \$'s, four #'s and five *'s?

A2: $n = 16$, $k = 4$ groups; $n_1 = 3, n_2 = n_3 = 4, n_4 = 5$
Thus, we can make a total of
$\frac{16!}{3!4!4!5!} = 50.450.400$ different code signals.

Variations without repetition

0.2 Variations

We now consider arrangements of n distinguishable items from which we select only $k < n$ items and line them up. It is obvious that there are

$$V_{n,k} := (n)_k \equiv n \cdot (n-1) \cdot \ldots \cdot (n-k+1)$$

ways to do this. Note that the order of selection matters and each item can be selected only once. Therefore, this type of arrangement is also called **variation without repetition**.

Example

Q: How many different Top 3-Placements exist for the participating teams in the quarter-final of a tournament?

A: $n = 8, k = 3$. Hence there are

$$V_{8,3} = (8)_3 = 8 \cdot 7 \cdot 6 = 336$$

different possibilities for the first 3 placements.

Variations with repetition

Let us now consider the situation when choosing $k < n$ items and we allow for replications among the k items. It is clear from the above formula for $V_{n,k}$ that the product on the right-hand side must be modified then as

$$V_{n,k}^{rep} := \underbrace{n \cdot n \cdot \ldots \cdot n}_{k \text{ factors}} = n^k$$

Example

Q: Given a padlock which has options for four digits that range from $0 \ldots 9$, what is the number of code variations for this padlock?

A: Obviously, the first digit can have 10 values, the second digit can have 10 values, the third digit can have 10 values and the final fourth digit can also have 10 values. So, there are

$$V_{10,4} = 10 \cdot 10 \cdot 10 \cdot 10 = 10^4 = 10.000$$

code variations.

0.3 Combinations

Finally, let us consider arrangements of $k < n$ items in which, contrary to variations, the order does not matter. So, for example, when rolling two dice the pairs (5,6) and (6,5) are considered the same with regard to the sum obtained (11). An arrangement in which the order does not matter is called a **combination**.
There are two types of combinations: with and without repetition.

0.3.1 Combinations Without Repetition

Suppose we have a class of n students and want to select k of them to form a basketball team. How many selections are there? Let's first approach the problem from a variation perspective: we have k spots to be filled from a potential of n candidates, thus there are

$$(n)_k = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}$$

ways to fill the k spots.

Combinations without repetition

Actually, however, the **order** with which a spot gets filled with a candidate is irrelevant. Thus, the above count contains $k!$ of duplicates. Hence, the number of **distinguishable choices** is given by

$$C_{n,k} = \frac{(n)_k}{k!} = \frac{n!}{k!(n-k)!} \equiv \binom{n}{k}$$

The term $\binom{n}{k}$ is known as **binomial coefficient** "n over k".

Example

Q: How many choices are there to form a basketball team of $k = 5$ members out of a class of $n = 24$ students?

A: There are

$$C_{24,5} = \binom{24}{5} = \frac{24!}{5!19!} = \frac{24 \cdot 23 \cdot 22 \cdot 21 \cdot 20}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 42504$$

ways to form such a team of 5 members.

Example (cont'd)

Using R, we easily compute $C_{n,k}$ with the "choose" function:

```
> choose(24,5)
[1] 42504
```

0.3.2 Combinations with repetition

A typical situation of such a scenario occurs in quality control of produced items. We take a sample of k items from a production total of n items. The order in which the items are selected is not relevant when it comes to evaluating the quality of the production.

Sampling without replacement leads us to the scenario of a combination without repetition, i.e. there would be $\binom{n}{k}$ different ways of getting a sample of size k.

However, when we sample with replacement then we are led to a scenario of a **combination with repetition**. Then there are more ways of creating a sample of size k since

Combinations with repetition

- the total size n out of which we draw remains constant
- single items may be selected again in subsequent draws.

The number of combinations with repetition comes out as

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

This can be verified by the principle of complete induction (over k, starting with the trivial case $k = 1$)

Example

Suppose we have 5 balls of different colours in an urn, from which we draw 3 balls with replacement.

Q: What is the total number of sampling protocols that might occur?

A: We have $n = 5$ and $k = 3$. Hence, there are 35 different protocols:

$$C_{5,3}^{rep} = \binom{5+3-1}{3} = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 35$$

Experiments

1. Probability, Probability Space

1.1 Random Events

Central notion: Experiment, Trial

- Experiment= realization of a set of conditions, it is reproducible and its outcomes are unpredictable (cannot be foreseen).
- The result of an experiment is a random event.
- The outcome is uncertain (due to factors which have not been taken into account).
- Reproducibility allows the systematic study of underlying principles/laws which only become visible after a large number of repetitions.

Random Events

Probability Theory = Theory of the principles/laws of chance (Zufall) in mass events

Mathematical notions

- An **event field (Ereignisfeld)** $\mathscr{E}$ represents the set of all possible outcomes (events) of an experiment
- The **sure event (sicheres Ereignis)** Ω necessarily occurs in any repetition of the experiment
- The **impossible event (unmögliches Ereignis)** $\emptyset$ never occurs
- A **random event (zufälliges Ereignis)**: $A, B, C, \ldots$ is an event which may occur, but is neither sure nor impossible (some event "between" $\emptyset$ and Ω)
- $\mathscr{E} = \{\emptyset, \Omega, A, B, C, \ldots\}$ denotes the **event field (Ereignisfeld)** associated with an experiment.

Example (Playing dice)

experimental outcomes when throwing a die: thrown numbers of points

$$\Omega = \{1, 2, 3, 4, 5, 6\}. \tag{1}$$

The event field (algebra) is given by

$$\begin{aligned}\mathscr{E} = \{\emptyset, \Omega, \overbrace{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}}^{\text{elementary events}}, \\ \{1,2\}, \{1,3\} \ldots \{5,6\}, \\ \{1,2,3\}, \{1,2,4\}, \ldots \{4,5,6\}, \\ \{1,2,3,4\}, \{1,2,3,5\}, \ldots \{3,4,5,6\}, \\ \{1,2,3,4,5\}, \{1,2,3,4,6\}, \ldots \{2,3,4,5,6\}\}.\end{aligned}$$

Each event is a subset of Ω.

1.2 Event Algebra

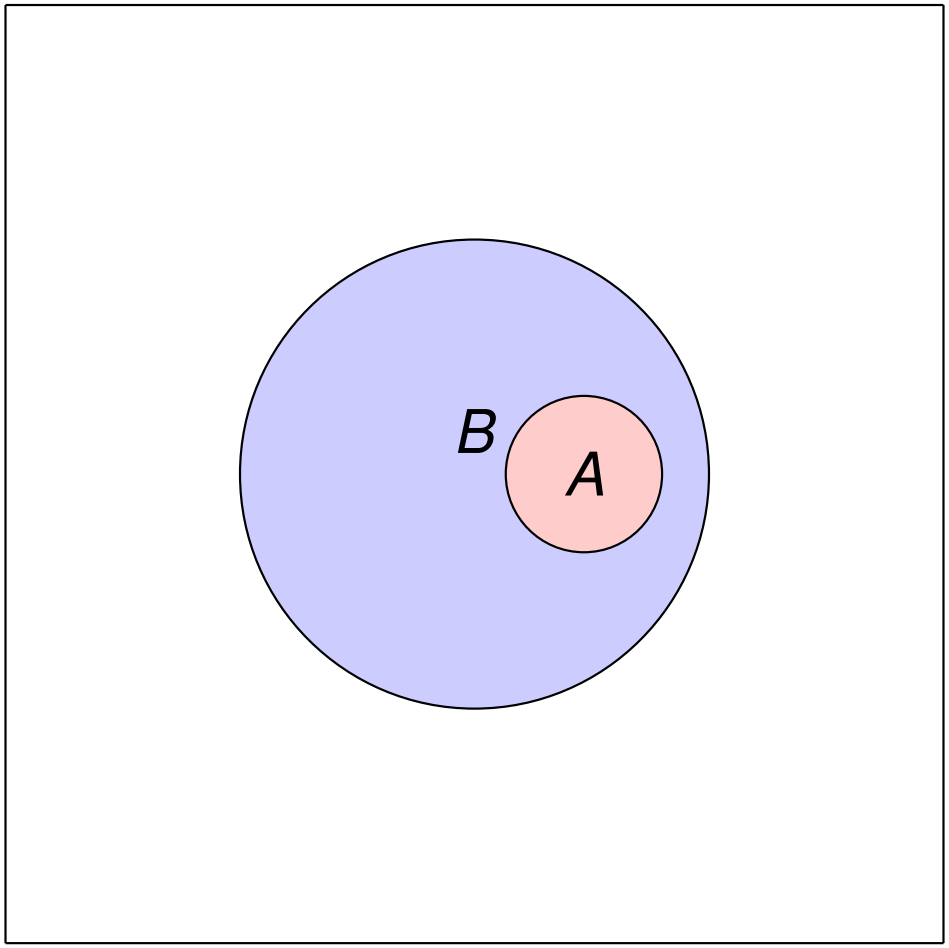
We introduce a partial ordering relation $\subseteq$ within $\mathscr{E}$ such that it becomes an algebra. We define

$$A \subseteq B \iff \begin{array}{l} A \text{ implies } B \\ \text{or } B \text{ comprises } A \end{array} \tag{2}$$

In case that $A \subseteq B$ it holds that with the ocurrence of A we always associate the occurence of B.
The crucial fact is that we may identify events with sets.
Then, in a set-theoretical sense, $A \subseteq B$ means that A is a subset of B.

Event algebra

Set-theoretical visualization of $A \subseteq B$ through **Venn Diagram**



Event algebra

Properties of relation $\subseteq$:

$$\emptyset \subseteq A \tag{3}$$
$$A \subseteq A \tag{4}$$
$$A \subseteq \Omega \tag{5}$$
$$\emptyset \subseteq \Omega \tag{6}$$
$$A \subseteq B \quad \wedge \quad B \subseteq C \Longrightarrow A \subseteq C \tag{7}$$

(4) means: $\subseteq$ is **reflexive**
(7) means: $\subseteq$ is **transitive**

Moreover, $\subseteq$ is **antisymmetric**
However, $\subseteq$ does not define a total **ordering relation**

$\subseteq$ defines only a **partial ordering** in $\mathscr{E}$

Event algebra

Example (Throwing a die)

Let be A the event "Die shows 2"
and B the event "Die shows an even number"
Then we have

$$A = \{2\} \tag{8}$$
$$B = \{2, 4, 6\} \tag{9}$$

Obviously, $A \subseteq B$, i.e. A implies B (2 is an even number).

Equality of events : For 2 events $A, B \in \mathscr{E}$,

$$A = B \tag{10}$$

means that either both A and B occur or none of them.

From set theory it is immediately clear that

$$A = B \Longleftrightarrow A \subseteq B \wedge B \subseteq A \tag{11}$$

Operations between random events

Sum of events: For two events $A, B \in \mathscr{E}$ we write

$$A \cup B \tag{12}$$

meaning that **at least one of them occurs**, i.e.
A **or** B **or both**)

More general, we write

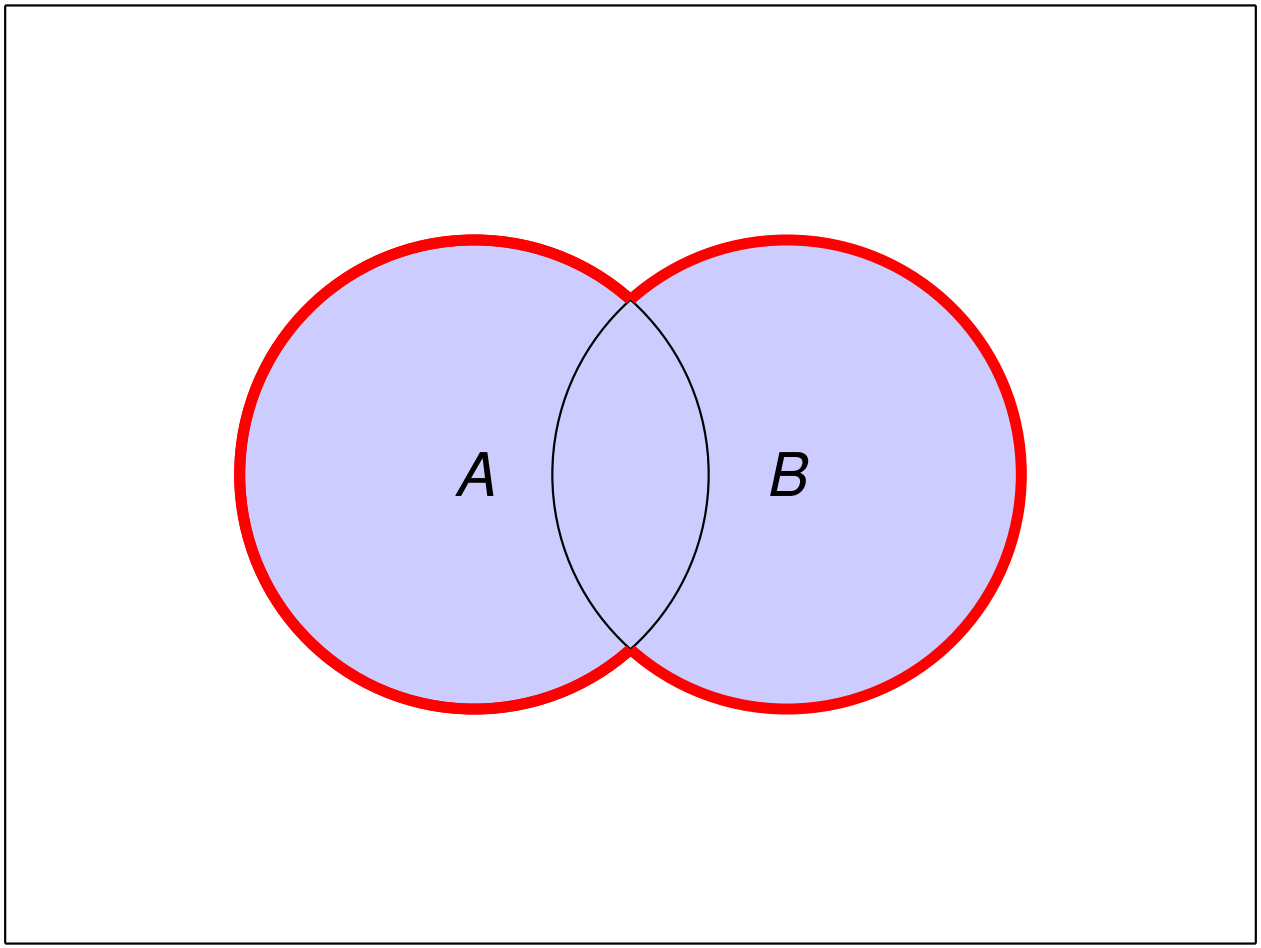
$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n \tag{13}$$

meaning that at least one of the events A_i occurs.

In a set-theoretical sense, the sum of events is just the **union** of the associated sets.

Operations between random events

Visualization: Sum of 2 events $A \cup B$



Operations between random events

Properties of the sum of events : For events $A, B \in \mathscr{E}$ it holds

$$A \cup B = B \cup A \tag{14}$$
$$A \cup (B \cup C) = (A \cup B) \cup C \tag{15}$$
$$A \cup \emptyset = A \tag{16}$$
$$A \cup A = A \tag{17}$$
$$A \cup \Omega = \Omega \tag{18}$$
$$A \subseteq A \cup B \tag{19}$$
$$B \subseteq A \cup B \tag{20}$$

(15) means: sum operation $\cup$ is **associative**

Operations between random events

Product of events: For two events $A, B \in \mathscr{E}$ we write

$$A \cap B \tag{21}$$

meaning that **both A and B occur**.

More general, we write

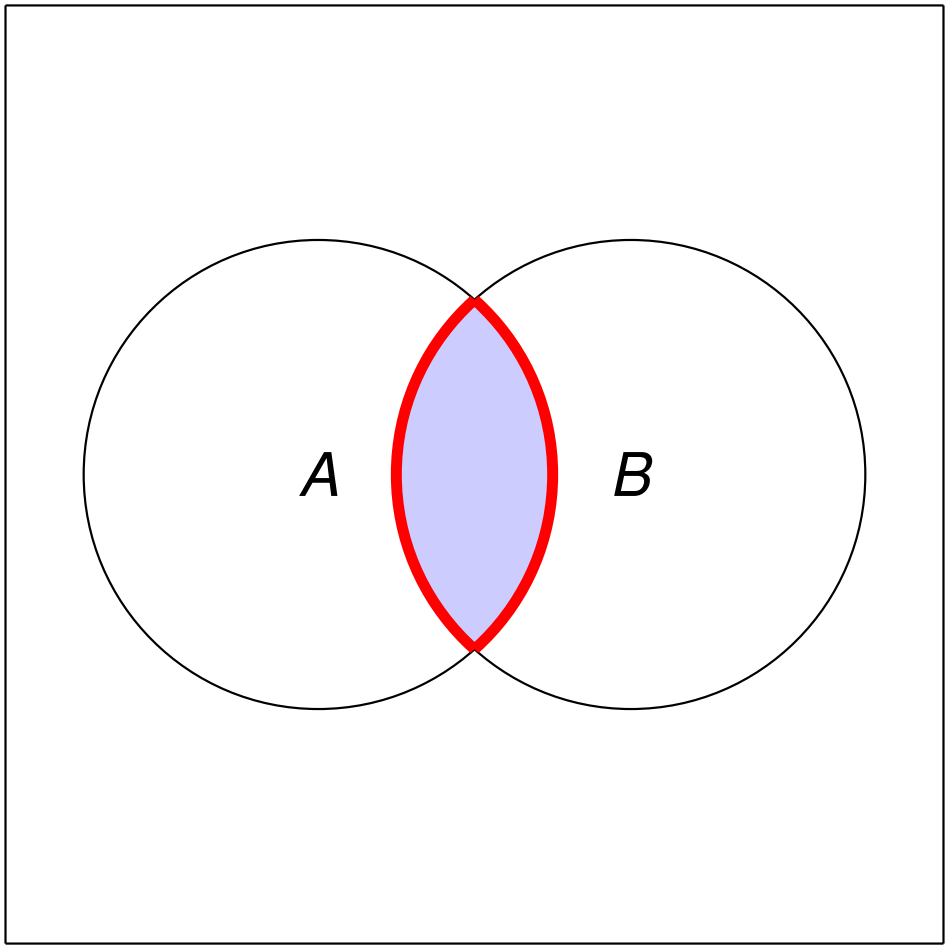
$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n \tag{22}$$

meaning that all of the events A_i occur.

In a set-theoretical sense, the product of events is just the **intersection** of the associated sets.

Operations between random events

Visualization: Product of 2 events $A \cap B$



Operations between random events

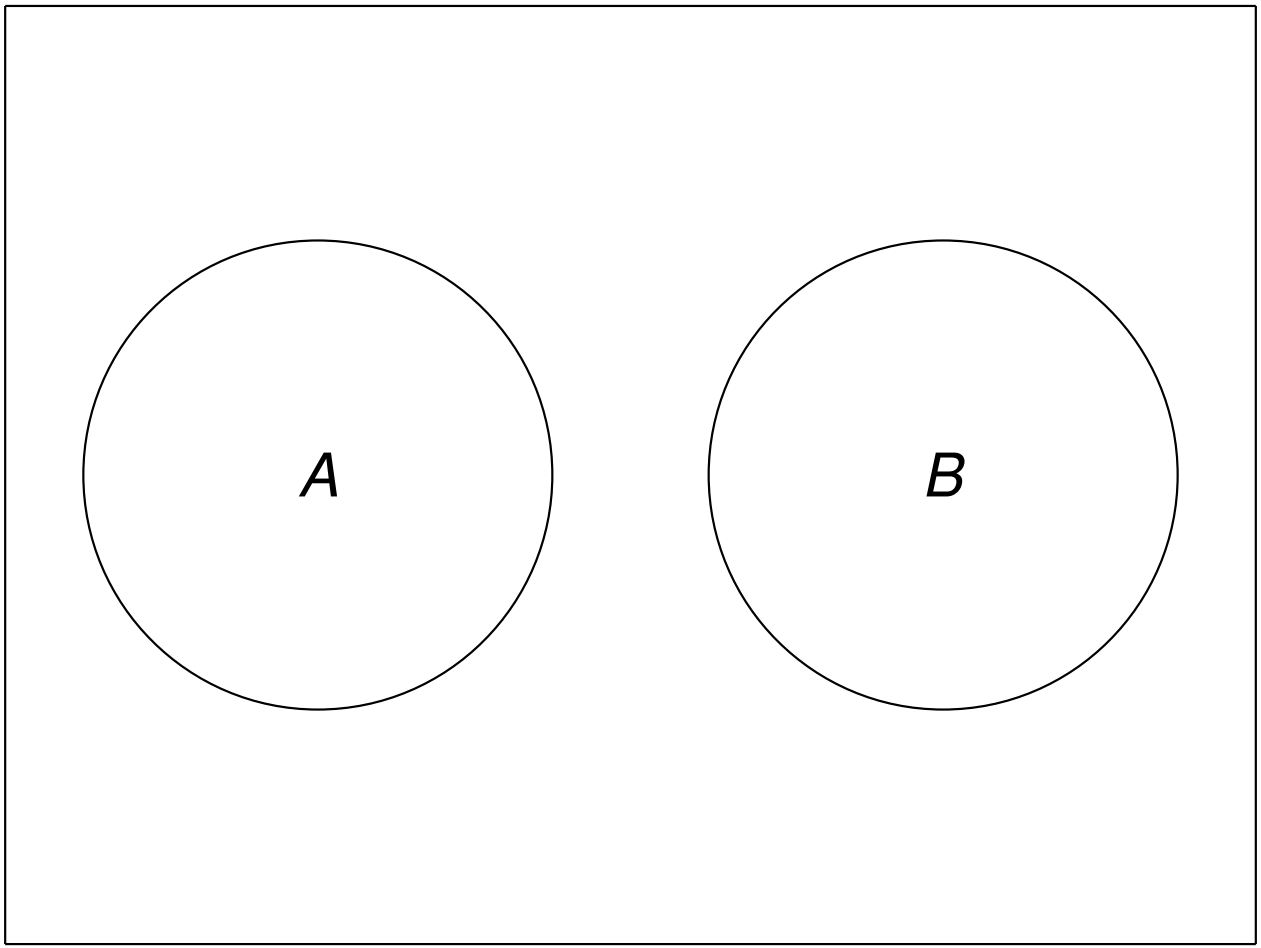
Properties of the product of events : For events $A, B \in \mathscr{E}$ it holds

$$A \cap B = B \cap A \tag{23}$$
$$A \cap (B \cap C) = (A \cap B) \cap C \tag{24}$$
$$A \cap \emptyset = \emptyset \tag{25}$$
$$A \cap A = A \tag{26}$$
$$A \cap \Omega = A \tag{27}$$
$$A \cap B \subseteq A \tag{28}$$
$$A \cap B \subseteq B \tag{29}$$

(23): commutativity, (24): associativity of the product operation $\cap$

Incompatibility of events

Two events $A, B \in \mathscr{E}$ are said to be **incompatible (unvereinbar)** or **disjoint** iff $A \cap B = \emptyset$. This means that the sets associated with the events do not intersect.



Operations between random events

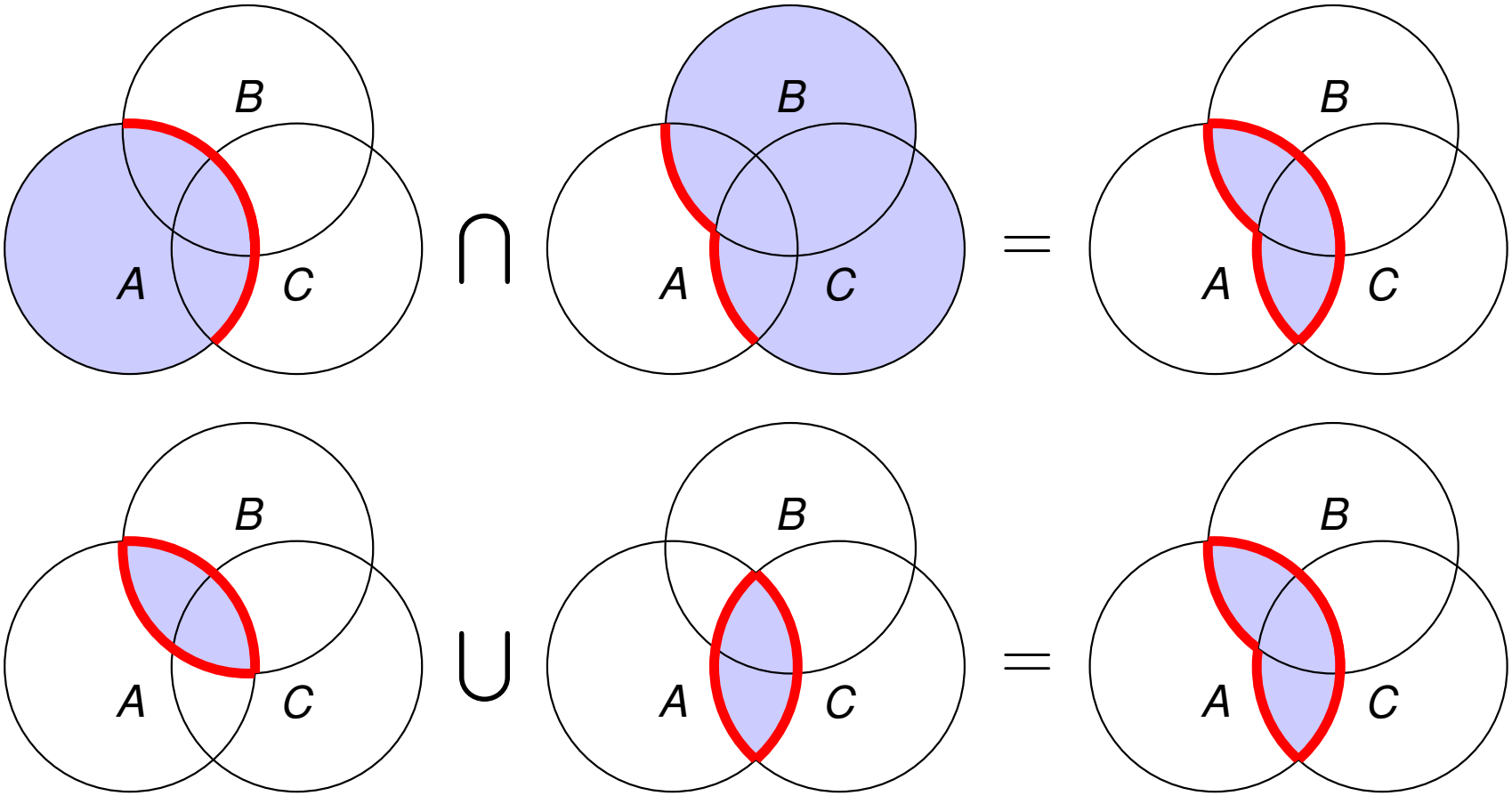
Distributive Laws : For events $A, B, C \in \mathscr{E}$ it holds

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{30}$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \tag{31}$$

Verify this!

Visualization of first distributive law

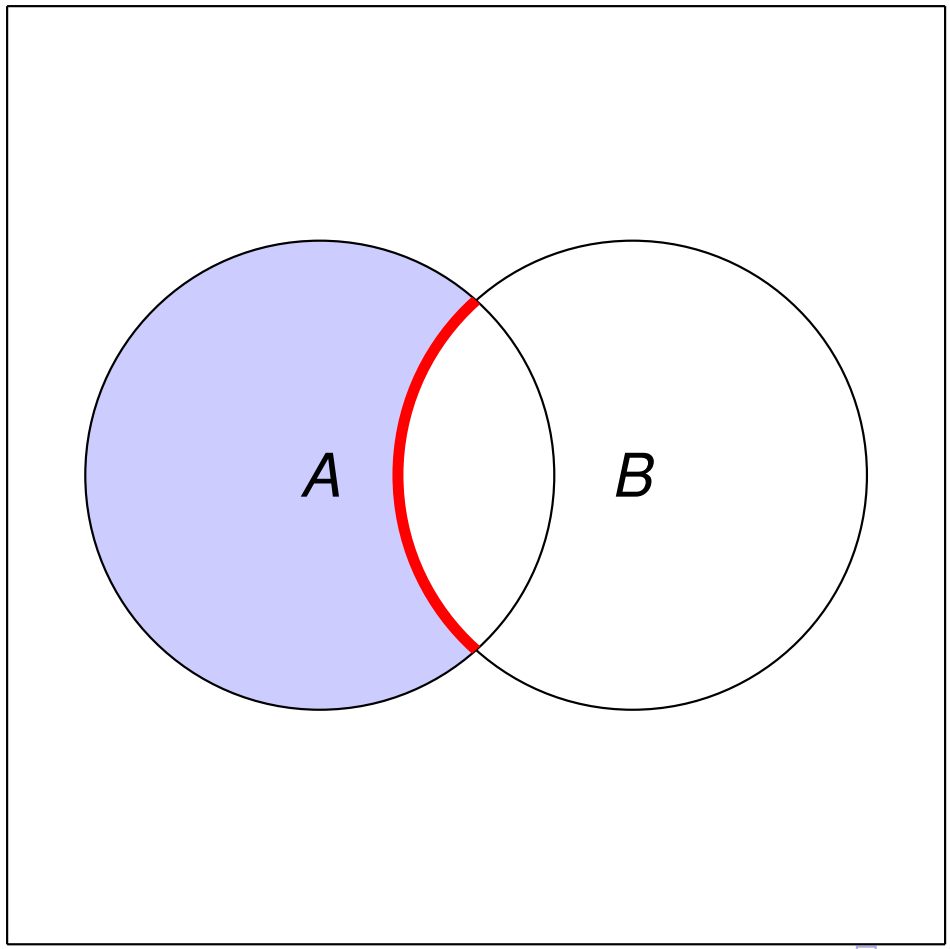


Operations between random events

Difference of events : For events $A, B \in \mathscr{E}$ we define

$$A \setminus B \tag{32}$$

as the event "A minus B". The difference means: A **occurs but not** B.



Operations between random events

Consequence : For events $A, B \in \mathscr{E}$ we have

$$A \backslash B = A \cap \overline{B} \tag{33}$$

where $\overline{B}$ means the complement of B (see definition below).

Properties of the difference of events : For events $A, B \in \mathscr{E}$ it holds

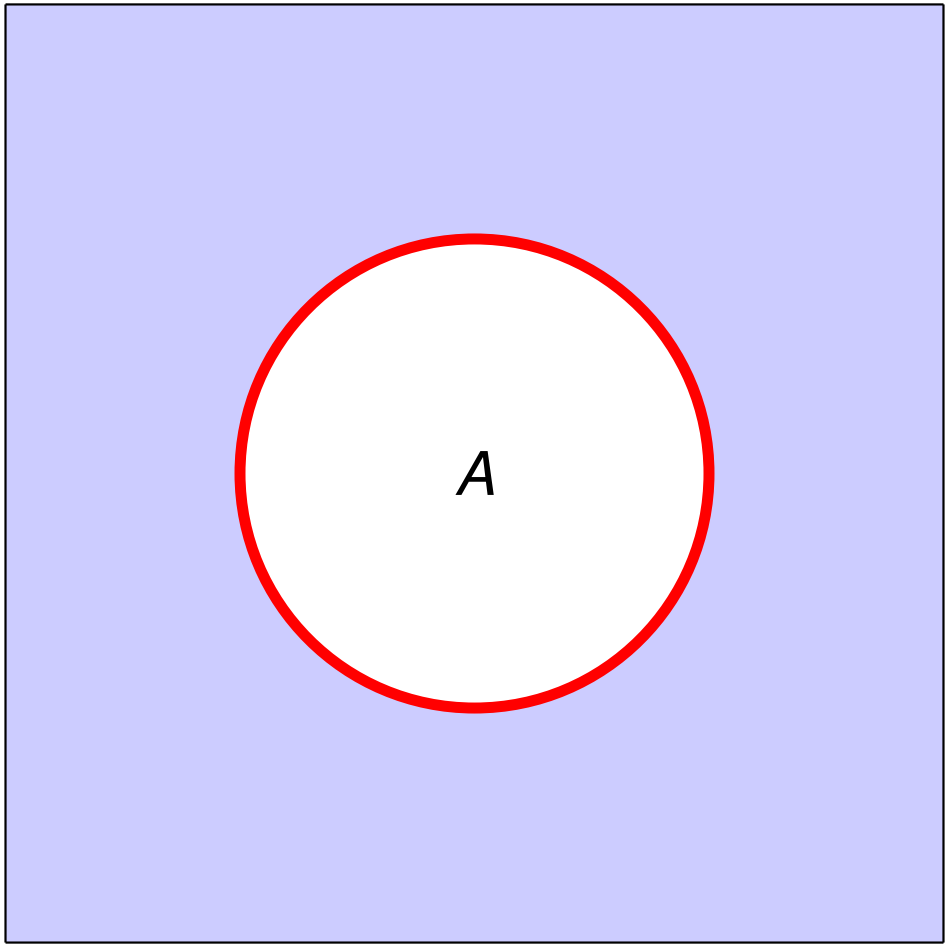
$$A \backslash A = \emptyset \tag{34}$$

$$A \backslash \emptyset = A \tag{35}$$

$$A \backslash B \subseteq A \tag{36}$$

Operations between random events

Complementary event : For any event $A \in \mathscr{E}$ there exists an event $\overline{A}$ which occurs if and only if A **does not occur**.



Complementary event

Properties of the complement :

$$A \cap \overline{A} = \emptyset \tag{37}$$
$$A \cup \overline{A} = \Omega \tag{38}$$
$$\overline{\emptyset} = \Omega \tag{39}$$
$$\overline{\Omega} = \emptyset \tag{40}$$
$$\overline{\overline{A}} = A \tag{41}$$
$$\overline{A} = \Omega \backslash A \tag{42}$$
$$A \subseteq B \implies \overline{B} \subseteq \overline{A} \tag{43}$$

Complementary event

Example (Throwing a die)

Let be A the event that the die shows an even number of points and B the event of having an odd number of points. Then we have

$$A = \{2,4,6\} \tag{44}$$
$$B = \{1,3,5\} \tag{45}$$

This immediately implies that $\overline{A} = B$, as well as $\overline{B} = A$ and $A \cup B = \Omega$.

Consequence (De Morgan's rules) : For any events $A, B \in \mathscr{E}$ we have

$$\overline{A \cap B} = \overline{A} \cup \overline{B} \tag{46}$$
$$\overline{A \cup B} = \overline{A} \cap \overline{B} \tag{47}$$

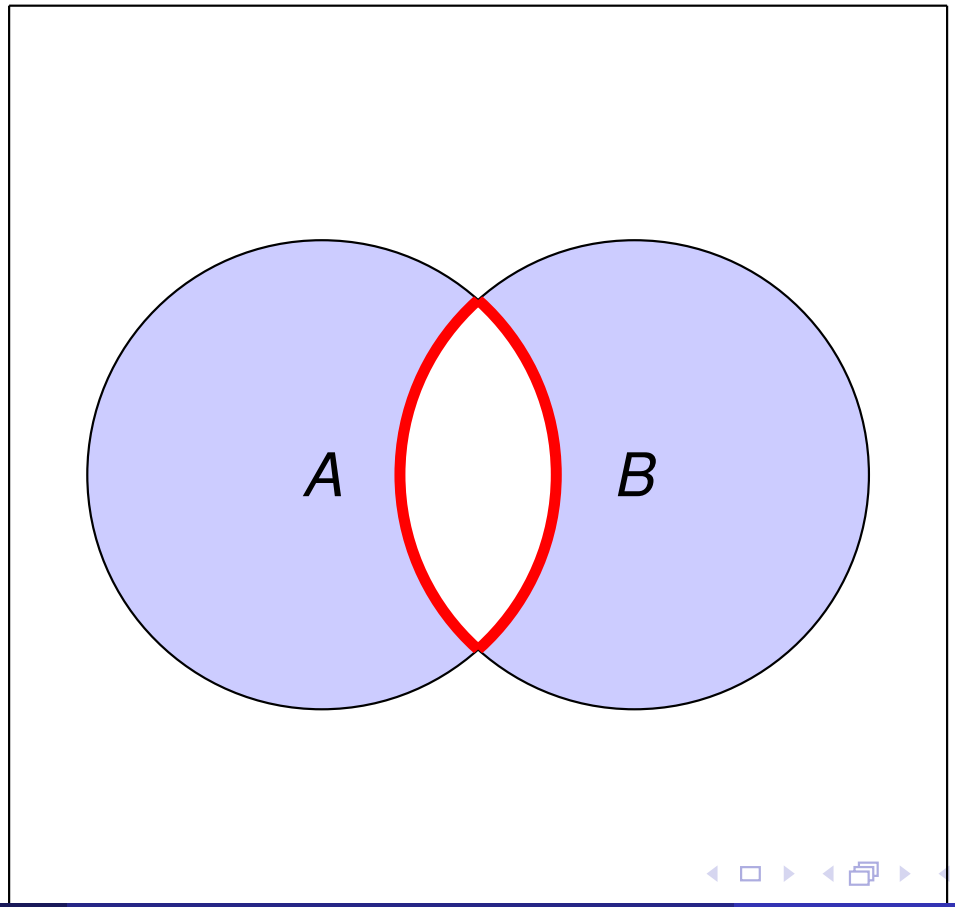
Verify this!

Symmetric Difference

Definition: The **symmetric difference** between two events A and B means that either A or B occurs, but not both.

$$A \Delta B = (A \backslash B) \cup (B \backslash A) \tag{48}$$

$$= (A \cap \overline{B}) \cup (\overline{A} \cap B) \tag{49}$$



Event Algebra

Definition 1: A set $\mathscr{E}$ of random events is called an **event algebra** iff it has the following properties:

- **Property I:** The sure event is an element of $\mathscr{E}$.

$$\Omega \in \mathscr{E} \tag{50}$$

- **Property II:** For all events $A, B \in \mathscr{E}$ their sum is contained in $\mathscr{E}$, too.

$$\forall A, B \in \mathscr{E} : A \cup B \in \mathscr{E} \tag{51}$$

- **Property III:** For all events from $\mathscr{E}$ the complementary event is contained in $\mathscr{E}$ as well.

$$\forall A \in \mathscr{E} : \overline{A} \in \mathscr{E} \tag{52}$$

Event Algebra

In case that $\mathscr{E}$ is infinite, we addionally request the so-called **σ-property**

$$\{A_i\}_{i\in\mathbb{N}} \subseteq \mathscr{E} \quad \Rightarrow \quad \bigcup_{i=1}^{\infty} A_i \in \mathscr{E} \tag{53}$$

Consequences: If $\mathscr{E}$ is an event algebra then it holds:

$$\emptyset \in \mathscr{E} \tag{54}$$

$$\forall A, B \in \mathscr{E} \quad : \quad A \cap B \in \mathscr{E} \tag{55}$$

$$A \backslash B \in \mathscr{E}$$

$$A \,\Delta\, B \in \mathscr{E}$$

$$\{A_i\}_{i\in\mathbb{N}} \subseteq \mathscr{E} \quad \Rightarrow \quad \bigcap_{i=1}^{\infty} A_i \in \mathscr{E} \tag{56}$$

Event Algebra

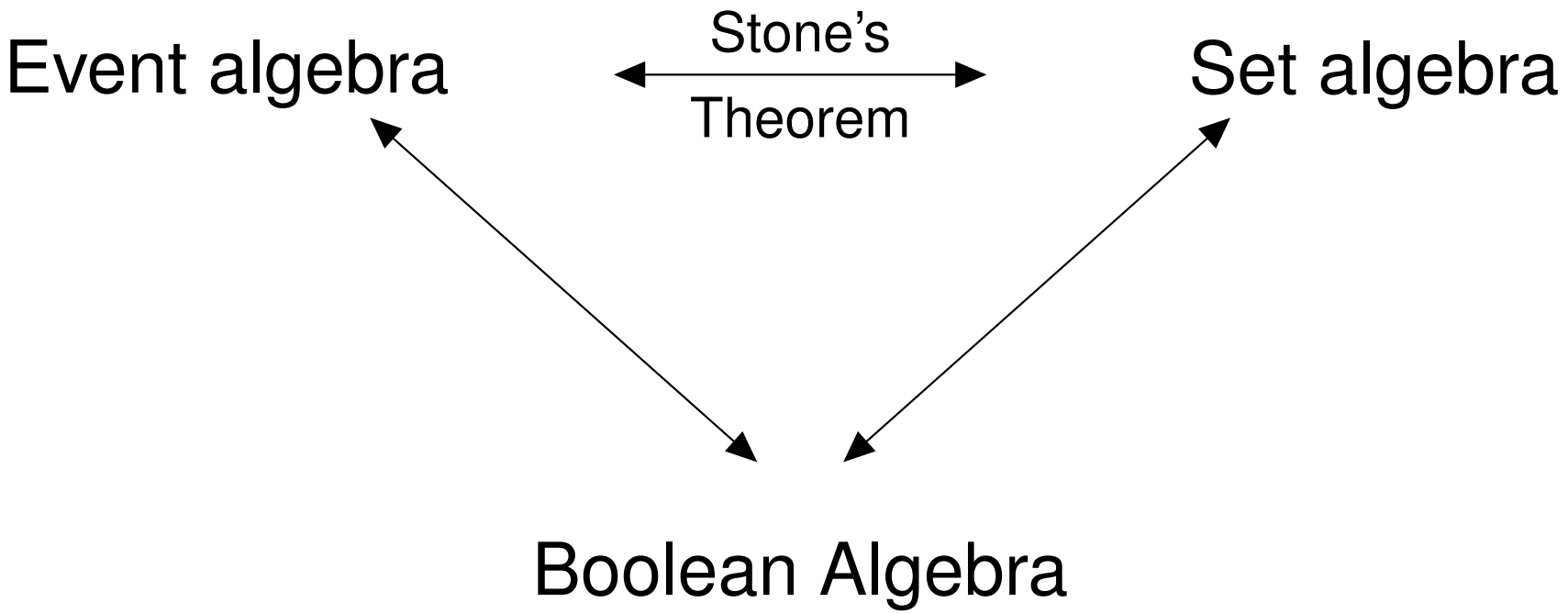
Theorem 1: An event algebra is a **partially ordered set** of random events, which is algebraically **closed w.r.t. product, sum and complement** operations.

Observing the fact that any event algebra is isomorphic to some set algebra, we can say even more: any event algebra also represents a **Boolean algebra** since

- $\cup$ defines an "Addition $+$ " and $\cap$ a "Multiplication $*$ " operation satisfying the principles of closedness, associativity and distributivity
- there exist neutral elements w.r.t. addition and multiplication:
  - $\emptyset$ is neutral w.r.t. $\cup$ (see (16))
  - Ω is neutral w.r.t. $\cap$ (see (27))
- for any element there exists a complementary element.

Summary

The previous results can be summarized as follows:



Atoms

Question: Are there "smallest" events in an event field $\mathscr{E}$?

Definition: 1. Let be $\mathscr{E}$ a finite event algebra and $A, B \in \mathscr{E}$. Then we call A **smaller than** B iff A is a proper subset of B, meaning that

$$A \subset B \iff A \subseteq B \wedge A \neq B \tag{57}$$

2. The event $A \in \mathscr{E}$ is called an **atom** iff there is no event $B \in \mathscr{E}$ such that $B \neq \emptyset$ and $B \subset A$. Otherwise, A is called a **composite event**.

Consequence: If $A \in \mathscr{E}$ is an atom, then any $B \in \mathscr{E}$ satisfies either one of the following two conditions:

$$A \cap B = \emptyset \tag{58}$$

$$A \subseteq B \tag{59}$$

Atoms

Q: Can any event of an event algebra be represented as a sum of atoms?

Example (Throwing a die)

The elementary events $\{1\}, \{2\}, \ldots, \{6\}$ are atoms.

Let A, B and C be the following events: A = "The die shows an even number of points", B = "The die shows an odd number of points", C = "The number of points is less than four". Then

$$A = \{2\} \cup \{4\} \cup \{6\} = \{2, 4, 6\} \tag{60}$$

$$B = \{1\} \cup \{3\} \cup \{5\} = \{1, 3, 5\} \tag{61}$$

$$C = \{1\} \cup \{2\} \cup \{3\} = \{1, 2, 3\} \tag{62}$$

$$\Omega = \{1\} \cup \{2\} \cup \{3\} \cup \{4\} \cup \{5\} \cup \{6\} = \{1, 2, 3, 4, 5, 6\} \tag{63}$$

Atoms

General Result: Any event in a finite event algebra can be represented as a sum of atoms. The representation is unique, up to the order of the summands.

Consequence: The atoms $A_1, \ldots A_n$ of a finite event algebra form a so-called **complete system** meaning that

$$A_i \cap A_j = \emptyset, \quad i \neq j \in \{1, ..., n\} \tag{64}$$

$$\bigcup_{i=1}^{n} A_i = \Omega \tag{65}$$

Q: How many elements are there in an event algebra $\mathscr{E}$ with n atoms ?

To answer this, we first consider an example.

Example (Throwing a die)

Let

$$\mathscr{E} = \mathscr{P}(\Omega) \tag{66}$$
$$\Omega = \{1,2,3,4,5,6\}. \tag{67}$$

where $\mathscr{P}(\Omega)$ means the **power set (Potenzmenge)** of Ω. Then

$$\begin{aligned}\mathscr{E} = \{ &\underbrace{\emptyset}_{\binom{6}{0}}, \underbrace{\{1\},\ldots\{6\}}_{\binom{6}{1}}, \underbrace{\{1,2\},\ldots\{5,6\}}_{\binom{6}{2}}, \\ &\underbrace{\{1,2,3\},\ldots\{4,5,6\}}_{\binom{6}{3}}, \underbrace{\{1,2,3,4\},\ldots\{3,4,5,6\}}_{\binom{6}{4}}, \\ &\underbrace{\{1,2,3,4,5\},\ldots\{2,3,4,5,6\}}_{\binom{6}{5}}, \underbrace{\Omega}_{\binom{6}{6}}\}\end{aligned} \tag{68}$$

Example (Throwing a die, cont'd)

Ω consists of 6 elements, the power set $\mathscr{P}(\Omega)$ thus contains

$$\binom{6}{0} + \binom{6}{1} + \binom{6}{2} + \binom{6}{3} + \binom{6}{4} + \binom{6}{5} + \binom{6}{6} = 64 = 2^6 \tag{69}$$

elements.
Hence, with 6 elements in Ω there are exactly 2^6 subsets in $\mathscr{E}$.

Consequence: Any finite event algebra $\mathscr{E}$ with $n \geq 1$ atoms contains exactly 2^n elements, i.e.

$$|\mathscr{E}| = 2^n. \tag{70}$$

Verify this!

1.3 Probability Space (Wahrscheinlichkeitsraum)

1.3.1 Possible definitions of probability

- **subjective probability** (Savage, de Finetti)
  Evaluation of chances, degree of belief (conviction), betting odds (Wettquoten)
  For example, the odds for "Winning the game" are 40 : 60.
- **frequentist interpretation** as relative frequency (von Mises, 1931):
  We observe the event A in n repetitions of an experiment.
  The frequentist definition of probability then assigns the value

$$P(A) = \lim_{n\to\infty} \frac{h_n(A)}{n} \tag{71}$$

  where $h_n(A)$ denotes the **absolute frequency** with which the event A occurs.
- **axiomatic definition** according to Kolmogorov (1933)

Probability concepts

Note: The frequentist definition thus defines the probability $P(A)$ as the limit of the **relative frequency**

$$w_n(A) = \frac{h_n(A)}{n} \tag{72}$$

of the occurrence of A. For the relative frequency $w_n(A)$ we observe the following properties:

$$0 \leq \quad w_n(A) \quad \leq 1 \tag{73}$$

$$w_n(\Omega) = 1, \quad w_n(\emptyset) \quad = 0 \tag{74}$$

$$A \subseteq B \quad \Rightarrow \quad w_n(A) \leq w_n(B) \tag{75}$$

These properties already reflect the essential properties of the axiomatic probability according to Kolmogorov's definition, which we will consider below.

1.3.2 Kolmogorov's System of Axioms

Let be given a non-empty set Ω (universe, basic set) with elements $\omega \in \Omega$.
We select a (Borel) set algebra $\mathscr{E}$ (σ-algebra) of subsets of Ω, i.e. $\mathscr{E} \subseteq \mathscr{P}(\Omega)$ with the following properties:

$$\emptyset, \Omega \in \mathscr{E} \tag{76}$$

$$\forall A, B \in \mathscr{E} : A \cup B \in \mathscr{E}, A \cap B \in \mathscr{E} \tag{77}$$

$$\forall A \in \mathscr{E} : \overline{A} \in \mathscr{E} \tag{78}$$

$$\forall \{A_i\}_{i \in \mathbb{N}} \subseteq \mathscr{E} : \bigcup_{i=1}^{\infty} A_i \in \mathscr{E} \tag{79}$$

Note: The first three properties define a **set algebra**, the last property extends the set algebra into a σ**-Algebra** or **Borel-Algebra**.

Kolmogorov's Probability Space

Interpretation of $\mathscr{E}$:

First, recall that we can identify events with their associated sets (set algebra $\simeq$ event algebra, Stone's Theorem).

From this fact it becomes clear that

- The above event algebra $\mathscr{E}$ consists of random events.
- Ω is the sure event.
- $\emptyset$ is the impossible event.

The elements $\omega \in \Omega$ are called **elementary events**.
It is often the case that $\{\omega\} \in \mathscr{E}$ for all elementary events ω, but this is not a necessary requirement.

Kolmogorov's Probability Space

Definition of probability on $\mathscr{E}$:

- **Axiom I**: Each element $A \in \mathscr{E}$ is assigned a non-negative number $P(A)$, which we call the **probability** of the occurrence of A,

$$P(A) \geq 0 \quad \forall A \in \mathscr{E} \tag{80}$$

- **Axiom II**: The probability measure is **normalized**,

$$P(\Omega) = 1 \tag{81}$$

- **Axiom III**: For any sequence $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathscr{E}$ with $A_i \cap A_j = \emptyset$ for $i \neq j$ the σ**-additivity** holds,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \tag{82}$$

Kolmogorov's Probability Space

The axioms I - III just mean that, mathematically, the probability $P(\cdot)$ is a **normed measure**.

The triplet $[\Omega, \mathscr{E}, P]$ is called **Kolmogorov's probability space**.

Weakenings of the probability concept comprise

- abandoning normalization of $P(\cdot)$: this leads to generalized probability distributions as used, occasionally, in Bayesian Statistics
- (Choquet) capacities, lower and upper probabilities
- abandoning σ-additivity (Fuzzy Theory, Belief Theory)

1.3.3 Properties of $P(\cdot)$

(P I): Probability of complements

$$P(\emptyset) = 0 \tag{83}$$
$$P(\overline{A}) = 1 - P(A) \quad \forall A \in \mathscr{E} \tag{84}$$

This follows directly from Axiom III observing that $A \cup \overline{A} = \Omega$, $A \cap \overline{A} = \emptyset$ and $\overline{\Omega} = \emptyset$.

Generalization: For a complete system of events $\{A_1 \ \ldots, A_n\}$ with $n \geq 2$, where $A_i \cap A_j = \emptyset$ for $i \neq j$ and $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$, it holds

$$P(A_1) + P(A_2) + \cdots + P(A_n) = 1 \tag{85}$$

Verify this for the "playing dice" experiment!

Properties of $P(\cdot)$

(P II): Monotonicity of $P(\cdot)$

$$A \subseteq B \Rightarrow P(A) \leq P(B) \quad \forall A, B \in \mathscr{E} \tag{86}$$

Proof:

$$\begin{aligned} B = \Omega \cap B &= \left(A \cup \overline{A}\right) \cap B \\ &= (A \cap B) \cup \left(\overline{A} \cap B\right) \\ &= A \cup \left(\overline{A} \cap B\right) \end{aligned} \tag{87}$$

With $A \cap \left(\overline{A} \cap B\right) = \emptyset$ we get from Axiom III

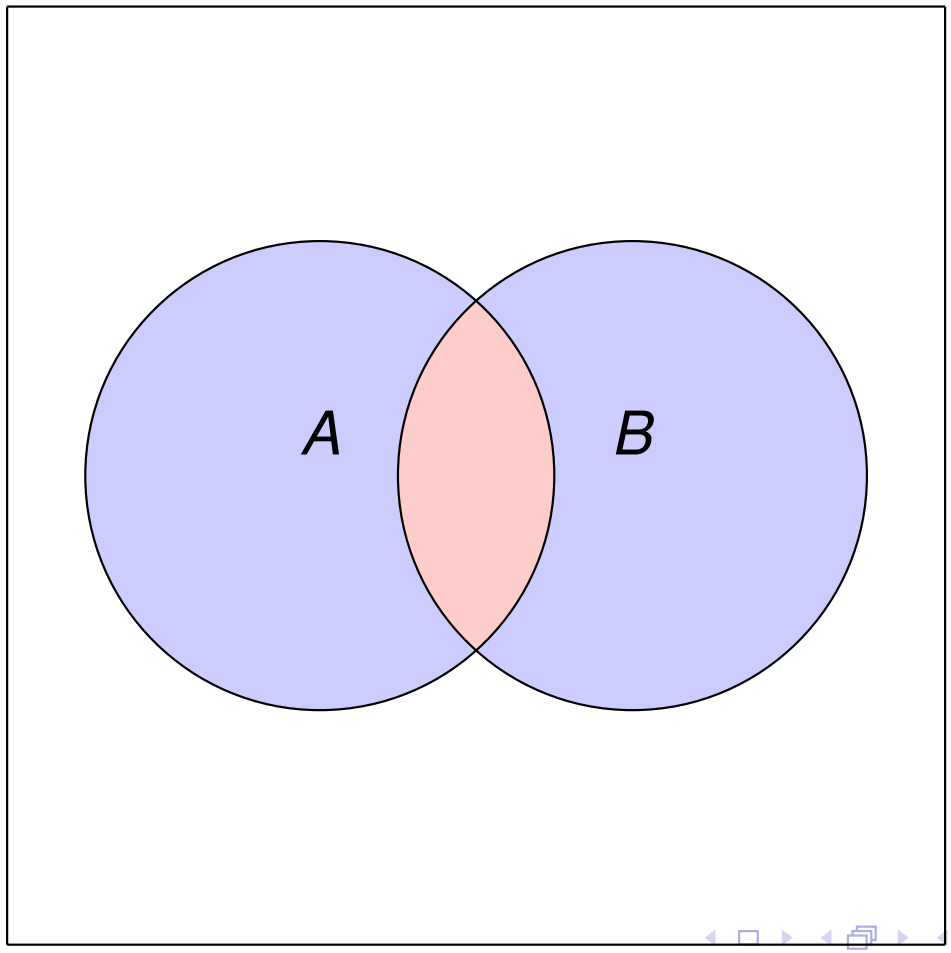
$$\begin{aligned} P(B) &= P\left(A \cup \left(\overline{A} \cap B\right)\right) \\ &= P(A) + \underbrace{P(\overline{A} \cap B)}_{\geq 0} \\ &\geq P(A) \end{aligned} \tag{88}$$

Properties of $P(\cdot)$

(P III): General addition rule

For all events $A, B \in \mathscr{E}$ it holds

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{89}$$



Properties of $P(\cdot)$

Proof: We decompose A, B and $A \cup B$ into disjoint events.

$$\begin{aligned} A &= A \cap \Omega = A \cap (B \cup \overline{B}) = (A \cap B) \cup (A \cap \overline{B}) && (90) \\ B &= \Omega \cap B = (A \cup \overline{A}) \cap B = (A \cap B) \cup (\overline{A} \cap B) && (91) \\ A \cup B &= (A \cap \overline{B}) \cup (\overline{A} \cap B) \cup (A \cap B) && (92) \end{aligned}$$

Applying Axiom III again, we obtain

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \overline{B}) && (93) \\ P(B) &= P(A \cap B) + P(\overline{A} \cap B) && (94) \\ P(A \cup B) &= P(A \cap \overline{B}) + P(\overline{A} \cap B) + P(A \cap B) && (95) \end{aligned}$$

Next, we subtract the equations 93 and 94 from eq. 95. This implies

$$P(A \cup B) - P(A) - P(B) = -P(A \cap B) \quad (96)$$

Properties of $P(\cdot)$

Addendum:

- In case that $A \cap B = \emptyset$ (i.e. A and B are incompatible/disjoint) we have $P(A \cap B) = P(\emptyset) = 0$ and then the general addition rule reduces to Axiom III:

$$P(A \cup B) = P(A) + P(B) \tag{97}$$

- For arbitrary events $A_1, A_2, \ldots A_n$, $n \geq 2$, the addition rule generalizes as follows:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{k=1}^{n} (-1)^{k-1} \cdot S_{k,n} \tag{98}$$

where

$$S_{k,n} = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} P\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) \tag{99}$$

Properties of $P(\cdot)$

- For $n = 2$ eq. (98) specializes to the result stated in (P III). Verify this!
- Formulate the result given in (98) for $n = 3$!

Obviously, we have for arbitrary sequences $\{A_i\}_{i\in\mathbb{N}}$ of events

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i) \tag{100}$$

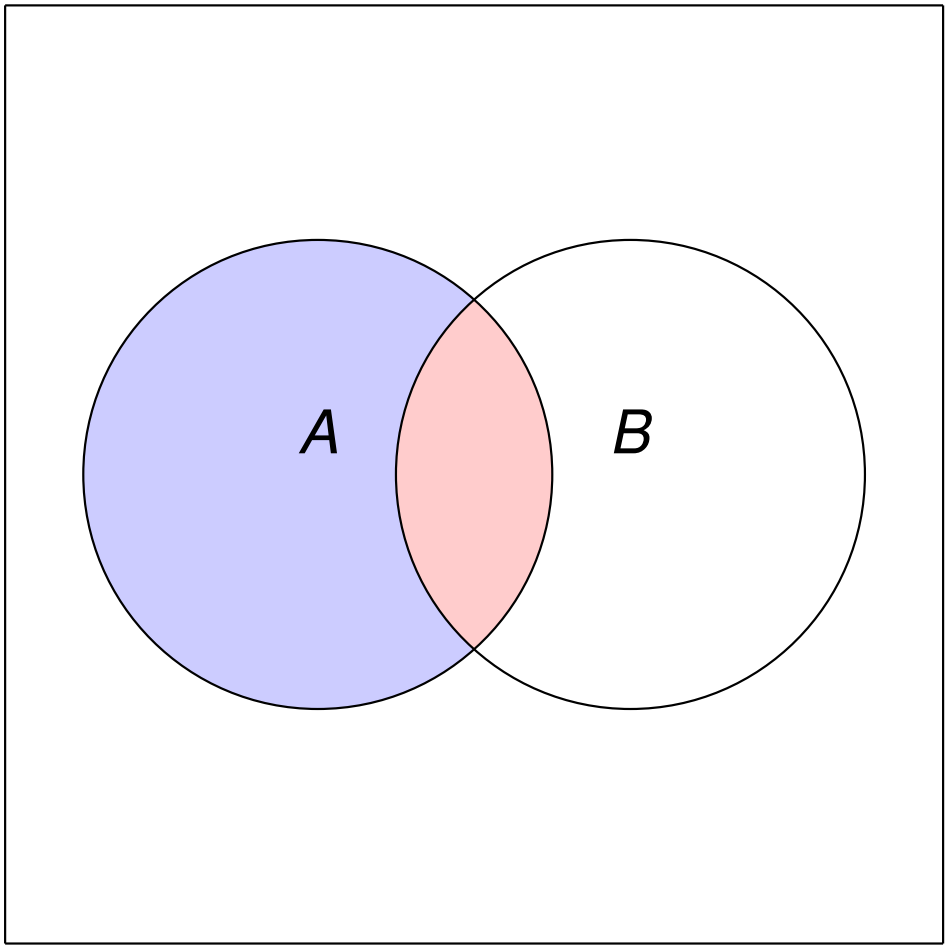
(P IV): Probability of the difference set

$$P(A \setminus B) = P(A) - P(A \cap B) \tag{101}$$

$$P(A \,\Delta\, B) = P(A) + P(B) - 2 \cdot P(A \cap B) \tag{102}$$

Properties of $P(\cdot)$

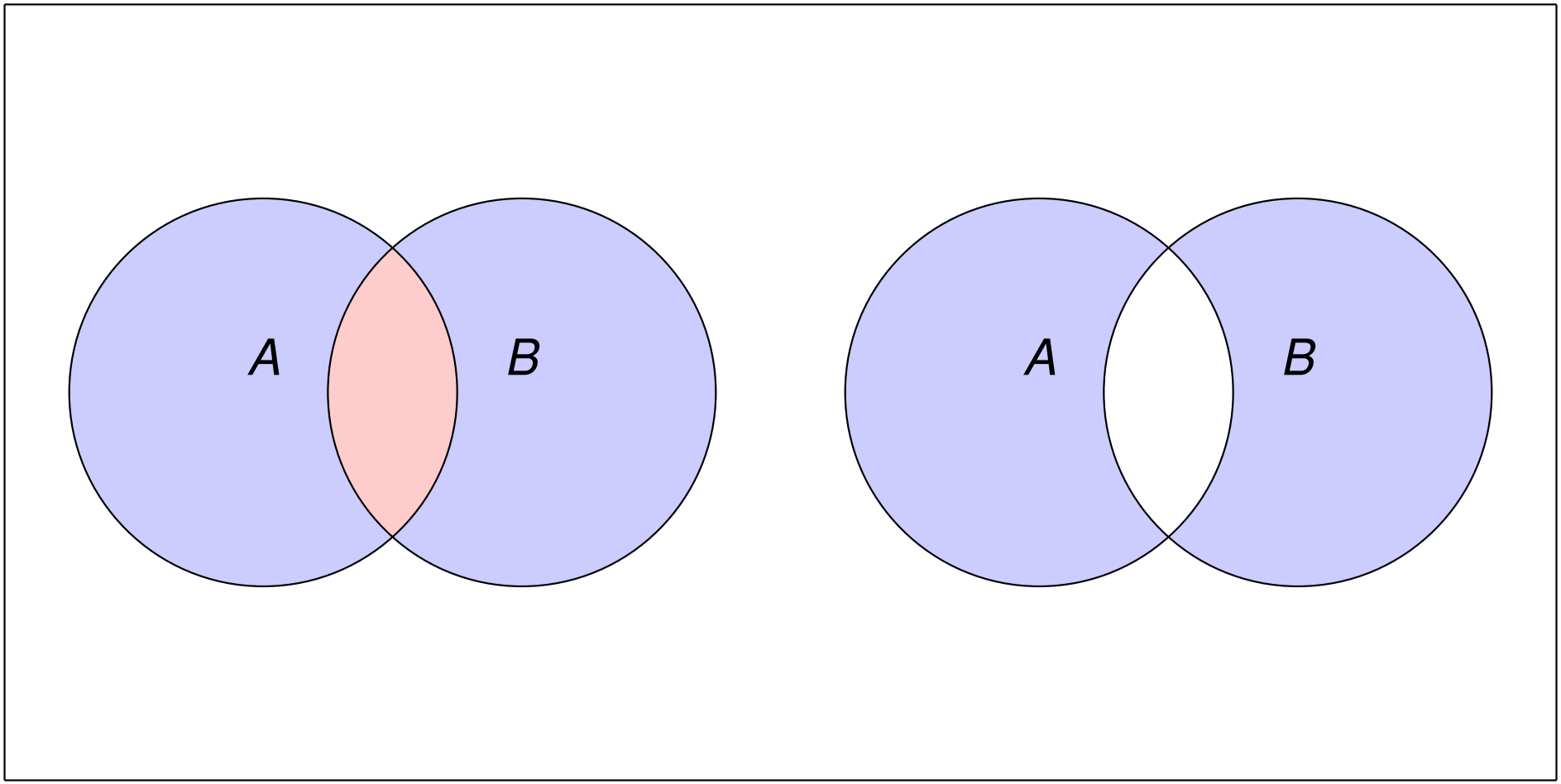
Graphical illustration of eq. (101):



Give a mathematical proof!

Properties of $P(\cdot)$

Graphical illustration of eq. (102):



Give a mathematical proof!

1.4 Classical and Geometric Probability

1.4.1 Classical Probability Space

Consider a **finite** event algebra $\mathscr{E}$ with a complete system $\{A_1, \ldots, A_N\}$ of N atoms.
The basic **assumption** of the classical concept of probability is that **all atoms have the same probability** p, where $0 < p < 1$.
From the completeness we have

$$1 = P(\Omega) = \sum_{i=1}^{N} \underbrace{P(A_i)}_{=p} = N \cdot p \tag{103}$$

This implies that $N \cdot p = 1$ or, equivalently, $p = \frac{1}{N}$. Thus, we know the probability of a single atom:

$$P(\text{single atom}) = \frac{1}{\#\text{atoms}} \tag{104}$$

Classical Concept of Probability

Since every event $B \in \mathscr{E}$ can be represented as a sum of atoms, it holds

$$P(B) = P\left(\bigcup_{j=1}^{m} A_j\right) = \sum_{j=1}^{m} P(A_j) = m \cdot p = \frac{m}{N} \tag{105}$$

Definition: For any event $B \in \mathscr{E}$ we define its **classical probability** as

$$P(B) = \frac{\text{number of atoms which are favorable for } B}{\text{number of all possible atoms}} \tag{106}$$

Briefly, we say "favorable divided by possible cases". This rule goes back to Laplace and is therefore also termed as **Laplace's rule of (classical) probability**.

Example (Maxwell-Boltzmann Statistic)

For illustration, we consider the problem of distributing n "things" to N "shelves". The total number of such possibilities is N^n. (We here assume that the n things are distinguishable.) Each of the possible combinations will be represented as an n-tuple of integers $(i_1, i_2, \ldots i_n)$, where i_j stands for the number of the shelf, in which the j- th thing is located.

For example, let us distribute $n = 2$ bank notes (50 and 100 Euros) to $N = 3$ shelves. There are $3^2 = 9$ different ways of doing this.

$$\begin{array}{lll} (1,1) & (1,2) & (1,3) \\ (2,1) & (2,2) & (2,3) \\ (3,1) & (3,2) & (3,3) \end{array}$$

E.g. $(3,2)$ means that the 50 Euro note is located in the shelf 3 and 100 Euro note in the shelf 2. Note the difference to combination $(2,3)$.

Example (cont'd)

Since we assume the combinations to be equiprobable, the probability for each single combination reads

$$p = \frac{1}{N^n} \tag{107}$$

In our bank note example this is just $p = \frac{1}{9}$.

Next, we consider a single specified shelf and answer the question

Q: How many ways are there to distribute **exactly** k things to this shelf, where $k \in \{0, 1, \ldots, n\}$ and the order of placement does not matter.

A:There are $\binom{n}{k}$ possible choices (combinations without repetition).

Now, for the remaining $N - 1$ shelves there are exactly $n - k$ objects left over.

Example (cont'd)

Thus, the number of combinations (=**favorable cases**) with exactly k objects in a particular shelf is given by

$$\binom{n}{k} \cdot (N-1)^{n-k} \tag{108}$$

Consequently, according to Laplace's rule,

$$p_k = \frac{\binom{n}{k} \cdot (N-1)^{n-k}}{N^n} \tag{109}$$

describes the probability that the (random) distribution of n distinguishable objects to N shelves leads to having exactly k out of n objects in a single shelf, $k = 0, 1, \ldots, n$.

This probability can be rewritten as

Maxwell-Boltzmann statistic

Example (cont'd)

$$p_k = \frac{\binom{n}{k} \cdot (N-1)^{n-k}}{N^n} = \binom{n}{k} \cdot \frac{(N-1)^{n-k}}{N^k \cdot N^{n-k}} = \binom{n}{k} \cdot \left(\frac{1}{N}\right)^k \cdot \left(1 - \frac{1}{N}\right)^{n-k} \tag{110}$$

This is but the probability mass of the so-called **Binomial Distribution** with parameters n and $p = \frac{1}{N}$, which we will discuss in more detail in Section 2.3 below.

In physical applications, this distribution is well-known as **Maxwell-Boltzmann statistic**, where "things" refer to molecules, electrons, photons or other particles and "shelves" refer to possible (grid) locations of the particles in space.

Example (Statistical Quality Control)

Consider a production process where N devices are produced among which there are M defective parts, where $0 \leq M \leq N$. We draw a **sample of size** n, where $n < N$.

Q: What is the probability of having exactly m defective parts in the sample?

First, we observe that there are $\binom{M}{m}$ possible combinations for drawing m defectives out of the total of M defective parts. The remaining $n - m$ devices in the sample are non-defective. The total number of non-defective devices in the production is $N - M$. Therefore, there are exactly $\binom{N-M}{n-m}$ possibilities for drawing $n - m$ "good" devices out of the $N - M$ well-functioning devices.
To obtain the number of favorable cases (for having m defectives and $n - m$ non-defectives in the sample) we just need to combine these two different types of drawings, i.e.

Example (cont'd)

combine the drawings of the "good" and "bad" ones. This yields

$$\binom{M}{m} \cdot \binom{N-M}{n-m} \tag{111}$$

favorable cases. On the other hand, the total number of possible drawings (cases) is $\binom{N}{n}$.

Applying Laplace's rule, we obtain the probability we have searched for in the above question **Q**:

$$\frac{\binom{M}{m} \cdot \binom{N-M}{n-m}}{\binom{N}{n}}. \tag{112}$$

This is known as **hypergeometric probability**. This probability distribution will be discussed in more detail in Section 2.3 below.

Lottery Problems

Example (Lotto "6 out of 45")

Lotto problems can be handled analogously to those of statistical quality control.

Q: What is the probablity of getting 4 numbers correctly in a lottery drawing 6 out of 45 (neglecting the additional number)?

First note that there are $\binom{6}{4} = 15$ possible ways of having 4 numbers chosen correctly out of the 6 numbers actually drawn in a lottery. The remaining 2 numbers should not be part of the winning draw. They are part of the $45 - 6 = 39$ "wrong" numbers; there are $\binom{39}{2} = 741$ such combinations. Hence, there are

$$\binom{6}{4} \cdot \binom{39}{2} = 15 \cdot 741 = 11\ 115 \tag{113}$$

favorable cases for having a lotto ticket with four correct numbers and two wrong numbers.

Lottery Problems

Example (cont'd)

The number of possible cases (= total number of different lotto tickets) is given by

$$\binom{45}{6} = 8\ 145\ 060 \tag{114}$$

According to Laplace's rule, we finally obtain the winning probability

$$\frac{\binom{6}{4} \cdot \binom{39}{2}}{\binom{45}{6}} = 0.001365 \tag{115}$$

For comparison, the probabilities of having five or six correct numbers in a drawing are given by 0.000 028 729 and 0.000 000 123, respectively.

Q: What is the probability of getting a Jackpot in "Euromillions"?

1.4.2 Geometric Probability

Classical Probability is defined as the ratio between the number of favorable atoms and the total number of atoms. Thereby it is assumed that the total number of atoms is finite or, at least, countably finite. However, in many applications the event algebra cannot be described by finite sets, e.g. when events refer to areas, volumes and the like.

Definition: Let be $\Omega \subset \mathbb{R}^n$ with $n \geq 1$ and $0 < \lambda(\Omega) < \infty$, then we define the **geometric probability** of any A, $A \in \mathscr{P}(\Omega)$, by

$$\begin{aligned} P(A) &= \frac{\lambda(A)}{\lambda(\Omega)} \\ &= \frac{\text{measure of favorable area}}{\text{measure of total area}} \end{aligned} \tag{116}$$

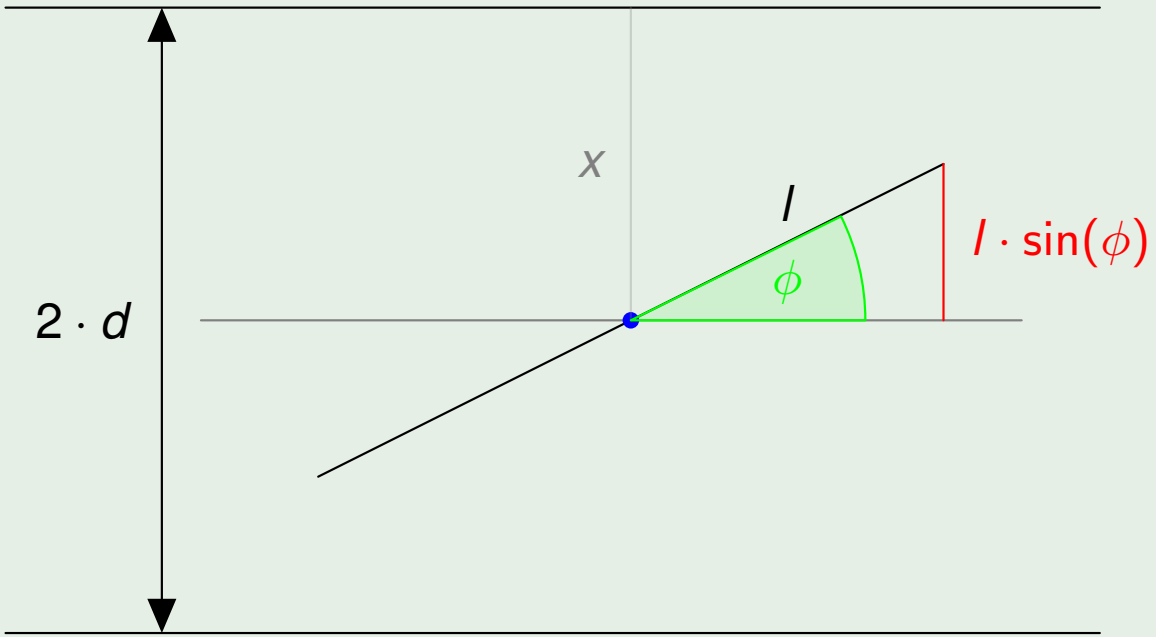
where $\lambda(\cdot)$ refers to the Lebesgue (volume) measure.

Example: a) $\Omega \subset \mathbb{R}^1$, $A = [0, 4]$. Then $\lambda(A) = 4$ = length of A.

b) $\Omega \subset \mathbb{R}^2$, $A = [0,4]x[0,2]$ a rectangle with side lengths 4 and 2. Then $\lambda(A) = 4 \cdot 2 = 8$ = area of A.

Example (Buffon's needle problem)

Consider a plane with horizontally parallel straight lines that have all equal distance $2 \cdot d$ from each other. Drop a needle of length $2 \cdot l$, $l \leq d$, on the plane randomly, and check whether it cuts one the lines.



Example (cont'd)

Q: What is the probability of A = "Needle cuts one of the lines"?

The basic set (universe) Ω can be represented as

$$\Omega = \{(x, \phi) \mid 0 \leq x \leq d,\ 0 \leq \phi \leq \pi\} \tag{117}$$

This set indeed covers all possible positions of the needle's midpoint (apart from horizontal displacements).
The needle cuts a line iff

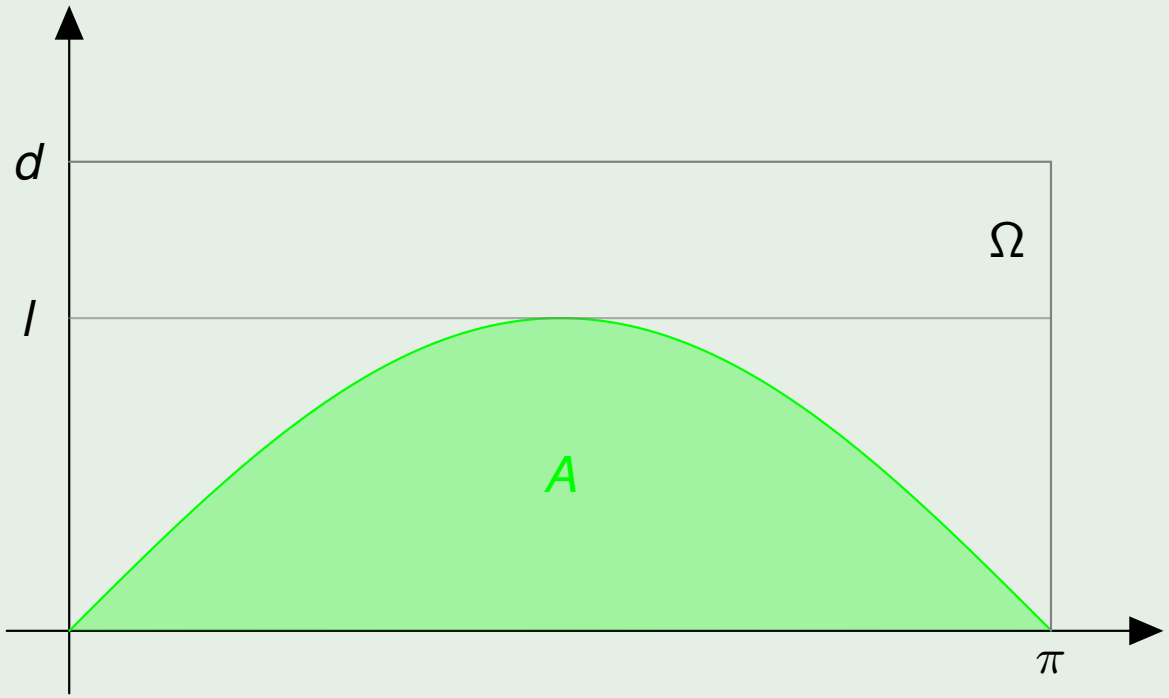
$$0 \leq x \leq l \cdot \sin(\phi) \tag{118}$$

This inequality (event) defines a subarea A of Ω as illustrated below.

The (Lebesgue) measure of the universe Ω is easily obtained as

$$\lambda(\Omega) \quad = \quad d \cdot \pi \tag{119}$$

Example (cont'd)



The Lebesgue measure of A is just the green area:

$$\lambda(A) = \int_0^{\pi} l \cdot \sin(x) dx \tag{120}$$

Example (cont'd)

The (geometric) probability of our event A is then determined as

$$
\begin{aligned}
P(A) = \frac{\lambda(A)}{\lambda(\Omega)} &= \frac{\int_0^{\pi} l \cdot \sin(x) dx}{d \cdot \pi} \\
&= \frac{1}{d \cdot \pi} \cdot \int_0^{\pi} l \cdot \sin(x) dx \\
&= \frac{l}{d \cdot \pi} \cdot [-\cos(x)]_0^{\pi} \\
&= \frac{2 \cdot l}{d \cdot \pi}
\end{aligned}
\tag{121}
$$

In the special case where $l = d/2$ we get $P(A) = 1/\pi$. This result was the origin of **Monte Carlo Simulation** ideas.

Example (Computation of Integrals)

Consider the problem of computing the integral

$$I = \int_0^1 f(x)dx, \; f(x) = \sqrt{1 - x^2} \tag{122}$$

The idea is to interpret this as Lebesgue measure $\lambda(A)$ of an associated event (area) A as in eq. (120) and then using the notion of geometric probability to obtain

$$I = \lambda(A) = P(A) \cdot \lambda(\Omega) \tag{123}$$

with a suitably chosen rectangle $\Omega = [a, b]x[c, d]$. This rectangle must fully cover the area A under the function $f(x) = \sqrt{1 - x^2}$.
Here, it is natural to choose $\Omega = [0, 1]x[0, 1]$ as the unit square.

Example (cont'd)

Next we generate a large set of "uniformly distributed" points in Ω and approximate $P(A)$ by the relative frequency of points falling into the area A.

$$\int_a^b f(x) \cdot dx \approx \text{relative frequency} \cdot \text{area of rectangle } \Omega \quad (124)$$

$$\approx \frac{\text{number of points in } A}{\text{total number of points in } \Omega} \cdot \lambda(\Omega) \quad (125)$$

The quality of the approximation can be improved by increasing the number of points and the use of more sophisticated models of a "random" distribution of points. We will hear about the uniform distribution and other distribution models later.
We are now going to demonstrate the above ideas using the R-system.

```
> n <- 20000
> f <- function(x) {sqrt(1-x^2)}
> a <- 0; b <- 1; c <- 0; d <- 1
> x <- runif(n,a,b)
> y <- runif(n,c,d)
> count <- 0
> for (i in 1:n){if (y[i]< f(x[i])){count <- count+1}}
> integral <- (count/n)*(b-a)*(d-c)
> integral #0.785
> pi/4 #0.7853982
```

Example (Rendezvous Problem)

Two persons A and B decide to have a meeting. However, they do not fix a time, instead they just agree to meet within a fixed time interval of one hour length. He who comes first will wait 15 minutes for the second person. In case that the second person does not show up within the 15 minutes interval the first person leaves and the meeting does not take place. What is the probability that the two persons will actually meet?

Let x and y denote the arrival times of Person A and Person B, respectively. The waiting time d is 15 minutes, i.e. $d = 0.25$ hours. For a successful meeting it must hold

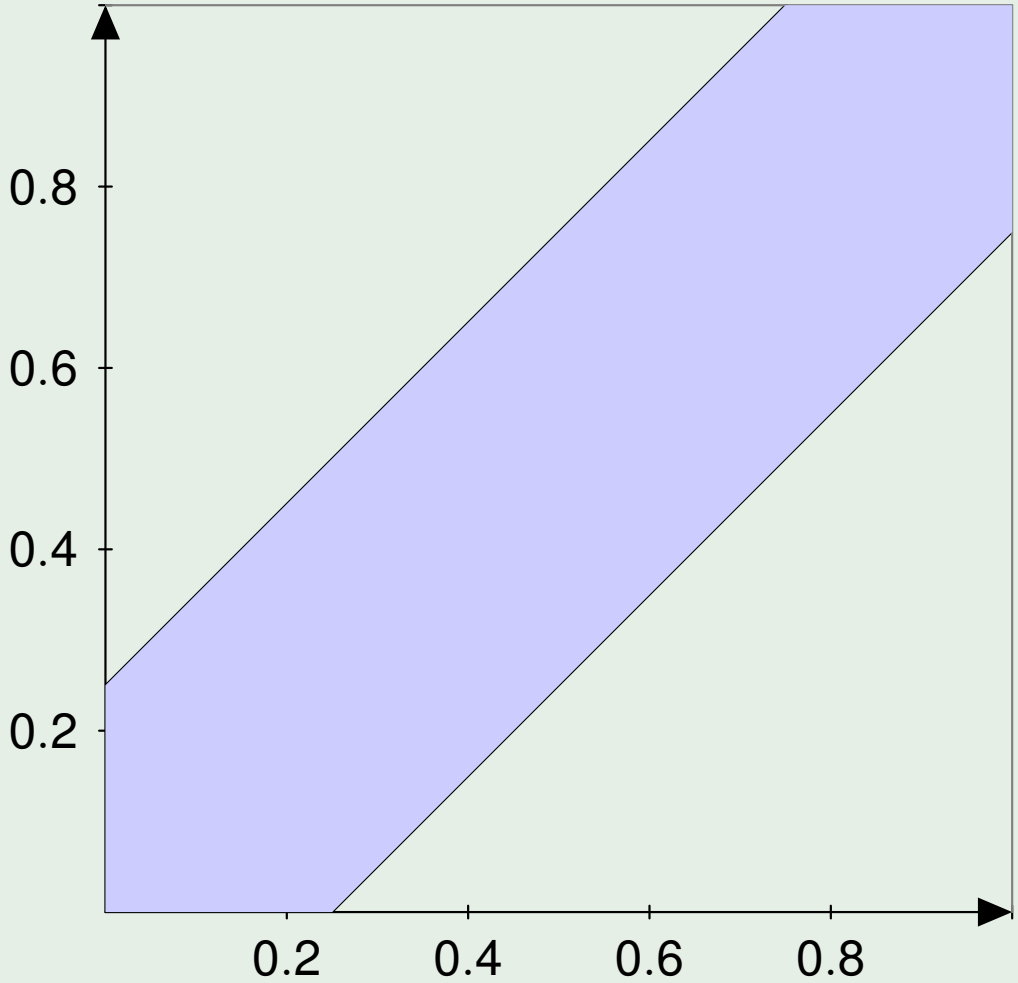
$$|x - y| \leq d \tag{126}$$

This is equivalent to stating that

$$x - d \leq y \leq x + d \tag{127}$$

Example (cont'd)

The situation can be displayed graphically.



Geometric Probability

Example (cont'd)

The geometric probability of a successful meeting results from the ratio of the blue shaded area (=favorable area) and the area of the unit square Ω (=possible area).

$$\begin{aligned} P(\text{'meeting takes place'}) &= \frac{\lambda(\text{'blue area'})}{\lambda(\Omega)} \\ &= 1 - \lambda(\text{'no meeting'}) \\ &= 1 - (1-d)^2 \\ &= 1 - 0.75^2 \\ &= 0.4375 \end{aligned} \tag{128}$$

observing that $\lambda(\Omega) = 1$.

Further applications of the concept of geometric probability can be found in statistical image analysis of materials or in medical imaging.

1.5 Conditional Probability

Up to now we have considered probabilities of "fixed" events. In many practical applications, however, there are relevant cases where probabilities will change after obtaining additional information.

Definition: Let be $A, B \in \mathscr{E}$ and $P(B) > 0$. Then we call

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\text{joint probability}}{\text{probability of condition}} \tag{129}$$

the **conditional probability of** A **under the condition** B.

Example (Drawing lots)

Consider a bowl filled with lots, which are either *white*, *red* or *green*. Additionally, we have three categories of lots: winnings, blanks (Nieten) and free tickets (Freilose). We assume that each lot in the bowl has the same chance to be drawn.

Example (cont'd)

The distribution of 100 lots looks as follows:

Color \ Value	B = Blank	W = Winning	F = Free Ticket	$\sum$
white	12	8	6	26
red	6	10	4	20
green	17	16	21	54
$\sum$	35	34	31	100

Thus, we have

$$P(\text{'white lot'}) = \frac{26}{100} \tag{130}$$

$$P(\text{'red lot'}) = \frac{20}{100} \tag{131}$$

$$P(\text{'green lot'}) = \frac{54}{100} \tag{132}$$

Example (cont'd)

Likewise, we get

$$P(\text{Blank}) = \frac{35}{100} \tag{133}$$

$$P(\text{Winning}) = \frac{34}{100} \tag{134}$$

$$P(\text{Free Lot}) = \frac{31}{100} \tag{135}$$

Should we now get the additional information that there are only red lots left in the bowl, then the probability of a winning changes:

$$P(\text{Winning}|\text{Red Lot}) = \frac{P(\text{Winning} \cap \text{Red Lot})}{P(\text{Red Lot})} \tag{136}$$

Example (cont'd)

which then becomes

$$P(\text{Winning}|\text{Red Lot}) = \frac{\frac{10}{100}}{\frac{20}{100}} = \frac{10}{20} = 0.5$$

This is the conditional probability of drawing a winning under the condition (information) that it is a red one.

Thereby, the probability $P(\text{Winning} \cap \text{Red Lot})$ in the numerator of eq. (136) denotes the probability of having both a winning **AND** a red lot. There are exactly 10 such lots out of the total of 100 lots, i.e. the joint probability becomes $\frac{10}{100} = 0.1$. The unconditional probability of obtaining a red lot is $\frac{20}{100} = 0.2$, according to eq. (131).

Note: The probabilities given in eqs. (130) - (135) are **unconditional** probabilities (sometimes also called **marginal** probabilities).

Conditioning

Consequence: Let be $A, B \in \mathscr{E}$ with $P(A) \cdot P(B) > 0$. Then it holds

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} \tag{137}$$

Proof: This follows immediately from the symmetry of the $\cap$ operation and our definition of conditional probability:

$$P(A \cap B) = P(A|B) \cdot P(B) \tag{138}$$
$$P(B \cap A) = P(B|A) \cdot P(A) \tag{139}$$

Remark: Let $B \in \mathscr{E}$ be a fixed event with $P(B) > 0$. Then

$$P_B(\cdot) = P(\cdot|B) \tag{140}$$

defines a probability measure on $[\Omega, \mathscr{E}]$.

This can be shown by verifying the Kolmogorov axioms for $P_B(\cdot)$.
In the special case of $B = \Omega$ this becomes trivial, since
$P_\Omega(A) = \frac{P(A \cap \Omega)}{P(\Omega)} = P(A)$ for all $A \in \mathscr{E}$.

Remark: Eq. (138) can be easily generalized to $n > 2$ conditions $B_1, \ldots, B_n \in \mathscr{E}$ with $P(B_1 \cap \cdots \cap B_n) > 0$ as follows

$$\begin{aligned} P(B_1 \cap B_2 \cap \cdots \cap B_n) = & \quad P(B_n|B_{n-1} \cap \cdots \cap B_1) \\ & \cdot \quad P(B_{n-1}|B_{n-2} \cap \cdots \cap B_1) \\ & \cdot \quad P(B_{n-2}|B_{n-3} \cap \cdots \cap B_1) \\ & \cdot \quad \ldots \\ & \cdot \quad P(B_2|B_1) \cdot P(B_1) \end{aligned} \tag{141}$$

This can be shown by complete induction.

For three events A, B and C we have, e.g.

$$P(A \cap B \cap C) = P(A|B \cap C) \cdot P(B|C) \cdot P(C) \tag{142}$$

Example (Birthday Problem)

Q: What is the probability that in a group of n people at least two of them have the same birthday? Thereby, it is assumed that each day of the year is equally probable for a birthday.

It is simpler to compute the probability of the complementary event $\overline{A}$= "No two people have the same birthday".

To this, let B_n denote the event that person n does not have the same birthday as any of persons 1 through $n-1$, where $n > 1$. Then it becomes immediately clear that

$$P(\overline{A}) = P(B_1) \cdot P(B_2|B_1) \cdot \ldots \cdot P(B_n|B_{n-1} \cap \cdots \cap B_1) \qquad (143)$$
$$= \frac{365}{365} \cdot \frac{364}{365} \cdot \ldots \cdot \frac{365-(n-1)}{365}$$

Thus, we get

Example (cont'd)

$$P(A) = 1 - P(\overline{A}) = \quad 1 - \prod_{k=0}^{n-1} \frac{365-k}{365} \tag{144}$$

Remarkably, the chances for meeting two or more people with the same birthday exceed 50% already when $n = 23$. See simulations at

`https://csferrie.medium.com/is-the-birthday-paradox-real-93ea6dc16e36`

Write an R-function **birthday**!

1.5.1 Total Probability

Consider a Kolmogorov Probability Space $[\Omega, \mathscr{E}, P]$ and a complete system of events $\{A_1, \ldots, A_n\}$, i.e.

$$A_i \cap A_j = \emptyset \text{ for } i \neq j \tag{145}$$

$$A_1 \cup A_2 \cup \cdots \cup A_n = \Omega \tag{146}$$

Total Probability

Let $B \in \mathscr{E}$ an arbitrary event, $B \neq \emptyset$, then the events $A_1 \cap B, \ldots, A_n \cap B$ are incompatible as well, since

$$(A_i \cap B) \cap (A_j \cap B) = (A_i \cap A_j) \cap B = \emptyset \cap B = \emptyset \tag{147}$$

for all $i, j = 1, ..., n$. Moreover,

$$(A_1 \cap B) \cup \cdots \cup (A_n \cap B) = (A_1 \cup \cdots \cup A_n) \cap B = \Omega \cap B = B \tag{148}$$

From the property of σ-additivity it then follows

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i) \tag{149}$$

where the last equation follows from the definition of conditional probability, cp. eq. (139). Summarizing, we have proven

Total Probability

Theorem 2: (Law of total probability)

For any complete event system $\{A_1, \ldots, A_n\}$ it holds

$$P(B) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i) \quad \forall B \in \mathscr{E} \tag{150}$$

Remarks:

- The events A_i do not necessarily have to be atoms
- The above theorem is also true for (countably) infinite and complete event systems

$$P(B) = \sum_{i=1}^{\infty} P(B|A_i) \cdot P(A_i) \tag{151}$$

Total Probability

Example (Chemical analogue)

We have n bowls with different (chemical) solutions, summing up to 1 liter.

Bowls	Bowl 1	Bowl 2	Bowl 3	Bowl 4
Percentages in %	20	40	30	10
Concentrations in %	20	60	30	40

Now we are mixing the 4 solutions. The total concentration then becomes

$$0.2 \cdot 0.2 + 0.4 \cdot 0.6 + 0.3 \cdot 0.3 + 0.1 \cdot 0.4 = 0.41$$

This is but the same result that we obtain directly from the law of total probability:

$$\sum_{i=1}^{4} \underbrace{P(B|A_i)}_{\text{concentration}} \cdot \underbrace{P(A_i)}_{\text{percentage}}$$

Total Probability

We are now in the position to derive a far-reaching result which has become known as **Bayes' Theorem** for finite events.

Theorem 3: (Bayes's Formula)

Let $\{A_i\}_{i=1,2,...,n}$ be a complete system of events. Then it holds for any event $B \in \mathscr{E}$ such that $P(B) > 0$

$$P(A_k|B) = \frac{P(B|A_k) \cdot P(A_k)}{\sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)} \tag{152}$$

Proof: According to our definition of conditional probability we have

$$P(A_k|B) = \frac{P(B|A_k) \cdot P(A_k)}{P(B)} \tag{153}$$

for all $k = 1, ..., n$. The theorem then follows already by plugging in the formula of total probability for $P(B)$ as given in eq. (150).

Bayes's Formula

Interpretation:

- A_k are a-priori hypotheses and $P(A_k)$ are called prior probabilities
- conditional probabilities $P(A_k|B)$ are called posterior probabilities after "observing" event B.

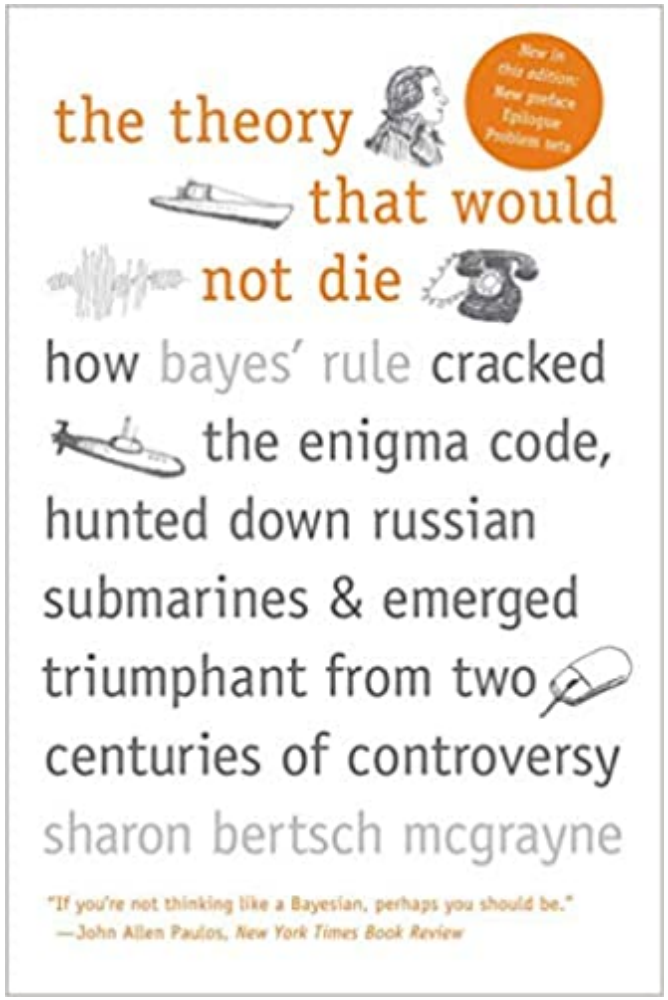
In Bayesian statistics, B stands for the observed data and the hypotheses A_k stand for the model parameters.
Bayesian Statistics finds many applications in areas where "root cause analysis (Ursachenforschung)" is requested:

- medical and technical diagnostics
  observation (failure) $\rightarrow$ searching for the cause of failure
- criminal science investigations and jurisdiction
  indications $\rightarrow$ drawing conclusions on suspects
- research in science, technology, medicine, business and engineering hypotheses $\rightarrow$ experiments/data $\rightarrow$ new hypotheses
  ...

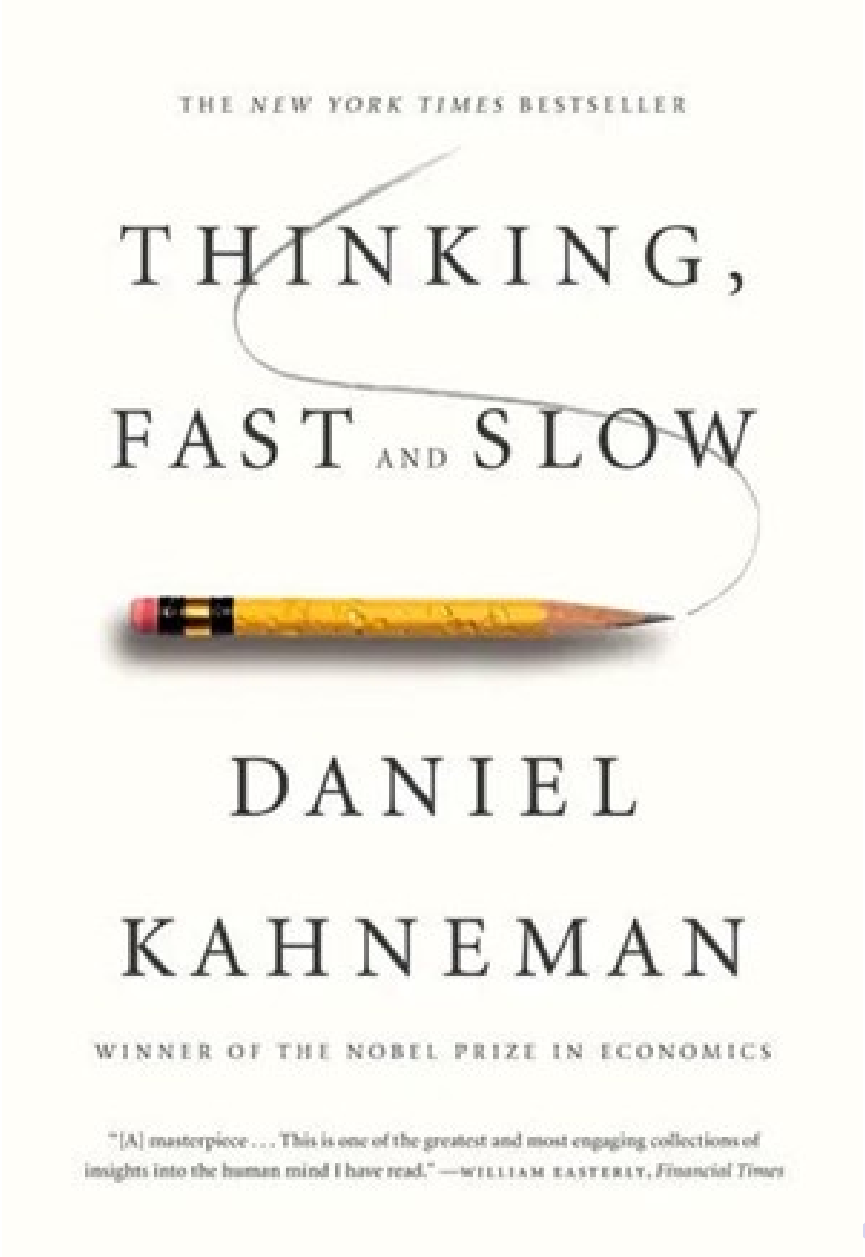
Bayes's Formula

Further applications in art sciences, literature, music and archaeology: dispute of origins and/or authorship

Literature recommendation:



Expert Opinions and Thinking



Bayes's Formula

Example

The German AIDS-Foundation recommends „autotest VIH“, which comes with the CE-sign of the EU and thus confirms the suitability of the test for laymen (www.autotest-sante.com). The operating manual claims that this test is very accurate, in particular, it says

1. the test has a **sensitivity** of 100%, i.e. a person tests 100 % positively when he/she is HIV-infected.
2. the **specificity** of the test is 99.8%, i.e. the probability that the person tests negatively when he/she is not HIV-infected is 0.998.

The last statement means that the false-alarm-rate is just 0.2%. Moreover, it is known that only 11.400 persons in the total population of about 69 million German inhabitants suffer from HIV.

Assume that a person has tested positively.

Q: What is the probability that this person is indeed infected?

Example (cont'd)

We define the following events: HIV = "person suffers from HIV", $\overline{HIV}$ means that the person is not infected. The event POS means a "positive test result", whereas $\overline{POS} = NEG$ stands for a negative test result.

The above statements tell us that

$$P(POS|HIV) = 1.0, \; P(NEG|\overline{HIV}) = 0.998$$

$$P(HIV) = \frac{11400}{69000000} = 0.0001652174.$$

According to Bayes's Theorem, we obtain

$$P(HIV|POS) = \frac{P(POS|HIV) \cdot P(HIV)}{P(POS|HIV) \cdot P(HIV) + P(POS|\overline{HIV}) \cdot P(\overline{HIV})} \quad (154)$$

Bayes's Formula

Example (cont'd)

leading to the (surprisingly low, why?) result

$$P(HIV|POS) = \frac{1.0*0.00016522}{1.0*0.00016522+0.002*0.9998348} = 0.07632.$$

In classification contexts we often encounter the notions of *false positive* and *false negative*. In our case, the first notion $P(POS|\overline{HIV})$ refers to a positive test, although the person actually is not infected and the second one, $P(NEG|HIV)$, means that the infection has not been detected by the test. These notions play an important role in e.g. cybersecurity and pattern recognition contexts (access control, facial recognition, motion video control, fraud detection,...)

In a medical context, $P(POS|infection)$ is usually called the *sensitivity* and $P(NEG|noinfection)$ is called the *specificity* of the test procedure.

Extensions to Bayes Theorem

Sensitivity, also known as recall or true positive rate (TPR) measures the proportion of actual positives correctly identified by the model. Sensitivity is particularly important when the cost of missing positive cases is high, such as in medical diagnoses. For the previous example, it is P(Test Positive|Disease).
Specificity, also known as the true negative rate (TNR), measures the proportion of actual negatives correctly identified by the model. It is particularly important in situations where the cost of false positives is high, such as spam detection or certain medical screenings. For the previous example, it is P(Test Negative|No Disease).

Both TPR and TNR combine to create ROC-AUC. The ROC curve and the area under the curve (AUC) metric can be directly interpreted in the Bayesian framework. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings, which aligns with varying decision thresholds on the posterior probabilities in Bayesian classifiers.

Why Bayesian statistics is so important

Although Bayesian statistics has been around for a long time, it's only recently that it is starting to be widely used. A big reason for this is advances in computational science. Classical statistical methods make wide-ranging assumptions about the context of events that are being modelled — often they involve a uniform prior which assumes that any context at all is possible with equal likelihood. This makes them computationally more simple, and so they have been much preferred by statisticians historically because they offered realistic, practical routes to estimation of likelihood.

Bayesian approaches are computationally more complex, which has prevented their practical use in the past due to limited computational power available to us. Things are changing, and we now have access to computational power that allows us to use Bayes' Theorem to take into account prior context when estimating likelihood of events, and to adjust that context and recalculate when we learn new information.

Extensions to Bayes Theorem

This is how our brains work in general — we adapt to new information all the time — so Bayesian methods are a real step forward in modeling our comprehension of the world.
If you are interested in learning more about the concepts of Bayesian Statistics, I highly recommend the book 'Statistical Rethinking' by Richard McElreath.

Bayesian techniques are fundamental in various **machine learning** algorithms, providing a probabilistic framework for decision-making under uncertainty.

Let's explore some common applications:

1. Naive Bayes Classifier

The Naive Bayes classifier is a simple yet powerful probabilistic classifier based on Bayes' theorem. It assumes independence between the features given the class label. This assumption is rarely true in real-world data, yet the classifier often performs well.

Extensions to Bayes Theorem

Types of Naive Bayes Classifiers

- Gaussian Naive Bayes: Assumes that features follow a normal distribution
- Multinomial Naive Bayes: Used for discrete data like word counts in text classification
- Bernoulli Naive Bayes: Used for binary/boolean features

Example: Spam Email Classification

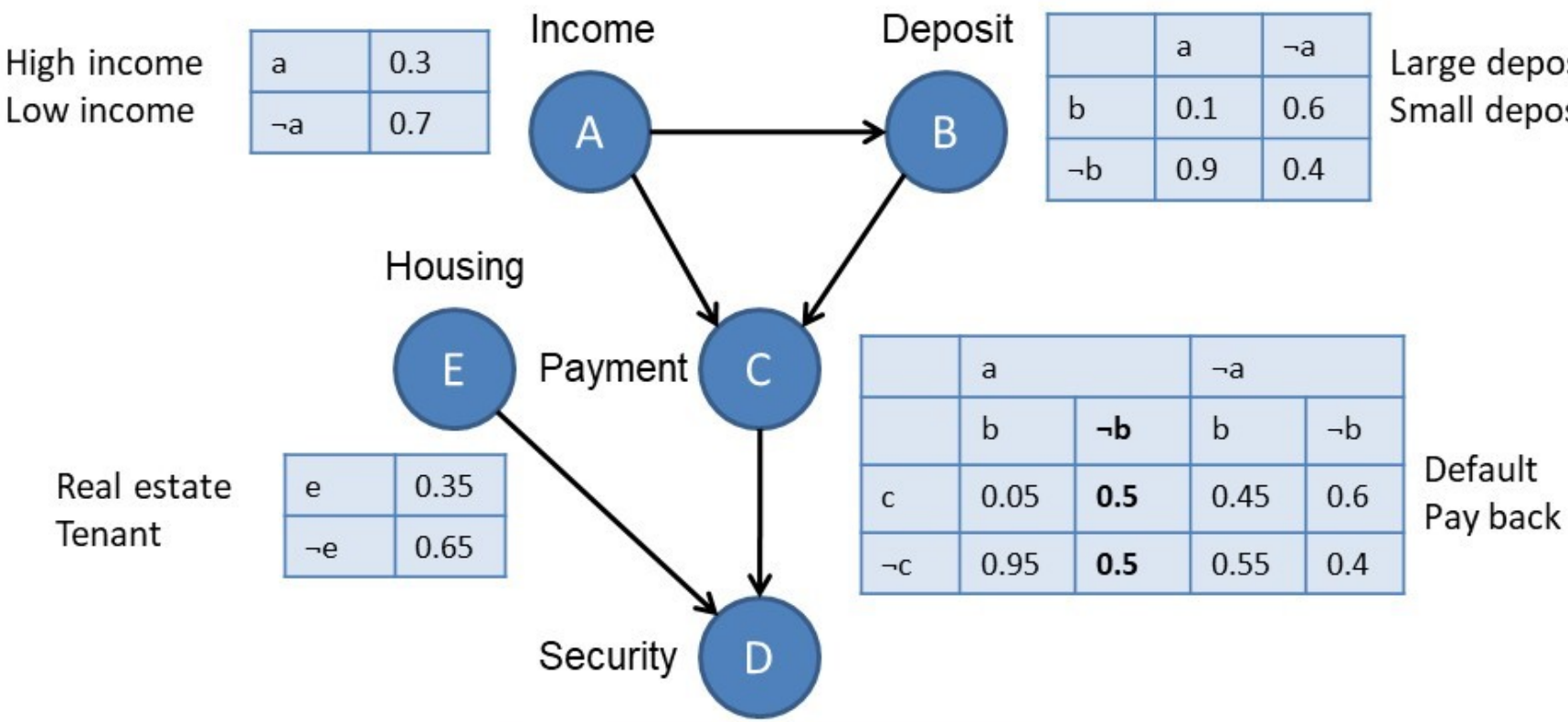
Extensions to Bayes Theorem

2. Bayesian Networks

A Bayesian network is like a smart map that shows how different things are connected and influence each other.
The network contains Nodes, Arrows, and Conditional Probability Tables. .

- Nodes (Circles): Each node represents a different information or event.
- Arrows (Connections): The arrows between nodes show how one piece of information affects another.
- Probabilities: Each connection has a probability that tells you how likely it is for one event to cause another.
- Updating Beliefs: When you get new information, like seeing that some person has increased his deposits and/or securities, you can use the network to update your beliefs about this person's ability to pay back the loan.

Bayesian Networks



High income
Low income

a	0.3
¬a	0.7

	a	¬a
b	0.1	0.6
¬b	0.9	0.4

Large deposit
Small deposit

Real estate
Tenant

e	0.35
¬e	0.65

	a		¬a	
	b	**¬b**	b	¬b
c	0.05	**0.5**	0.45	0.6
¬c	0.95	**0.5**	0.55	0.4

Default
Pay back

	c		¬c	
	e	¬e	e	¬e
d	0.01	0.5	0.75	0.31
¬d	**0.99**	0.5	**0.25**	0.69

Security given
No security

Causal Discovery

An extensive guide that will walk you through applications, libraries, and dependencies of causal discovery approaches can be found at

`https://towardsdatascience.com/an-extensive-starters-guide-for-causal-discovery-usin`

Extensions to Bayes Theorem

3. Forecasting:

For example, Bayesian time series forecasting involves using Bayesian statistical methods to predict future values in a time series.
PyMC3 is a popular Python library for Bayesian statistical modeling and probabilistic machine learning. Applying the Bayesian method with time-series forecasting quantifies the uncertainty as a full posterior distribution of predictions, not just point estimates, which allows for robust handling of uncertainty in predictions.
For a simple example to forecast monthly sales data using a Bayesian approach, see
`https://medium.com/@ryassminh/practical-\`
`bayesian-inference-for-data-scientists-b48aaca9395a`

Using Bayesian Modeling to Predict The Champions League
`https://towardsdatascience.com/using-bayesian-\`
`modeling-to-predict-the-champions-league-8ebb069006ba`

1.6 Independence of events

In everyday life, the notion of independence of two events is used quite often. However, this is usually done in a loose way. In this section we will deal with a mathematically precise definition of independence.

1.6.1 Multiplication rule

In the sequel let $A, B \in \mathscr{E}$ be arbitrary events such that $P(A) \cdot P(B) > 0$, meaning that both $A, B \neq \emptyset$.

Definition: Two events A, B are said to be **independent** from each other iff

$$P(A \cap B) = P(A) \cdot P(B) \tag{155}$$

This definition makes sense observing that

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A) \tag{156}$$

Independence

Analogously, we also have $P(B|A) = P(B)$ for independent events A, B.

Example

In a previous example we had analyzed the probabilities of having a "red lot" and a "winning lot".

$$P(R) = 0.2 \tag{157}$$
$$P(W) = 0.34 \tag{158}$$
$$P(R \cap W) = 0.1 \tag{159}$$
$$P(R) \cdot P(W) = 0.2 \cdot 0.34 \tag{160}$$

The events R (= "red lot") and W (= "winning lot") are **not** independent, since

$$P(R \cap W) \neq P(R) \cdot P(W) \tag{161}$$

Independence

Remark: Note the difference between independence and incompatibility of events.
Incompatibility of two events A, B means $P(A \cap B) = 0$, and thus

$$P(A \cup B) = P(A) + P(B) \tag{162}$$

whereas **independence** of A, B means

$$P(A \cap B) = P(A) \cdot P(B) \tag{163}$$

Corollary: If A, B are independent events then the following events are independent, too:

- A and $\overline{B}$
- $\overline{A}$ and B
- $\overline{A}$ and $\overline{B}$.

Proof: We first prove the independence of A and $\overline{B}$. It holds

$$A = A \cap (B \cup \overline{B}) = (A \cap B) \cup (A \cap \overline{B}) \tag{164}$$

$A \cap B$ and $A \cap \overline{B}$ are incompatible. Hence, we have

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \overline{B}) \\ &= P(A) \cdot P(B) + P(A \cap \overline{B}) \end{aligned} \tag{165}$$

This implies, however,

$$\begin{aligned} P(A \cap \overline{B}) &= P(A) - P(A) \cdot P(B) = P(A) \cdot [1 - P(B)] \\ &= P(A) \cdot P(\overline{B}) \end{aligned} \tag{166}$$

Analogously, we can prove the independence of $\overline{A}$ and B. This, in turn, implies that $\overline{A}$ and $\overline{B}$ must be independent, too.

Independence

We now extend the notion of independence to more than two events. Before we do so, we first show, by means of an example, that pairwise independence of all events is not enough to claim their complete independence.

Example

We are rolling two dice. Let A denote the event of having an even number of points with the first die and B the event of having an odd number with the second die. Further, let C denote the event that both dice either show even numbers or odd numbers. Obviously,

$$P(A) = \frac{3}{6} = \frac{1}{2} \tag{167}$$

$$P(B) = \frac{3}{6} = \frac{1}{2} \tag{168}$$

$$P(C) = \frac{2}{4} = \frac{1}{2} \tag{169}$$

Example (cont'd)

where we have 4 possible combinations for event C, namely the pairs $(even, even), (even, odd), (odd, even)$ and (odd, odd), out of which the first and last pair are favorable cases. Moreover,

$$P(A \cap B) = \frac{9}{36} = \frac{1}{4} \tag{170}$$

$$P(A \cap C) = \frac{9}{36} = \frac{1}{4} \tag{171}$$

$$P(B \cap C) = \frac{9}{36} = \frac{1}{4} \tag{172}$$

However, it holds

$$P(A \cap B \cap C) = 0 \neq \frac{1}{8} = P(A) \cdot P(B) \cdot P(C) \tag{173}$$

i.e. we have pairwise independence, but no total independence.

Total Independence

Definition: Events $A_1, \ldots, A_n$ where $n \geq 2$ are said to be **completely independent** (totally independent) iff

$$P\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) = P\left(A_{i_1}\right) \cdot P\left(A_{i_2}\right) \cdot \cdots \cdot P\left(A_{i_k}\right) \tag{174}$$

for all k-tuples $(i_1, \ldots, i_k)$ from $\{1, 2, 3, \ldots, n\}$ with $k = 2, 3, \ldots, n$.

Example

Consider a device consisting of 250 serial components $B_1, \ldots, B_{250}$. Each component B_i fails with the same probability $p_i = 0.003$ in a given time interval $[0, t]$. The failures of the components B_i are assumed to be totally independent.

Q: What is the failure probability of the whole device within $[0, t]$?

Due to the serial arrangement of the components the whole system already fails when just a single component fails. From the total independence of the failures we thus have

Example (cont'd)

$$
\begin{aligned}
P(\text{system failure}) &= 1 - P(\text{no system failure}) \qquad (175)\\
&= 1 - \prod_{i=1}^{250}(1 - p_i)\\
&= 1 - (1 - 0.003)^{250}\\
&= 0.528
\end{aligned}
$$

The following table lists some system failure probabilities for various component failure probabilities p_i.

p_i	0.003	0.002	0.001	0.0001	0.00001
P(system failure)	0.528	0.394	0.221	0.025	0.0025

This example demonstrates the usefulness of including spare parts (in parallel mode).

Total Independence

Remark: Clearly, total independence implies pairwise independence. The converse is not true in general, as we have seen.

Example

A very interesting example of using the multiplication rule in infinite-dimensional settings is given in

`https://www.cantorsparadise.com/a-simple-geometric-approach-to-the-zeta-function-ddce`

1.6.2 The Bernoulli scheme

We now consider Bernoulli experiments. These are experiments, in which there are only two mutually exclusive results that are of interest, represented by events A and $\overline{A}$, respectively. Let us perform n indepent Bernoulli experiments. The i-th experiment can then be described by the sub-algebra

$$\mathscr{E}_i = \{\emptyset, A, \overline{A}, \Omega\} \tag{176}$$

We now assume that the probability of the occurrence of A is constant throughout the series of n experiments, i.e.

$$P(A) = p \tag{177}$$
$$P(\overline{A}) = 1 - p \tag{178}$$

with some predefined $p > 0$.

Bernoulli scheme

The observed sequence of results after having performed the Bernoulli trials could then read as follows

$$(A, \overline{A}, \overline{A}, \overline{A}, A, A, \ldots, \overline{A}). \tag{179}$$

We call such an n- tuple a protocol of the experiment.

Example

Q1: What is the probability of a particular protocol with occurrence of A exactly in the trials $i_1, i_2, \ldots, i_k \in \{1, \ldots, n\}$?
The simple answer is

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k} \cap \overline{A}_{i_{k+1}} \cap \cdots \cap \overline{A}_{i_n}) = p^k \cdot (1-p)^{n-k} \tag{180}$$

where A_i stands for the occurrence of A in the i-th Bernoulli trial.

Often, we are not interested in the particular order of the results in the sequence of the trials, but only in the total number of the occurrences of A in n trials.

Example (cont'd)

Q2: What is the probability that we observe the event A exactly k times in n Bernoulli trials?

Let $H_n(A)$ denote the absolute frequency of the occurrence of A. Then it follows

$$\begin{aligned} P(H_n(A) = k) &= \sum_{(i_1,\ldots i_k)\in\{1,\ldots,n\}} P(A_{i_1} \cap \cdots \cap A_{i_k} \cap \overline{A}_{i_{k+1}} \cap \cdots \cap \overline{A}_{i_n}) \\ &= \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \end{aligned} \tag{181}$$

Note: Here the order of occurrence of successes and failures, respectively, does not matter. The probability only depends on $H_n(A)$.

Example (Lottery 6 out of 45)

Q1: For the lottery game "6 out of 45", what is the probability of the event A = "Guessing 4 numbers correctly twice in the course of 52 weekly lottery drawings"?

The weekly drawings form a Bernoulli scheme (since they are independent of each other) with $n = 52$ and

$$p = \frac{\binom{6}{4} \cdot \binom{39}{2}}{\binom{45}{6}} = 0.001365 \tag{182}$$

cp. eq. (115). From this it follows immediately that

$$\begin{aligned} P(A) &= \binom{52}{2} \cdot p^2 \cdot (1-p)^{50} \\ &= 0.0023 \end{aligned} \tag{183}$$

Example (cont'd)

Q2: For the lottery game "6 out of 45", what is the probability of the event B = "Guessing all the 6 numbers correctly once in the course of 20 years of 52 weekly lottery drawings"?

Again, we have a Bernoulli scheme with $n = 52 \cdot 20 = 1040$ and

$$p = \frac{\binom{6}{6} \cdot \binom{39}{0}}{\binom{45}{6}} = 0.000000123 \quad (184)$$

From this we get

$$P(B) = \binom{1040}{1} \cdot p^1 \cdot (1-p)^{1039} \quad (185)$$

$$= 0.000128 \quad (186)$$

Playing the lottery 50 years, the probability would increase to 0.00032.

1.6.3 Multinomial Scheme

We will now extend the notion of a Bernoulli (Binomial) scheme, which only distinguishes between the occurrence/non-occurrence of just one specific event, to situations where we have to take account of more than two alternative outcomes in each trial.

Let $\{A_1, \ldots, A_r\}$ be a **complete system** of events occurring with probabilities $p_i = P(A_i) > 0; i = 1, ..., r$. The completeness implies that

$$\sum_{i=1}^{r} p_i = p_1 + p_2 + \cdots + p_r = 1 \tag{187}$$

Let $k_i = H_n(A_i)$ denote the absolute frequency of occurrence of the event A_i in a total of n independent trials, $i = 1, ..., r$. Then we have

$$k_1 + k_2 + \cdots + k_r = n \tag{188}$$

Extension of Bernoulli scheme

Now, we are interested in the probability of the protocol

$$(\underbrace{A_1, ..., A_1}_{k_1}, \underbrace{A_2, ..., A_2}_{k_2}, ..., \underbrace{A_r, ..., A_r}_{k_r}) \tag{189}$$

where each A_i occurs exactly k_i times, $i = 1, ..., r$. Obviously, it holds

$$P\left(H_n(A_1) = k_1, H_n(A_2) = k_2, \ldots, H_n(A_r) = k_r\right) \tag{190}$$

$$= \frac{n!}{k_1! \cdot k_2! \ldots k_r!} \cdot \prod_{i=1}^{r} p_i^{k_i}$$

$$= n! \cdot \frac{p_1^{k_1} \cdot p_2^{k_2} \ldots p_r^{k_r}}{k_1! \cdot k_2! \ldots k_r!}$$

For $r = 2$ we arrive at the Bernoulli scheme again, with $k_1 = k, k_2 = n - k$ and $p_1 = p, p_2 = 1 - p$:

Multinomial scheme

$$
\begin{aligned}
n! \cdot \frac{p_1^{k_1} \cdot p_2^{k_2}}{k_1! \cdot k_2!} &= n! \cdot \frac{p^k \cdot (1-p)^{n-k}}{k! \cdot (n-k)!} \qquad (191)\\
&= \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k}\\
&= \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}
\end{aligned}
$$

Example (Rolling two dice)

We are rolling 2 dice $n = 10$ times. We consider the following three events:

$$
\begin{aligned}
A_1 &= \text{'Sum of points equals 11'} \qquad (192)\\
A_2 &= \text{'Sum of points equals 12'} \qquad (193)\\
A_3 &= \text{'Sum of points is } < 11\text{'} \qquad (194)
\end{aligned}
$$

Example (cont'd)

The probabilities of these events read

$$P(A_1) = \frac{2}{36}, P(A_2) = \frac{1}{36}, P(A_3) = \frac{33}{36} \tag{195}$$

Q: Wat is the probability that we have three times a sum of 11 points, two times a sum of 12 points and five times less than 11 points?

$$P\left(H_{10}(A_1) = 3, H_{10}(A_2) = 2, H_{10}(A_3) = 5\right) = \tag{196}$$

$$= 10! \cdot \frac{\left(\frac{2}{36}\right)^3 \cdot \left(\frac{1}{36}\right)^2 \cdot \left(\frac{33}{36}\right)^5}{3! \cdot 2! \cdot 5!}$$

$$= 0.0002157919$$

R-command: > dmultinom(x=c(3,2,5), prob=c(2/36,1/36, 33/36))

2. Random Variables and Probability Distributions

The notion of a random variable is of central importance in probability theory and statistics.

Goal: Describe the random mechanism generating the data by a function which assigns a probability mass to any given outcome of an experiment (data).

We say that the random variable is distributing the probability mass to the various outcomes of an experiment (collection of data).

We will distinguish between **discrete** and **continuous** random variables.

Example (Rolling a fair die)

This experiment can be described by the random variable

$$X \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix} \tag{197}$$

This is an example of a discrete random variable. In general, a discrete random variable can take only (countably) finitely many values.

Describe the random variable counting the sum of points which we get when we are rolling two fair dice!

In general, any discrete random variable can be described as a matrix consisting of two rows, the first row lists the values (outcomes) which can occur and the second row lists the corresponding probabilities of the outcomes:

Random Variables

$$X = \begin{pmatrix} x_1 & x_2 & \dots & \dots & x_{n-1} & x_n \\ p_1 & p_2 & \dots & \dots & p_{n-1} & p_n \end{pmatrix} \tag{198}$$

where

$$p_i = P(X = x_i) \; ; \; i = 1, ..., n. \tag{199}$$

Clearly, we must have

$$p_i > 0; i = 1, ..., n \quad \textit{and} \quad p_1 + p_2 + \cdots + p_n = 1 \tag{200}$$

To formalize things in more mathematical depth, we consider again our Kolmogorov Probability Space $[\Omega, \mathscr{E}, P]$.

We define a random variable as a mapping X

$$X : \Omega \to M \quad : \quad X(\omega) \in M, \;\; \forall \omega \in \Omega \tag{201}$$

into some (measurable) space $[M, \mathscr{M}]$.

The problem is then how to transfer the probability P from Ω to M?

Example (Rolling two dice)

We are rolling two dice and we are interested in the probability distribution of the random variable X = Total sum of points.
The following table lists all possible outcomes of this experiment

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

From this table we immediately get the frequencies of the different outcomes

Example (cont'd)

Sum of points	2	3	4	5	6	7	8	9	10	11	12
Frequencies	1	2	3	4	5	6	5	4	3	2	1

We thus have

$$M = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\} \tag{202}$$

and the original universe (sure event) reads

$$\Omega = \{(1,1), (1,2), (1,3), \ldots (6,6)\} \tag{203}$$

Finally, it is easy to find the probabilities associated with X:

Sum of points	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Random Variables

Coming back to the general notion of a random variable X as a mapping

$$X : \Omega \to M \quad : \quad X(\omega) \in M, \ \ \forall \omega \in \Omega \tag{204}$$

the following question arises

Q: Can we transfer the probability measure P from Ω to M and if so, how can we do it?

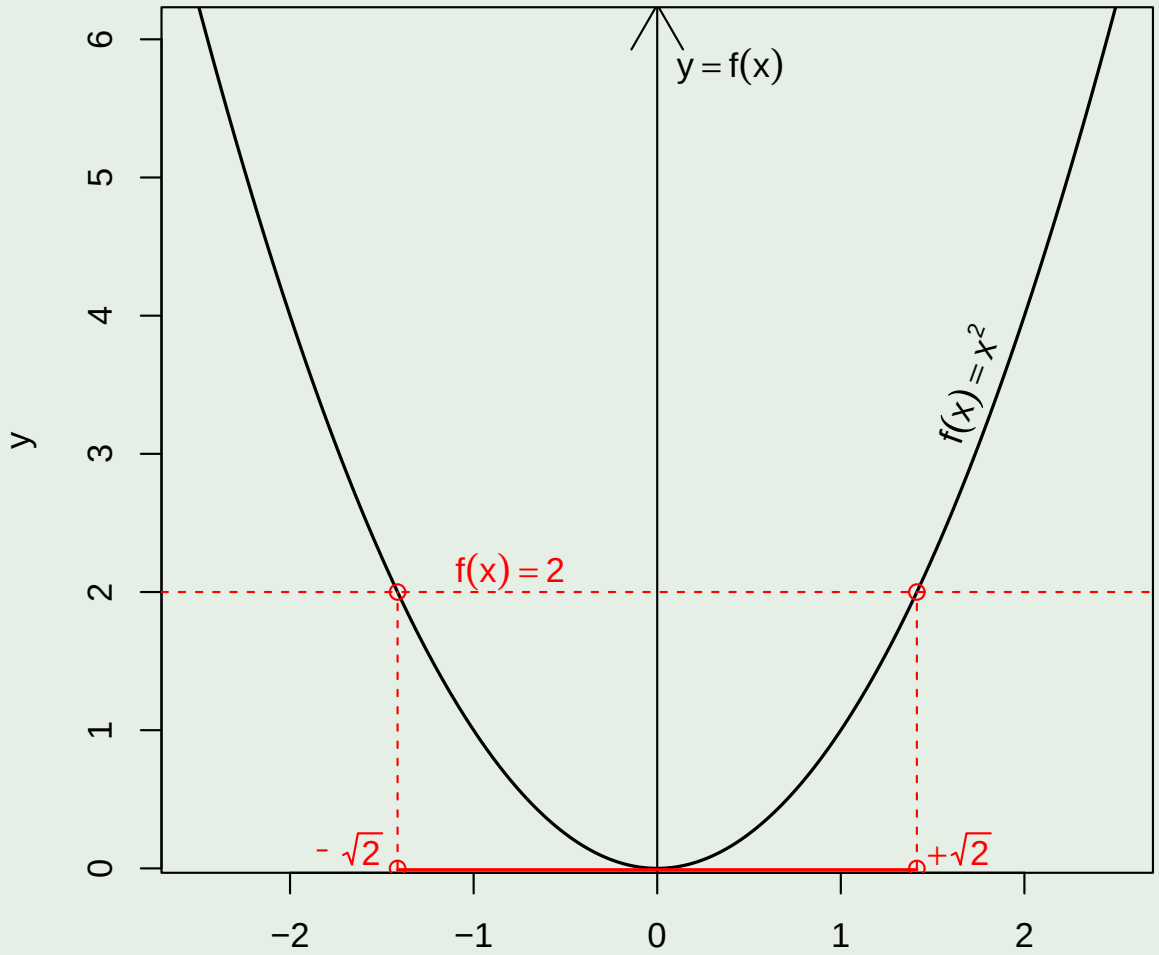
Clearly, the transfer can be made if the inverse mapping X^{-1} is well defined, i.e. the probability measure P can be transferred to M iff all inverse images are again contained in the σ-algebra $\mathscr{E}$.

The **inverse image** of $B \in \mathscr{M}$ defined as

$$X^{-1}(B) = \{\omega \in \Omega \ : \ X(\omega) \in B\} \tag{205}$$

Example (Inverse image)

Consider the function $y = f(x) = x^2, x \in \mathbb{R}$.



Inverse Image

Example (cont'd)

We want to find $f^{-1}(2)$ and $f^{-1}([0,2])$.

Clearly,

$$\begin{aligned} f^{-1}(2) &= \{x \in \mathbb{R} : f(x) = 2\} \\ &= \{x \in \mathbb{R} : x^2 = 2\} \\ &= \{-\sqrt{2}, +\sqrt{2}\} \end{aligned} \tag{206}$$

and

$$\begin{aligned} f^{-1}([0,2]) &= \{x \in \mathbb{R} : f(x) \in [0,2]\} \\ &= \{x \in \mathbb{R} : x^2 \in [0,2]\} \\ &= \left[-\sqrt{2}, +\sqrt{2}\right] \end{aligned} \tag{207}$$

Random Variables

Definition: The mapping $X : \Omega \to M$ is called a **random variable** of the measurable space $[M, \mathscr{M}]$, if it holds

$$X^{-1}(B) \in \mathscr{E} \quad \forall B \in \mathscr{M} \tag{208}$$

The set of all inverse images

$$\sigma(X) := \{X^{-1}(B) \ : \ B \in \mathscr{M}\} \tag{209}$$

is called the σ**-algebra induced by** X.

According to the above definition, X is a random variable of the measurable space $[M, \mathscr{M}]$ if and only if

$$\sigma(X) \subseteq \mathscr{E} \tag{210}$$

Random Variables

Now we are in the position to transfer the probability measure P to the image space:

$$[\Omega, \mathscr{E}, P] \to [M, \mathscr{M}, P_X] \tag{211}$$

Definition: (Probability Distribution)

$$\forall B \in \mathscr{M} \ : \ P_X(B) = P\left(X^{-1}(B)\right) \tag{212}$$

$$= P(\{\omega \in \Omega \ : \ X(\omega) \in B\}) \tag{213}$$

P_X is called the **induced probability measure** (induced by P) or **probability distribution of** X on $[M, \mathscr{M}]$.

Remark: In the sequel we will always have $M = \mathbb{R}$ and $\mathscr{M}$ the associated Borel Algebra of $\mathbb{R}$ (the set of all semiopen intervals of the form $(-\infty, x], x \in \mathbb{R}$).

Probability Distribution

Other important special cases are

- If $M = \mathbb{R}^n$ then X is called **random vector**.
- If M is a function space ($C[a, b]$, $L_2[a, b]$, ...), then X is called a
  - **stochastic process** for functions of one real-valued variable (e.g. time series)
  - **random field** for functions of $d \geq 2$ variables (e.g. spatial or spatio-temporal fields).
- If M is a set of closed subsets from $\mathbb{R}^n$, then X is called a **random closed set** (RACS). This is subject of the theory of stochastic geometry.

Example (Rolling two dice, cont'd)

Again, we are rolling two dice and we are interested in the probability distribution of the random variable X = Total sum of points.
The following table again lists all possible outcomes of this experiment, now in reverse order:

	1	2	3	4	5	6
6	7	8	9	10	11	12
5	6	7	8	9	10	11
4	5	6	7	8	9	10
3	4	5	6	7	8	9
2	3	4	5	6	7	8
1	2	3	4	5	6	7

The original event space is

Example (cont'd)

$$\Omega = \{(1,1),(1,2),(1,3),\ldots(6,6)\} \tag{214}$$

with cardinality $|\Omega| = 36$. For the image space of X we have

$$M = \{2,3,4,5,6,7,8,9,10,11,12\} \tag{215}$$

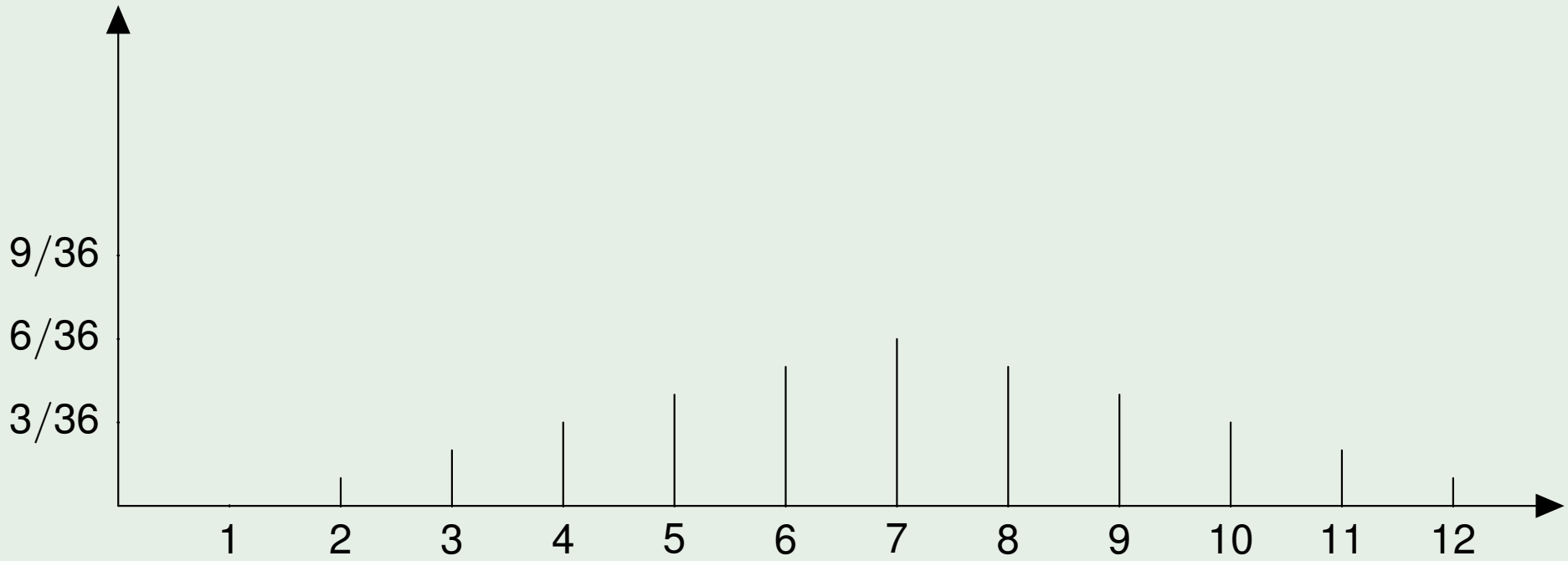
and the induced probability distribution on M is given as

$$X \sim \begin{pmatrix} 2 & 3 & \ldots & 7 & 8 & \ldots & 11 & 12 \\ \frac{1}{36} & \frac{2}{36} & \ldots & \frac{6}{36} & \frac{5}{36} & \ldots & \frac{2}{36} & \frac{1}{36} \end{pmatrix} \tag{216}$$

In a compact form, this can be written as

$$P(X = x) = \frac{6 - |7 - x|}{36} \cdot I_M(x);\ x \in \mathbb{R} \tag{217}$$

Example (cont'd)



2.2 Distribution Functions of Random Variables

We will now work with random variables as we do with functions.

Definition: The function

$$F_X(x) = P(X \leq x) \tag{218}$$

defined for all $x \in \mathbb{R}$ is called the **cumulative distribution function (cdf)** of the random variable X.

2.2.1 Properties of Distribution Functions

Theorem 4: Let be F_X the distribution function of a random variable X, then it holds:

(i) Boundedness: $0 \leq F_X(x) \leq 1 \quad \forall x \in \mathbb{R}$

(ii) Monotonicity: $x_1 \leq x_2 \quad \Rightarrow \quad F_X(x_1) \leq F_X(x_2)$

(iii) Normalization: $\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1$

(iv) Right- continuity: $\lim_{x \to x_0^+} F_X(x) = F(x_0) \quad \forall x_o \in \mathbb{R}$

Properties of cdf's

Proof: (i) This follows directly from the definition of $F_X(\cdot)$ as a probability, which is non-negative and normed by one (Kolmogorov's axioms I and II).

(ii) Observing that $(-\infty, x_1] \cap (x_1, x_2] = \emptyset$, we can decompose

$$(-\infty, x_1] \cup (x_1, x_2] \quad = \quad (-\infty, x_2] \tag{219}$$

which implies

$$\begin{aligned} F_X(x_2) &= P(X \leq x_2) = P\left(X \in (-\infty, x_1] \cup (x_1, x_2]\right) \\ &= P(X \leq x_1) + P\left(x_1 < X \leq x_2\right) \\ &= F_X(x_1) + \underbrace{P\left(x_1 < X \leq x_2\right)}_{\geq 0} \\ &\geq F_X(x_1) \end{aligned} \tag{220}$$

Properties of cdf's

(iii) Let be $\{x_i\}_{i=1,2,\ldots}$ a monotone increasing sequence such that

$$x_1 \leq x_2 \leq \cdots \leq x_n \to \infty \tag{221}$$

Now, defining sets $A_i = \{\omega \in \Omega \ : \ X(\omega) \leq x_i\}$, we immediately see that these sets form an increasing sequence as well,

$$A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq A_{n+1} \subseteq \ldots$$

from which, in turn, it follows that

$$\lim_{n\to\infty} P(X \leq x_n) = P\left(\bigcup_{n=1}^{\infty} A_n\right) = P(\Omega) = 1 = F_X(\infty).$$

Thus, we have proved that $\lim\limits_{x\to\infty} F_X(x) = 1$.

Analogously, we can show that $\lim\limits_{x\to-\infty} F_X(x) = 0$, using a monotone decreasing sequence

$$x_1 \geq x_2 \geq \cdots \geq x_n \to -\infty \tag{222}$$

Properties of cdf's

(iv) We finally prove that $F_X(\cdot)$ is right-continuous. Again, consider a sequence $\{x_i\}_{i=1,2,\ldots}$ with $x_1 \geq x_2 \geq \cdots \geq x_n \to x_0^+$. Then we have

$$\begin{aligned} \lim_{x \to x_0^+} F_X(x) &= P\left(\bigcap_{n=1}^{\infty} \{\omega \in \Omega \,:\, X(\omega) \leq x_n\}\right) \\ &= P(X \leq x_0) \\ &= F_X(x_0) \end{aligned} \tag{223}$$

Remark: The four properties listed in Theorem 4 fully characterize a (cumulative) distribution function $F_X(x)$. This means, conversely, that for any given function F, which is bounded from above and below ($F(\cdot) \in [0,1]$), monotone non-decreasing, continuous from the right and tends towards 0 and 1 when approaching the limits $\pm\infty$, respectively, there exists a random variable X such that $F_X = F$.

Properties of cdf's

We will now state an important theorem which tells us exactly how to compute probabilities with cdf's.

Theorem 5: Let X be a random variable and $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$ with $a < b$. Then it holds:

$$P_X((a,b]) = P(a < X \leq b) = F_X(b) - F_X(a) \tag{224}$$

and for each particular value b we have

$$P_X(\{b\}) = P(X = b) = F_X(b) - F_X(b - 0). \tag{225}$$

Proof: Eq. (224) follows from the fact that

$$\begin{aligned} F_X(b) &= P(X \leq b) \\ &= P(X \leq a) + P(a < X \leq b) \\ &= F_X(a) + P(a < X \leq b) \end{aligned} \tag{226}$$

The limiting case $a \to b - 0$ then yields eq. (225).

Note: If $F_X(x)$ is continuous at $x = b$, then eq. (225) simplifies to $P_X(\{b\}) = P(X = b) = 0$.

2.2.3 Types of distribution functions

We will distinguish between two types of random variables: **discrete** and **continuous** rv's, according to the associated image space.

Definition: Discrete random variable

A random variable X is said to be **discrete** if its image space is (countably) finite. Discrete rv's can be fully described by matrices consisting of two rows, the first row displays the values x_i taken by X and the second row the corresponding probabilities $p_i > 0$:

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n \dots \\ p_1 & p_2 & \dots & p_n \dots \end{pmatrix} \tag{227}$$

where $\sum_i \; p_i = 1$.

Discrete rv's

The probabilities p_i in (227) are associated with $F_X(\cdot)$ through

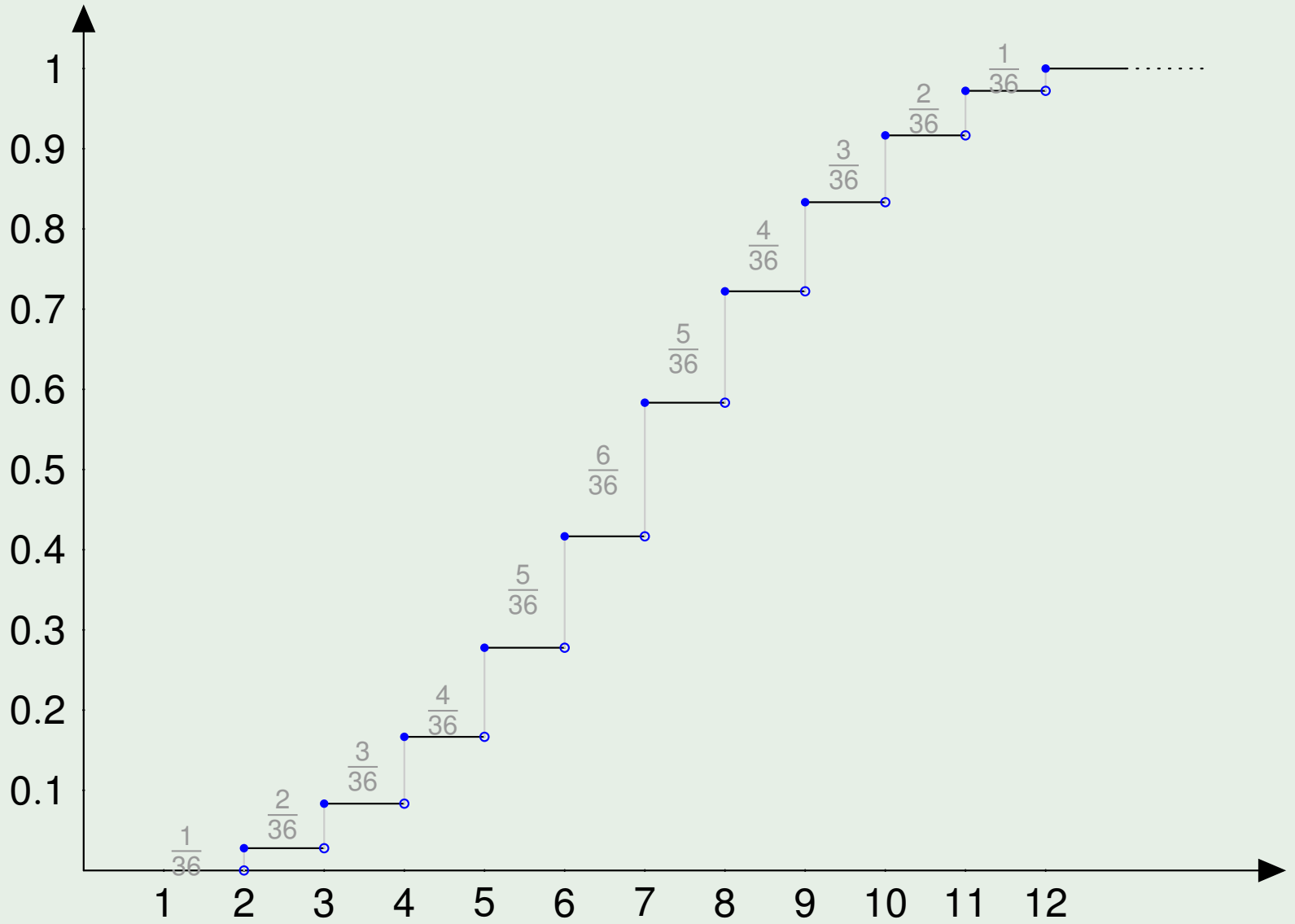
$$\begin{aligned} P(X = x_i) &= P(X \leq x_i) - P(X < x_i) \\ &= F_X(x_i) - F_X(x_i - 0) \\ &= p_i \end{aligned} \tag{228}$$

Therefore, the values x_i of X define the **discontinuities** of the distribution function F_X and the probablities p_i can be interpreted as "**jump heights**" of F_X.

Clearly, for all $x \notin \{x_1, x_2, \ldots, x_n, \ldots\}$ we have $P(X = x) = 0$.
Thus, there are only countably many outcomes of the discrete rv X with probability $\neq 0$.

Consequence: The distribution function F_X of a discrete rv X is a **step function** with jumps of height p_i at its realizations x_i; $i = 1, 2, ...$

Example (F_X for X = Number of points with two dice)



Implementation in R

Plotting a step function in *R* is done using the command **stepfun**. The syntax may be queried by typing

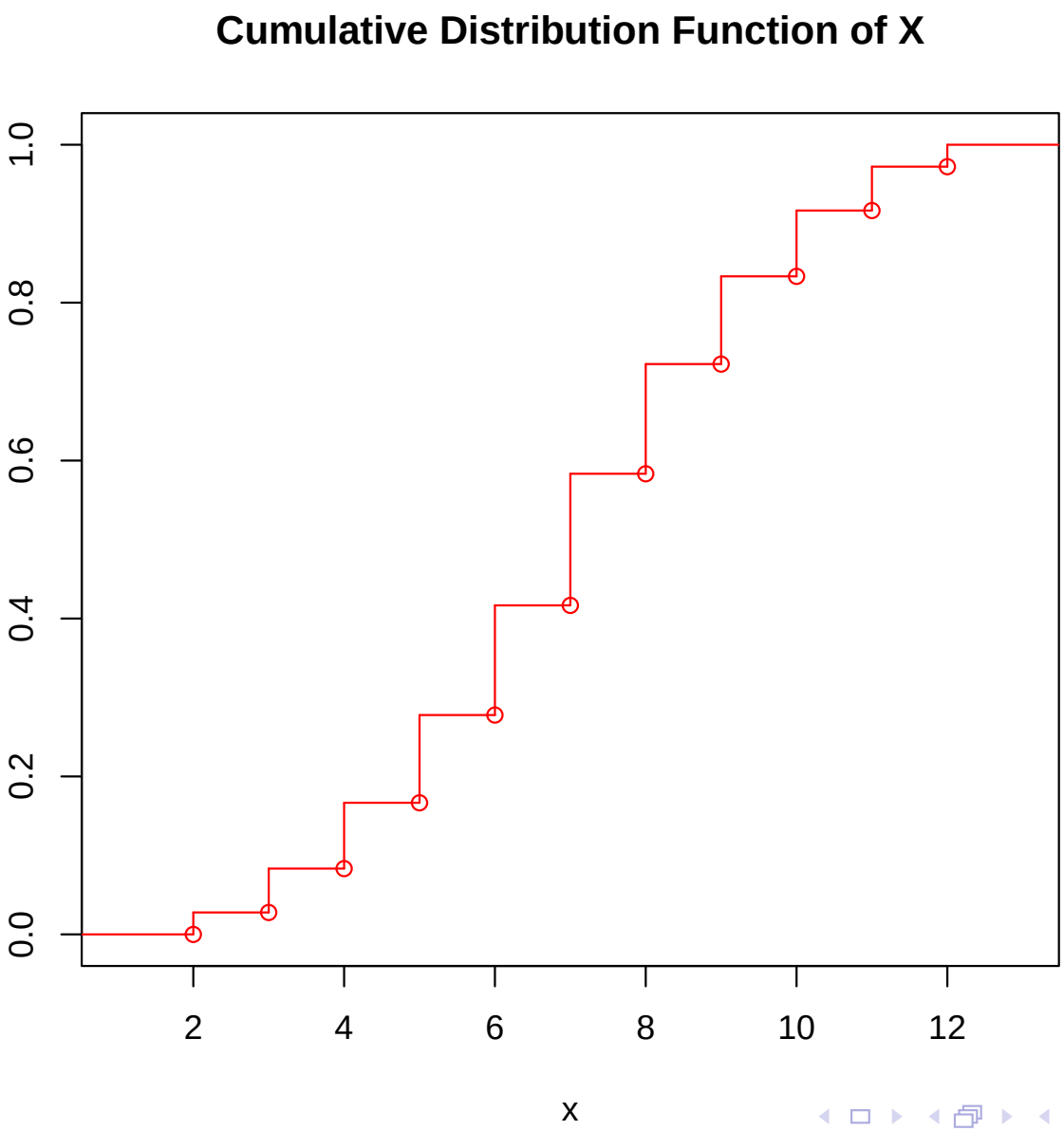
$>?stepfun$

Then we proceed as follows:

$$
\begin{aligned}
&> x = 1:13 \\
&> y = c(0,0,1:6,5,4,3,2,1,0)/36 \\
&> y = cumsum(y) \\
&> sf = stepfun(x, y, right = TRUE) \\
&> plot(sf, xval = 2:12, ylab = "", col = "red", main = \\
&+ "Cumulative\ Distribution\ Function\ of\ X")
\end{aligned}
$$

Remark: For statistical data analysis of a given data vector x, use the command **ecdf(x)** to plot the empirical cdf.

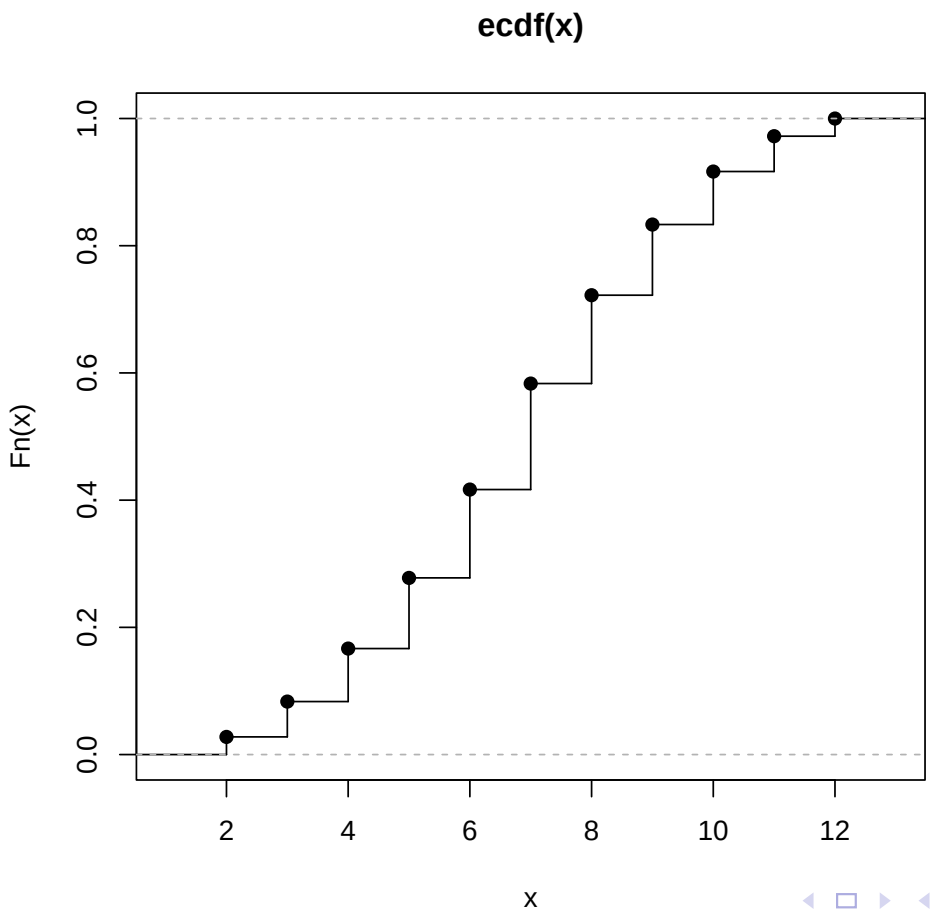
Implementation in R



Implementation in R

Alternatively, you can use the "ecdf" command as follows:

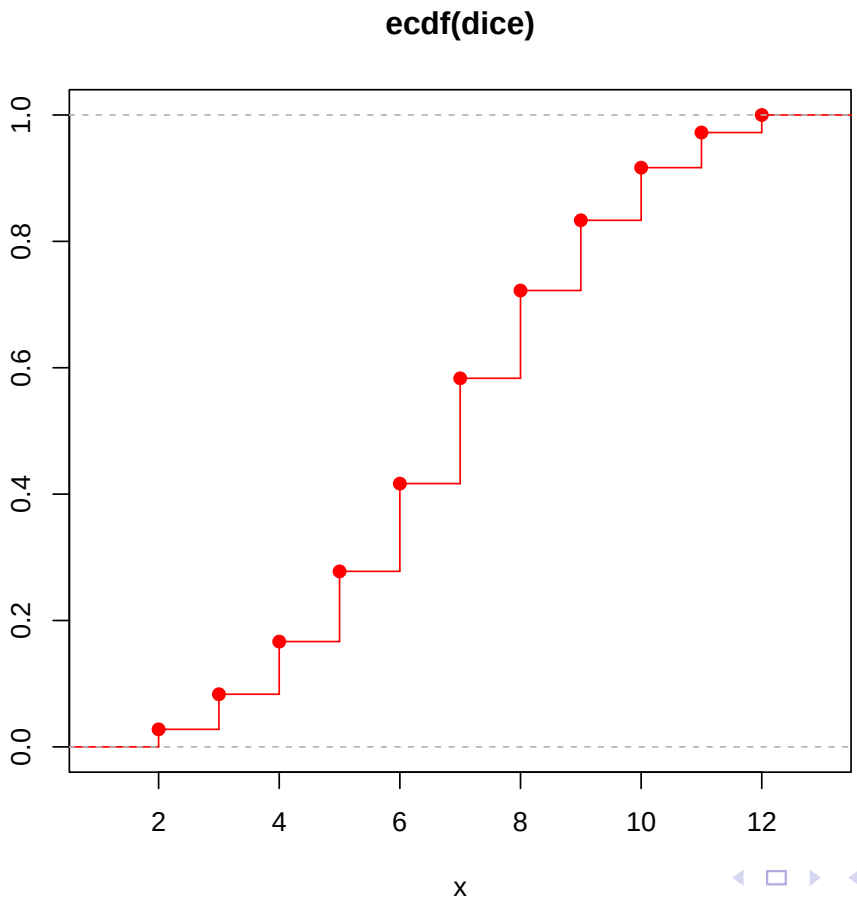
```
> x=c(2,3,3,4,4,4,5,5,5,5,6,6,6,6,6,7,7,7,7,7,7,
+ 8,8,8,8,8,9,9,9,9,10,10,10,10,11,11,12)
> plot.ecdf (x, verticals=T)
```



Implementation in R

Still better: Make use of the powerful R-functions **outer** and **table**

```
> dice= outer(1:6,1:6,FUN="+")
> table(dice)
> plot.ecdf(dice, ylab="", col="red", main="ecdf(dice)", verticals=T)
```



Computing probabilities using $F_X(x)$

Example (Rolling two dice)

Q1: What is the probability that the number of points is smaller than 10?

$$P(X < 10) = P(X \leq 9) = F_X(9)$$
$$= P(X = 2) + P(X = 3) + ... + P(X = 9)$$

Using R, we write

> x =2:12
> y= (6-abs(7-x))/36
> cumsum(y)[8]

i.e. $P(X \leq 9) = 0.8333$.

Q2: What is the probability that the number of points is at least 8?

$$P(X \geq 8) = 1 - P(X \leq 7) = 1 - (P(X = 2) + ... + P(X = 7))$$

In R we write: > 1 - cumsum(y)[6]
i.e. $P(X \geq 8) = 0.4166667$.

Probability Density Function

Definition: Continuous Random Variable

The random variable X is said to be **continuously distributed** if its image M is a continuous set (an interval) and there exists a nonnegative (Riemann-) integrable function $f(x)$ such that

$$F_X(x) = \int_{-\infty}^{x} f(t) \cdot dt \tag{229}$$

We then call f the **probability density function (pdf)** or, briefly, the **density** of X.

Remark: To stress the fact that f is the pdf of X, we often write f_X instead of just f. The pdf represents the continuous analogue to the probability mass function $f_X(x) = P(X = x)$ of a discrete rv X. We will deal more extensively with continuous rv's in Section 2.4.

2.3 Examples of discrete distributions

2.3.1 Binomial distribution

Definition: The random variable X with image space $\{0, 1, 2, \ldots, n\}$ is said to be **binomially distributed** with parameters n (a positive integer) and p $(0 < p < 1)$, if it holds

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots n \tag{230}$$

Briefly, we write

$$X \sim \text{Bi}(n, p) \tag{231}$$

Note: The binomial distribution describes experiments consisting of a series of n independent Bernoulli-trials, in which we observe either the success or failure as an outcome. In this context, p denotes the (constant) probability of success in each Bernoulli-trial. The integer k means the number of successes in the total of n trials.

Binomial Distribution

Corollary: The binomial distribution is normalized, since

$$\begin{aligned} \sum_{k=0}^{n} P(X=k) &= \sum_{k=0}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \\ &= [p+(1-p)]^n \\ &= 1 \end{aligned} \tag{232}$$

Important application areas of the binomial distribution refer to

- Bernoulli-Schemes (cp. eq. (181))
- Statistical Quality Control (SQC) **with** replacement

Binomial distribution in R: The probability mass function (pmf) and the cumulative distribution function of the binomial distribution Bi(n, p) can be accessed in R using

Binomial Distribution

- $P(X = k) =$ `dbinom(k,n,p)` computes the pmf
- $P(X \leq k) =$ `pbinom(k,n,p)` computes the cdf
- For simulation use `rbinom(m,n,p)`, which generates m realizations of $X \sim \text{Bi}(n, p)$

Example

Consider a multiple-choice-examination with 48 questions and 4 possible answers, each. A candidate tries to pass the exam by "simply guessing" the correct answer to each question. He/she then acts according to a Bernoulli scheme with parameters

$$n = 48 \tag{233}$$

$$p = \frac{1}{4} \tag{234}$$

Q1: What is the probability of answering at least half of the questions correctly?

Example (cont'd)

We need to compute the probability of having at least 24 successes. This is but

$$\begin{aligned} P(X \geq 24) &= 1 - P(X \leq 23) \\ &= 1 - \texttt{pbinom(23,48,0.25)} \\ &= 0.00017. \end{aligned} \tag{235}$$

Q2: What is the probability of answering at most one third of the questions correctly?

This probability is given by

$$\begin{aligned} P(X \leq 16) &= \texttt{pbinom(16,48,0.25)} \\ &= 0.9296. \end{aligned} \tag{236}$$

Example (cont'd)

Q3: How many questions need to be answered correctly in order to prevent someone from passing the exam by "pure guessing" with a probability of at most 0.01?

We need to find the smallest integer k such that

$$P(X \leq k) = \texttt{pbinom(k,48,0.25)} \geq 0.99 \tag{237}$$

Trying several integers, we find $k = 19$. However, there is a more concise method to determine this number:

```
qbinom(0.99, 48, 0.25)
```
(238)

leading to the same result, of course. We will get to know more about **quantiles** in a later section.

2.3.2 Hypergeometric distribution

Definition: Let be M, N and n positive integers with $M \leq N$. The random variable X with image space $\{0, 1, 2, \ldots, \min(n, M)\}$ is said to be **hypergeometrically distributed** with parameters M, N and n, if it holds

$$P(X = k) = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, \ldots, \min(n, M) \tag{239}$$

We briefly write this as

$$X \sim H(M, N, n) \tag{240}$$

Remark: This model is used to describe drawings from an urn proceeding **without replication** and finds application in SQC (sampling without replication) and wildlife/ecological monitoring programs (capture-recapture models).

Hypergeometric Distribution

In this model, the parameter M is of particular interest, meaning the number of failures in a production lot (number of ill persons in the total population, ill animals in a wildlife stock, endangered species in an ecological area, ...). This parameter is hard to determine when the total number N of objects is large, but it can be well estimated statistically. Further, also the fraction $\frac{M}{N}$ is of interest in practical applications (scrap rate, customer satisfaction/dissatisfaction rate, voting percentages, ...).

Hypergeometric distribution in R: The probability mass function and the cumulative distribution function of the hypergeometric distribution $\mathrm{H}(M, N, n)$ can be accessed in R using

- $P(X = k) =$ `dhyper(k,M,N-M,n)` computes the pmf
- $P(X \leq k) =$ `phyper(k,M,N-M,n)` computes the cdf
- For simulation, use `rhyper(100,M,N-M,n)` , which generates 100 realizations of $X \sim \mathrm{H}(M, N, n)$

Hypergeometric Distribution

Example (Lotto 6 out of 45)

Consider again the lottery game "Lotto 6 out of 45". Let the rv X be defined as "Number of correct guesses in a lottery draw". We are interested in the probability of having 4 correct numbers ("Vierer"). There are $N = 45$ numbers available on a lottery ticket, out of which $M = 6$ (correct ones) are picked in a weekly draw. On a ticket, $n = 6$ numbers can be sampled, out of which we want to have $k = 4$ correct ones. Thus,

$$\begin{aligned} P(X = 4) &= \frac{\binom{6}{4} \cdot \binom{45-6}{6-4}}{\binom{45}{6}} = \frac{\binom{6}{4} \cdot \binom{39}{2}}{\binom{45}{6}} \\ &= \texttt{dhyper(4,6,39,6)} \\ &= 0.001365. \end{aligned} \tag{241}$$

Clearly, when we focus on the symmetric "counter" event of having exactly 2 "wrong" numbers on the ticket, we arrive at the same result.

Hypergeometric Distribution

Example (cont'd)

Then, again, we have $N = 45$ available numbers on the ticket, out of which $M = 39$ are wrong. We sample $n = 6$ numbers out of which $k = 2$ are wrong (not picked in the lottery). This probability is given by `dhyper(2,39,6,6)` $= 0.001365$.

Relationship between hypergeometric and binomial distribution

In the very first trial (sample of size one) there is no difference between the two models. However, starting from the second trial, the lot (population) sizes under a hypergeometric sampling scheme successively reduce by one, since the sampled objects are not replaced. Therefore, under the hypergeometric sampling scheme, the probability of dawing a single object from the lot is **not constant**. On the contrary, for a binomial sampling scheme this probability remains constant, since any of the sampled objects is put back into the lot (sampling with replacement).

Approximation of the hypergeometric distribution

However, this difference becomes less essential in large populations, where $N >> n$. In this case we can well approximate the hypergeometric distribution by the binomial one.

Theorem 6: (Approximation of $\mathrm{H}(M, N, n)$ by $\mathrm{Bi}(n, p)$)

Let p be a given constant, $0 < p < 1$, then it holds

$$\lim_{\substack{N,M\to\infty \\ \frac{M}{N}=p}} \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}} = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \tag{242}$$

where $k = 0, 1, \ldots, \min(n, M)$.

Proof: We start with some preliminary facts. First,

$$\begin{aligned} \lim_{M\to\infty} \frac{\binom{M}{k}}{M^k} &= \lim_{M\to\infty} \frac{M \cdot (M-1) \ldots (M-k+1)}{k! \cdot M^k} \\ &= \lim_{M\to\infty} \frac{1}{k!} \cdot 1 \cdot \left(1 - \frac{1}{M}\right) \ldots \left(1 - \frac{k-1}{M}\right), \end{aligned} \tag{243}$$

Approximation of the hypergeometric distribution

which implies that

$$\lim_{M\to\infty} \frac{\binom{M}{k}}{M^k} = \frac{1}{k!} \tag{244}$$

Similarly, we have

$$\lim_{N\to\infty} \frac{\binom{N}{n}}{N^n} = \frac{1}{n!} \tag{245}$$

Additionally, we have

$$\lim_{\substack{N,M\to\infty \\ \frac{M}{N}=p}} \frac{\binom{N-M}{n-k}}{N^{n-k}} = \tag{246}$$

$$= \lim_{\substack{N,M\to\infty \\ \frac{M}{N}=p}} \frac{(N-M)(N-M-1)\ldots[(N-M)-(n-k)+1]}{(n-k)!N^{n-k}}$$

$$= \lim_{N,M\to\infty} \frac{1}{(n-k)!}\left(1-\frac{M}{N}\right)\left(1-\frac{M}{N}-\frac{1}{N}\right)\ldots\left(1-\frac{M}{N}-\frac{n-k-1}{N}\right)$$

Approximation of the hypergeometric distribution

which yields

$$\lim_{\substack{N,M\to\infty \\ \frac{M}{N}=p}} \frac{\binom{N-M}{n-k}}{N^{n-k}} = \frac{(1-p)^{n-k}}{(n-k)!} \tag{247}$$

With these preliminaries, we can proceed as follows

$$\lim_{\substack{N,M\to\infty \\ \frac{M}{N}=p}} \frac{\binom{M}{k}\cdot\binom{N-M}{n-k}}{\binom{N}{n}} = \lim_{\substack{N,M\to\infty \\ \frac{M}{N}=p}} \left(\frac{M}{N}\right)^k \cdot \frac{\binom{M}{k}}{M^k} \cdot \frac{N^n}{\binom{N}{n}} \cdot \frac{\binom{N-M}{n-k}}{N^{n-k}} \tag{248}$$

$$= \lim_{\substack{N,M\to\infty \\ \frac{M}{N}=p}} p^k \cdot \frac{\binom{M}{k}}{M^k} \cdot \frac{N^n}{\binom{N}{n}} \cdot \frac{\binom{N-M}{n-k}}{N^{n-k}}$$

It is well-known that the limit of a product equals the product of the limits (provided these exist). In our case, all the limits exist and we finally obtain

Approximation of the hypergeometric distribution

$$\lim_{\substack{N,M\to\infty \\ \frac{M}{N}=p}} p^k \cdot \frac{\binom{M}{k}}{M^k} \cdot \frac{N^n}{\binom{N}{n}} \cdot \frac{\binom{N-M}{n-k}}{N^{n-k}} = p^k \cdot \frac{1}{k!} \cdot n! \cdot \frac{(1-p)^{n-k}}{(n-k)!} \quad (249)$$

$$= \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k}$$

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

This proves our assertion.

Check this approximation through an example with large M and N! Assume, for example, $N = 10\hat{}6, M = 10\hat{}4, n = 200, k = 6$.

2.3.3 Poisson Distribution

This is an often used distribution for modeling occurrences of either rare and/or completely at random (CAR) events.

Definition: Poisson Distribution
The random variable X with image space (Wertebereich) $\{0, 1, \dots\}$ is said to have a **Poisson- Distribution** with parameter $\lambda > 0$, if it holds:

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \qquad k = 0, 1, 2, \dots \tag{250}$$

Briefly, we write

$$X \sim \text{Po}(\lambda) \tag{251}$$

Corollary: The Poisson distribution is normalized, since

$$\sum_{k=0}^{\infty} P(X = k) \quad = \quad \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \cdot e^{-\lambda} \tag{252}$$

$$= e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$
$$= e^{-\lambda} \cdot e^{\lambda} = 1$$

Example (Radioactive Decay)

Assume there are, initially, N atoms at time $t = 0$. Denoting by $P_k(t)$ the probability that exactly k of the atoms are decaying in the time interval $(0, t]$, then it is known from physics that

$$P_k(t) = \frac{(N \cdot \lambda \cdot t)^k}{k!} \cdot e^{-N \cdot \lambda \cdot t} \tag{253}$$

This, however, only holds for time intervals for which $t << \tau$, where τ stands for the radioactive half-life (Halbwertszeit). The decay rate (Zerfallskonstante) λ describes the physical properties of the radioactive material and is defined by $\lambda = \ln(2)/\tau$.

Poisson distribution

Further areas of application of the Poisson distribution:

- The probability of having exactly k objects (meteorite impacts, landslide occurrences,...) in an area of $|A|$ areal units is given by

$$P(X = k) = \frac{(\lambda \cdot |A|)^k}{k!} \cdot e^{-\lambda \cdot |A|} \tag{254}$$

  where λ stands for the average number of objects per areal unit.
- In the same vein, the number of misprints/errors in books, master and PhD theses etc. as well as the number of material faults per area or volume follows the Poisson distribution law.
- In sports, the number of goals scored in a game (soccer, ice hockey, baseball, ...) can be modeled by the Poisson distribution, where λ is typical for a single player or a whole team.

For more information on such applications see

`https://towardsdatascience.com/predicting-football-\match-result-using-poisson-distribution-ac72afbe36e0`

Poisson distribution

- The Poisson distribution is a basic model in epidemiology, e.g. for modeling the number of infected people in an area or the spatio-temporal spread of a disease (malaria, dengue, Covid,...)
- Finally, the Poisson distribution has lots of applications in queuing theory (Bedienungstheorie, Warteschlangentheorie), for example when modeling
- the food and beverage demands for restaurants, canteens (we have recently beat Facebook's "Prophet" tool)
- the number of customers at a cash point (in banks, supermarkets,...)
- the number of cars stopping at a crossroad or filling station
- the number of telephone calls per time
- the number of accesses to a website.

Poisson distribution

Poisson distribution in R

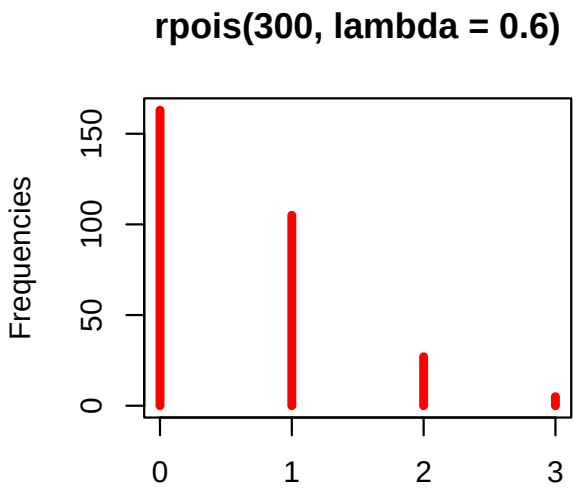
- $P(X = k) =$ `dpois(k,lambda)` computes the pmf
- $P(X \leq k) =$ `ppois(k,lambda)` computes the cdf
- For simulation `rpois(n, lambda)` generates n realizations of $X \sim \text{Po}(\lambda)$

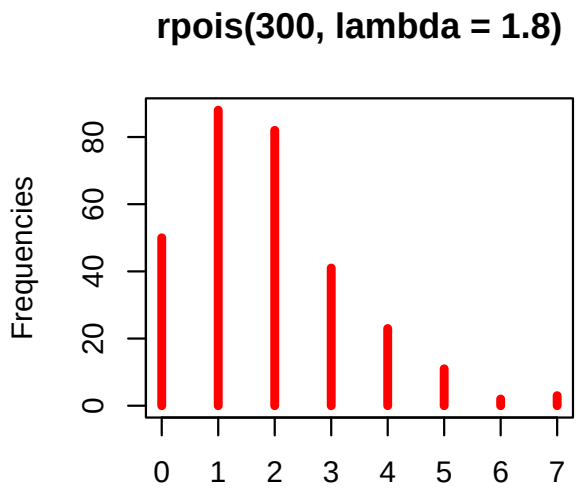
Example (Web access)

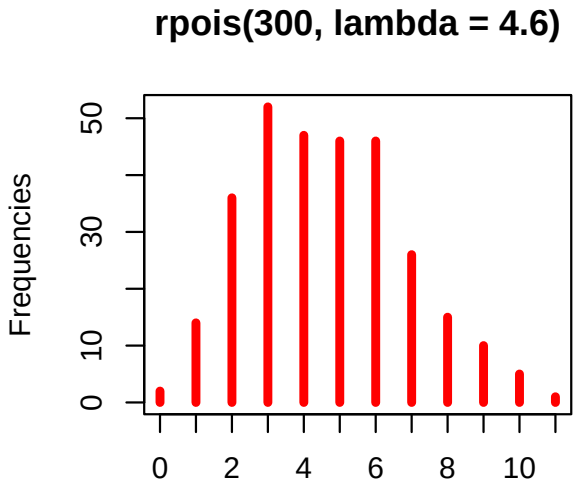
Suppose a website is accessed, on average, 40 times per hour. What is the probability that there will be more than 50 accesses at some hour?
We immediately have $\lambda = 40$. Thus, the probability becomes

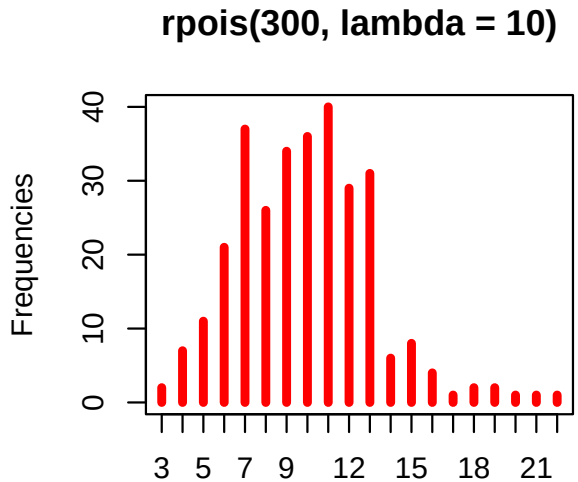
$$\begin{aligned} P(X > 50) &= 1 - P(X \leq 50) \\ &= 1 - \texttt{ppois(50,40)} \\ &= 0.0526 \end{aligned} \tag{255}$$

Poisson pmf's with different intensities λ









Approximation of the Binomial distribution

Interestingly, the binomial distribution can be approximated, under certain conditions, by the Poisson distribution. This is particulary helpful in cases of (very) large sample sizes. (Try to compute the outcomes of "choose(100000,80)" and "choose(100000,90)" in R!)

Theorem 7: (Poisson's limit theorem)

In the limiting situation, when $n \to \infty$, $p \to 0$ und $n \cdot p = \lambda$ with λ being constant, the binomial distribution approaches the Poisson distribution:

$$\lim_{\substack{n\to\infty \\ p\to 0 \\ n\cdot p=\lambda}} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \tag{256}$$

The proof proceeds on the basis of the famous **Stirling formula**:

$$n! \approx S(n) \text{ where } S(n) = \sqrt{2\pi n} \cdot n^n \cdot e^{-n} \tag{257}$$

More precisely, we have the following approximation bounds:

Approximation of the Binomial distribution

$$0 < ln(n!) - ln(S(n)) < \frac{1}{12n}$$

Check this for various choices of n, show that 100! has 158 digits!

The Poisson distribution is often called the "distribution of rare events", due to the limting approach $p \to 0$ in the above theorem. As a rule of thumb, the approximation of the binomial distribution by the Poisson distribution is sufficiently accurate when

$$n \geq 50, \quad p \leq 0.1, \quad n \cdot p \leq 9 \tag{258}$$

Check the conditions (258) for the examples on the Bernoulli experiments of

- opting for four correctly guessed numbers in a series of 52 weekly drawings in the lottery "6 out of 45"
- opting for a jackpot within a series of playing the lottery "6 out of 45" weekly over a time span of 20 years.

2.3.4 Geometric Distribution

Consider a series of (possibly infinite) Bernoulli experiments with (success) parameter $p > 0$. What is the chance that we have success already in the 1st, 2nd, 3rd, ... trial? Put another way, we are interested in the chances of having a small waiting time until the first success. The following simple distribution gives us the answer.

Definition: Geometric distribution
The rv X with image space $\{0, 1, 2, \dots\}$ is said to follow a **geometric distribution** with parameter $0 \leq p \leq 1$, if it holds

$$P(X = k) = (1 - p)^k \cdot p, \quad k = 0, 1, 2, \dots \tag{259}$$

Briefly, we write

$$X \sim \text{Geom}(p) \tag{260}$$

and interpret X as the number of failures before the first success.

Geometric Distribution

Corollary: The geometric distribution is normalized, since

$$\begin{aligned} \sum_{k=0}^{\infty} P(X=k) &= p \cdot \sum_{k=0}^{\infty} (1-p)^k \\ &= p \cdot \sum_{k=0}^{\infty} q^k, \end{aligned} \tag{261}$$

where $q = 1 - p$. Observing that the infinite sum over the powers of $q \in (0,1)$ represents a geometric series with (limiting) value $\frac{1}{1-q} = \frac{1}{p}$, the assertion follows.

Example (Rolling a die)

Suppose you and three other persons want to play ludo (Mensch ärgere dich nicht!), where he/she opens the game who first rolls "six". Suppose you would like to be "noble" and suggest to be the last in the row to roll the die.
What is the probability that you open the game?

Example

Clearly, rolling the die until the first six is produced is a Bernoulli experiment, where, $p = 1/6$ and n, at least in principle, need not to be bounded. Let X denote the number of failures before the first "six" occurs; each failure occurs with probability $q = 1 - p = 5/6$. The problem then is to find the probability that $X \in \{3, 7, 11, 15, \dots\}$:

$$\begin{aligned} \sum_{k=1}^{\infty} P(X = 4k-1) &= q^3 p + q^7 p + q^{11} p + \dots \qquad (262) \\ &= q^3 p (1 + q^4 + q^8 + \dots) \\ &= q^3 p / (1 - q^4) \ \ (\textit{geom. series with } a = q^4) \end{aligned}$$

Plugging in the values $p = \frac{1}{6}, q = \frac{5}{6}$, we finally obtain the solution

$$\sum_{}^{\infty} P(X = 4k-1) = 0.18629. \qquad (263)$$

Geometric Distribution

Geometric distribution in R

- $P(X = k) =$ `dgeom(k, p)` computes the pmf of having exactly k failures before the first success
- $P(X \leq k) =$ `pgeom(k, p)` computes the cdf (having at most k failures before the first success)
- For simulation `rgeom(m, p)` generates m realizations of $X \sim \text{Geom}(p)$

Example (Rolling the first "six" with a die)

Coming back to the problem of opening a ludo game with three other persons, using R we have

- The probability of opening the game as the last one in the queue for rolling a "six" reads

```
> k=1:1000; p=1/6
> sum(dgeom(4*k-1, p))
 [1] 0.1862891
```

Geometric Distribution

Example (cont'd)

This result, of course, coincides with eq. (263).

For the other three players we get

- ```
  > sum(dgeom(4*k-2, p))
  [1] 0.2235469
  ```
- ```
  > sum(dgeom(4*k-3,p))
  [1] 0.2682563
  ```
- ```
  > sum(dgeom(4*k-4,p))
  [1] 0.3219076
  ```

Thus, the first player has the biggest chance to open the game, while the fourth player has the least chance.
  ```

2.3.5 Negative binomial distribution

This distribution generalizes the geometric one. Let T_r denote the (random) number of trials until the r-th success in Bernoulli experiments, where $r \in \{1, 2, ...\}$ is some positive integer. To illustrate this, for the following sequence of results, with 1 = sucess, 0= failure,

$$0001000000100100000010000\ldots \tag{264}$$

$$T_1 = 4,\ T_2 = 11,\ T_3 = 14,\ T_4 = 21,\ T_5 = ?$$

Q: What is the distribution of T_r?

Clearly, the possible values of T_r are $r, r+1, r+2, \ldots$. For t in this range

$$\begin{aligned} P(T_r = t) &= \binom{t-1}{r-1} p^{r-1}(1-p)^{t-r} * p \qquad (265) \\ &= P(r-1 \textit{ succ. in first } t-1 \textit{ trials}) * P(\textit{succ. in trial } t) \end{aligned}$$

Negative Binomial

We immediately notice that

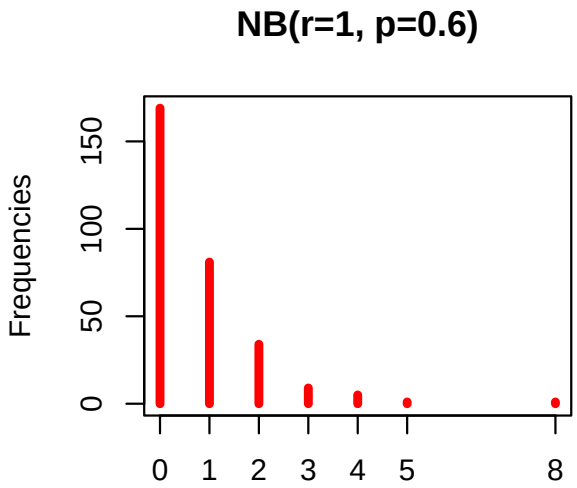
$$T_r \quad = \quad W_1 + W_2 + \cdots + W_r \tag{266}$$

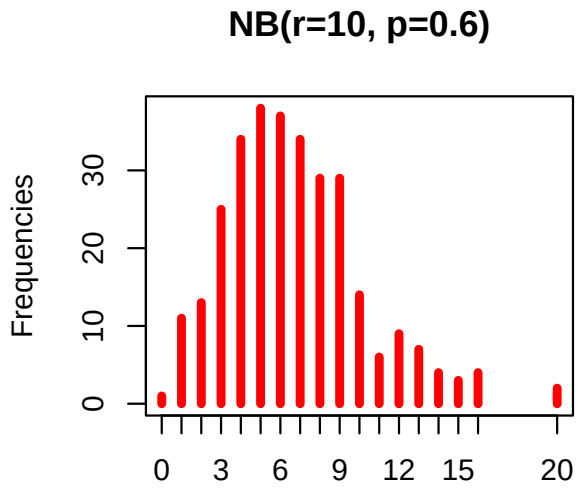
where W_i is the waiting time after the $(i-1)$ th success till the i th success; here $W_1 = 3, W_2 = 7, W_3 = 3, W_4 = 7$,...
It is intuitively clear that $W_1, W_2, W_3, \ldots$ are iid $\text{Geom}(p)$ distributed.
The distribution of $X = T_r - r$, the number of failures before the r th success, in independent Bernoulli trials with success probability p, is called negative binomial with parameters r and p. This is just the distribution of T_r, shifted from $\{r, r+1, r+2, \ldots\}$ to $\{0, 1, 2, \ldots\}$.

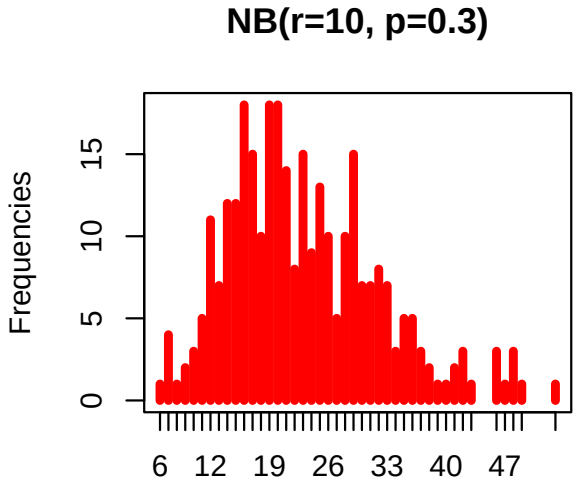
Definition: The rv X with image space $\{0, 1, 2, \ldots\}$ is said to follow a **negative binomial distribution** with parameters r and p, if it holds

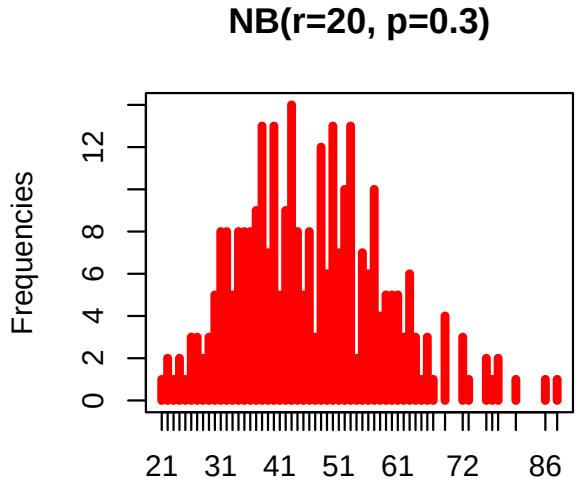
$$P(X = k) \quad = \quad \binom{k+r-1}{r-1} \cdot p^r (1-p)^k, \quad k = 0, 1, 2, \ldots \tag{267}$$

NegBinomials with different r and p









Negative Binomial

We will briefly write

$$X \sim \text{NB}(r, p) \tag{268}$$

and interpret X as the number of failures before the r th success.

Negative binomial distribution in R

- $P(X = k) =$ `dnbinom(k, r, p)` computes the pmf of having exactly k failures before the r th success
- $P(X \leq k) =$ `pnbinom(k, r, p)` computes the cdf (having at most k failures before the r th success)
- `qnbinom(prob, r, p)` computes the quantile k for given probability prob (do not mix it up with p)
- For simulation `rnbinom(m, r, p)` generates m realizations of $X \sim \text{NB}(r, p)$

Negative Binomial

Example (Playing Roulette)

Suppose you are playing the roulette game in a casino. You decide to place your bets exclusively on "Columns" (e.g. Column 34: 1,4,7,10,...,34) and to stop playing after having reached 10 wins.

Q: What is the probability that you have to play no more than 30 games in order to achieve your goal of ten wins?

Here we have $p = 12/37$, since there are 37 numbers in total $(0, 1, 2, \ldots, 36)$. We have to find $P(X \leq 20)$, where $X \sim \mathrm{NB}(r, p)$ with $r = 10$. We get

$$P(X \leq 20) \quad = \quad \texttt{pnbinom(20, 10, 12/37)=0.5264.} \tag{269}$$

To achieve your goal with at least 80% probability, you need at least 37 games, since

`qnbinom(0.8,10,p)` $= 27$.

Negative Binomial

Note: The negative binomial distribution finds application in modeling recurring events in ecology, epidemiology and health insurance (floods, landslides, wildfires, infectious disease waves, bone fractures, hospital stays,...).
For the special case of $r = 1$, eq. (267) reduces to

$$P(X = k) \quad = \quad p(1-p)^k, \quad k = 0, 1, 2, \ldots \tag{270}$$

i.e. the negative binomial distribution then reduces to the geometric distribution: $\mathrm{NB}(1, p) = \mathrm{Geom}(p)$.

2.4 Continuous Distributions

We will now extend the notion of discrete rv's, whose distribution functions are described by **step functions**, to continuous distributions.

Definition: The random variable X is said to be **continuously distributed** if its image space is a continuous set (interval) on the real line and there exists a non-negative, integrable function $f(x)$ such that

Continuous distributions

$$F_X(x) = \int_{-\infty}^{x} f(t)dt \tag{271}$$

The function $f(\cdot)$ is called **probability density function (pdf)** or simply **density function** of X.

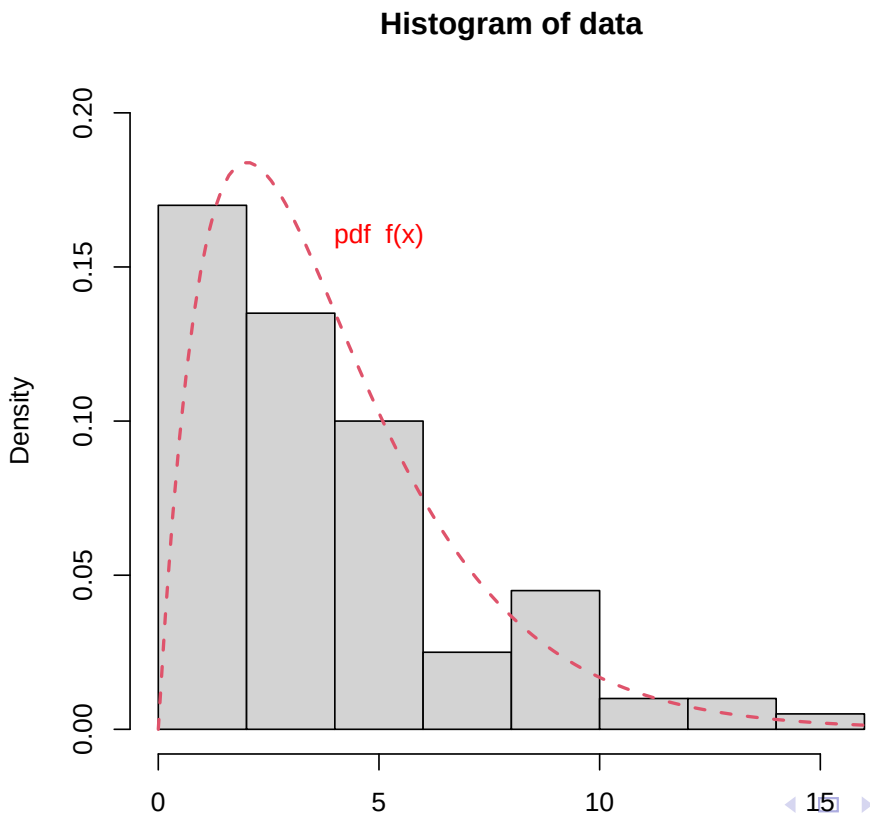
To indicate that the pdf in fact generates the distribution function $F_X(x)$, we will henceforth write $f_X(x)$ instead of just $f(x)$.
From the properties of F_X given in Theorem 4 of Section 2.2.1 it follows immediately that f_X must satisfy the two conditions

$$f_X(x) \geq 0 \;\; \forall x \in \mathbb{R} \quad \textit{and} \quad \int_{-\infty}^{+\infty} f_X(x)\, dx = 1 \tag{272}$$

The concept of a cont. distributed rv is an idealization which allows probabilities to be calculated by calculus (real analysis). Moreover, they often occur as limits for discrete models (see Ch. 4).

Continuous distributions

The empirical distribution of a data list $(x_1, x_2, \ldots, x_n)$ can be displayed in a histogram, which smoothes out the data to display the general shape of the empirical distribution. The histogram approximates the density $f_X(x)$ of the rv generating the data, and can be estimated from the data. Try `?density` in R!



Continuous distributions

The basic idea is that probabilities are defined by areas under the graph of $f_X(x)$, i.e. for all $a \leq b$

$$\begin{aligned} P(a < X \leq b) &= \int_a^b f_X(x)\, dx \\ &= F_X(b) - F_X(a). \end{aligned} \tag{273}$$

This implies that $F_X(\cdot)$ is the integrand (antiderivative, Stammfunktion) of the pdf $f_X(\cdot)$, which, in turn, is the continuous analogue to the pmf $P(X = x)$ of a discrete rv. However, for a continuous rv X we have

$$P(X = x) = 0 \quad \forall x \in \mathbb{R} \tag{274}$$

and, therefore, $P(a < X \leq b) = P(a \leq X \leq b) = P(a < X < b)$ are all the same.

2.4.1 The uniform distribution

Definition: The rv X is said to have a **uniform distribution** on the interval (a, b), if it has pdf

$$\begin{aligned} f_X(x) &= \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{1}{b-a} \cdot I_{(a,b)}(x) \end{aligned} \tag{275}$$

This in fact defines a pdf, since it satisfies the conditions given in (272), noting that the area under the graph of f_X is one. Further, we have

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases} \tag{276}$$

Uniform distribution

We will briefly write

$$X \sim \mathrm{U}(a, b) \tag{277}$$

For a uniform distribution, probabilities reduce to relative lengths: if $X \sim \mathrm{U}(a, b)$, then for $a < x_1 < x_2 < b$,

$$P(x_1 < X < x_2) = \frac{x_2 - x_1}{b - a} \tag{278}$$

For example, if X has uniform (0,2) distribution, then the probability that X is 1.2 correct to one decimal place is

$$P(1.15 < X < 1.25) = \frac{1.25-1.15}{2-0} = \frac{0.1}{2} = 0.05.$$

A simple rescaling transforms the interval (a, b) into $(0, 1)$. The density of a $\mathrm{U}(0, 1)$ distribution is simply 1 on $(0, 1)$, and 0 elsewhere:

Uniform distribution

$$X \sim \mathrm{U}(a,b) \quad \Longrightarrow \quad Y = (X-a)/(b-a) \sim \mathrm{U}(0,1) \tag{279}$$

and then any problem solving for X can be done on the basis of the simpler $\mathrm{U}(0,1)$ distribution:

$$Y \sim \mathrm{U}(0,1) \quad \Longrightarrow \quad X = a + (b-a)Y \sim \mathrm{U}(a,b) \tag{280}$$

Even more so, we will later see that through a relatively simple nonlinear transformation we can generate **any** cdf F_X on the basis of the uniform distribution $\mathrm{U}(0,1)$. For example, random number generation often proceeds on the basis of random numbers which are uniformly distributed as $\mathrm{U}(0,1)$.

Uniform distribution in R

- `dunif(x,a,b)` computes the pdf $f_X(x)$ of $X \sim \mathrm{U}(a,b)$
- `punif(x,a,b)` computes the cdf $F_X(x)$ of $X \sim \mathrm{U}(a,b)$

Uniform distribution

- `qunif(prob,a,b)` computes the quantile x of $X \sim \mathrm{U}(a, b)$ for a given probability prob
- For simulation `runif(m,a,b)` generates m realizations of $X \sim \mathrm{U}(a, b)$

Example

> ??punif

-
```
> punif(3,0,1)
[1] 1
```
-
```
> punif(3,1,4)
[1] 0.6666667
```
-
```
> dunif(3,1,4)
[1] 0.3333333
```
-
```
> dunif(6,1,4)
[1] 0
```

Example

- `> qunif(0.6666667,1,4)`
  `[1] 3`
- `> runif(5)`
  `[1] 0.2964735 0.2327178 0.3262376 0.9811382 0.1207741`

2.4.2 The normal distribution

Definition: The random variable X is said to have a **normal distribution** with parameters $\mu \in \mathbb{R}$ and σ^2, $\sigma > 0$, if it has pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot exp\left(-\frac{1}{2}(x-\mu)^2/\sigma^2\right), \quad -\infty < x < +\infty \quad (281)$$

We will briefly write

$$X \sim \mathrm{N}(\mu, \sigma^2) \quad (282)$$

Normal distribution

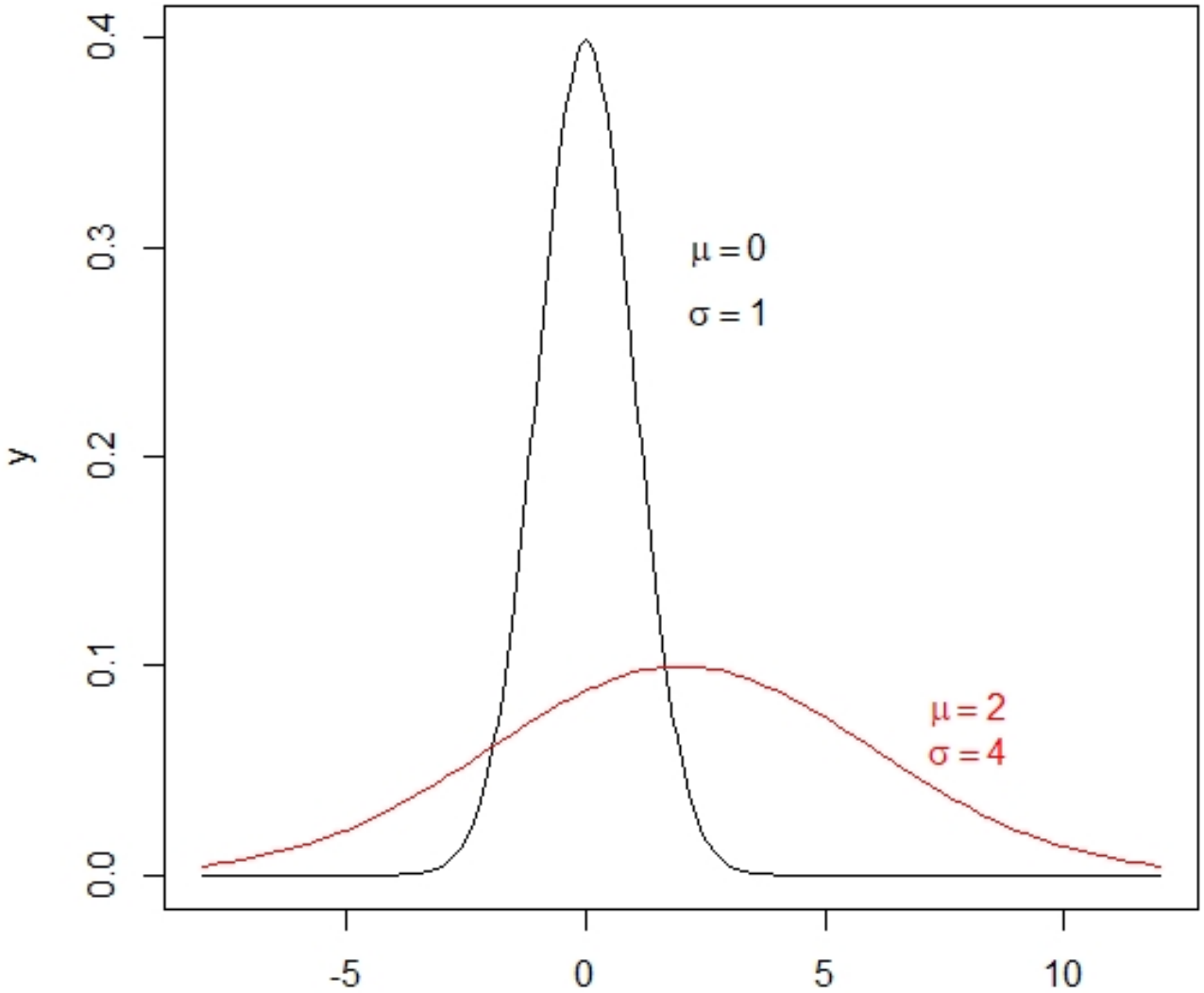
Historically, it was first discovered by Abraham DeMoivre (1667-1754), as the approximation of the binomial distribution $\mathrm{Bi}(n, p)$ for large n.
The normal distribution is also known as the **Gaussian** distribution.
Gauss (1777-1855) and Laplace (1749-1827) brought out the central role of the normal distribution in the theory of errors of observation.
Quetelet (1796-1874) and Galton (1822-1911) fitted the normal distribution to empirical data such as heights and weights in human and animal populations.
The normal distribution $\mathrm{N}(\mu, \sigma^2)$ can always be standardized such that $\mu = 0$ and $\sigma = 1$:

$$X \sim \mathrm{N}(\mu, \sigma^2) \quad \Longrightarrow \quad Y = (X - \mu)/\sigma \sim \mathrm{N}(0, 1) \tag{283}$$

and then any problem solving for X can be done on the basis of the simpler $\mathrm{N}(0, 1)$ distribution:

$$Y \sim \mathrm{N}(0, 1) \quad \Longrightarrow \quad X = \mu + \sigma Y \sim \mathrm{N}(\mu, \sigma^2) \tag{284}$$

Normal density functions



Normal distribution

The normal distribution piles up around μ for small values of σ^2, and becomes more and more spread out as σ^2 increases.

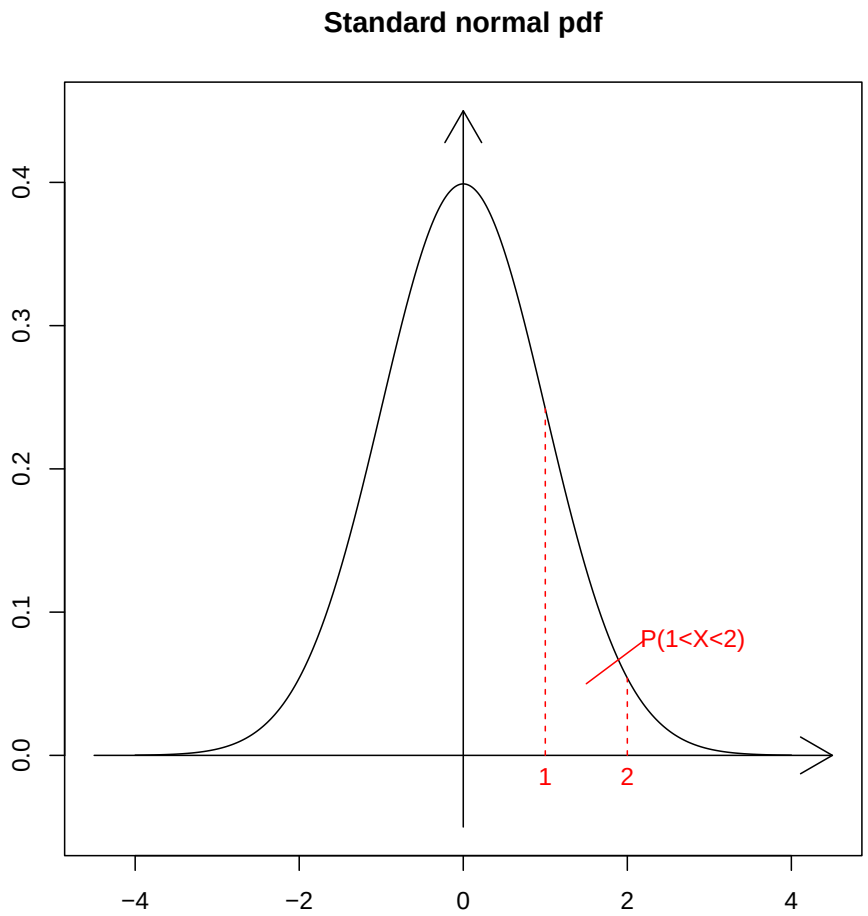
Note: There is no analytical formula for the distribution function F_X corresponding to the pdf f_X of $\mathrm{N}(\mu, \sigma^2)$ given in eq. (281),

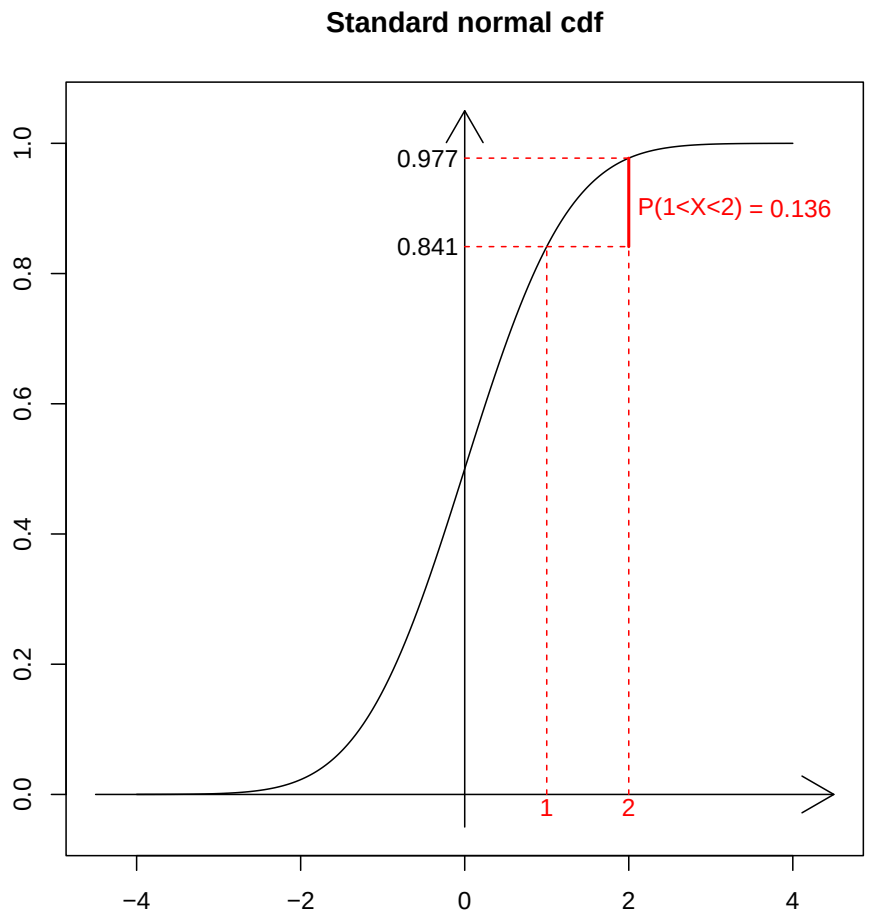
$$F_X(b) = \int_{-\infty}^{b} f_X(x)\, dx. \tag{285}$$

Therefore, all the probabilities $P(a \leq x \leq b) = F_X(b) - F_X(a)$ must be computed numerically. However, one can show that the normalization condition for the (standard) normal density is satisfied:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma}\, e^{-x^2/2}\, dx = 1. \tag{286}$$

Normal pdf and cdf





Normal distribution

In scientific and engineering applications the following σ **rules** for normally distributed rv's $X \sim \mathrm{N}(\mu, \sigma^2)$ are widely used:

- **One σ rule:** $P(-\sigma < X - \mu < +\sigma) = 0.6827$,
  i.e. about 68.3% of the data are within the 1σ range
- **Two σ rule:** $P(-2\sigma < X - \mu < +2\sigma) = 0.9545$,
  i.e. about 95.4% of the data are within the 2σ range
- **Three σ rule:** $P(-3\sigma < X - \mu < +3\sigma) = 0.9973$,
  i.e. about 99.7% of the data are within the 3σ range.

From the "2σ rule", the common statistical principle of determining 95% **confidence intervals** has been derived.

Normal distribution in R

Normal distribution in R:

All the necessary probabilities involving the normal distribution can be easily computed in R as follows

- `dnorm(x, mean= 3, sd= 2.4)` computes the pdf $f_X(x)$ of $X \sim \mathrm{N}(\mu = 3, \sigma^2 = 2.4^2)$
- `pnorm(x, mean= 0, sd= 1)` computes the cdf $F_X(x)$ of $X \sim \mathrm{N}(0, 1)$
- `qnorm(prob, mean= 5, sd= 0.4)` computes the quantile x of $X \sim \mathrm{N}(\mu = 5, \sigma^2 = 0.4^2)$ for a given probability prob
- For simulation `rnorm(k, mean= 5, sd= 0.4)` generates k realizations of $X \sim \mathrm{N}(\mu = 5, \sigma^2 = 0.4^2)$

Use this to check the "σ **rules**"!

Normal distribution in R

The basis for most statistical applications of the normal distribution is the **central limit theorem**, which states that the distribution of the sum (or average) of a large number of independent measurements will typically tend to follow the normal curve, even if the distribution of the individual measurements does not.

Example (Repeated measurements)

Suppose a long series of repeated measurements of the weight of (Lindt) chocolates yield results that are normally distributed with a mean of 100 g and a standard deviation of 1.5 g.

Q1: About what proportion of measurements are within the tolerance limit of 3 grams?

This is

$$P(97 \leq X \leq 103) = pnorm(103, 100, 1.5) - pnorm(97, 100, 1.5)$$
$$= 0.9544, \quad i.e.\ 95.44\%.$$

Normal distribution in R

Example (cont'd)

Q2: In 100 measurements, what is the probability that more than 10 measurements will fall outside the tolerance interval $(97g, 103g)$?

It seems reasonable to assume that each measurement is correct to within 3 grams with probability 0.9544, independently of all others. Out of 100 measurements, the number correct to within 3 grams has the binomial distribution $\mathrm{Bi}(100, 0.9544)$. Thus, the probability is

$$pbinom(89, 100, 0.9544) = 0.0059745, \ i.e.\ 0.60\%.$$

Q3: To which level should the standard deviation be reduced to guarantee that at least 98% of all the chocolates fall within the tolerance interval $(97g, 103g)$?

We have to make sure that the standard deviation sd is such that

$$pnorm(103, 100, sd) - pnorm(97, 100, sd) \geq 0.98.$$

This is, however, guaranteed as soon as $sd \leq 1.2895$:

Normal distribution in R

Example (cont'd)

We check this using the function `uniroot` for solving nonlinear equations in R.

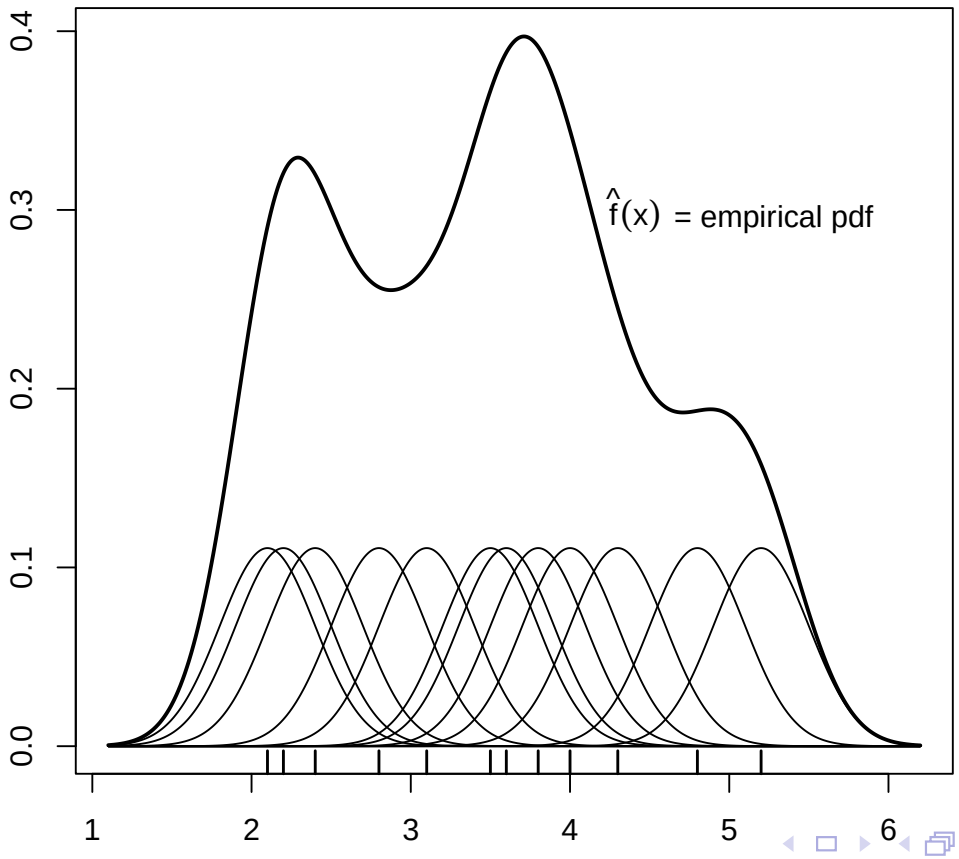
```
> uniroot(function(x) pnorm(103,100,x) -
pnorm(97,100,x) - 0.98001,
+ interval=c(1.0,1.5), tol=0.0001)
$root
 [1] 1.289472
$f.root
 [1] -8.236778e-08
$iter
 [1] 6
$estim.prec
 [1] 5e-05
```

Normal distribution in R

Note: The empirical pdf $\hat{f}_X(x)$ of the data $(x_1, \ldots, x_n)$ can be estimated by a superposition of Gaussian "kernel" densities

$$K_n(x) = \frac{1}{n} \cdot \frac{1}{\sqrt{2\pi} \cdot h} \cdot exp(-(x - x_i)^2/2h^2); \; i = 1, \ldots, n$$

data=c(2.1,2.2,2.4,2.8,3.1,3.5,3.6,3.8,4.0,4.3,4.8,5.2)



Normal distribution in R

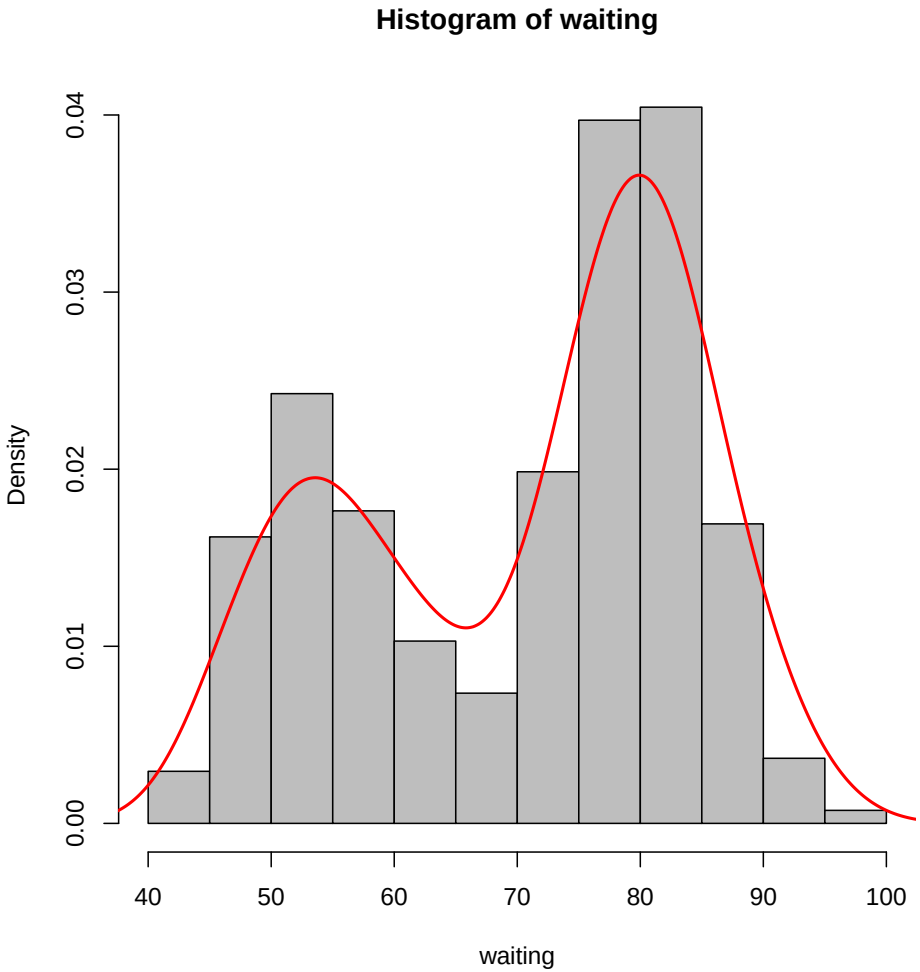
The empirical pdf $\hat{f}(x)$ is created on the basis of a so-called **kernel density estimate** defined as

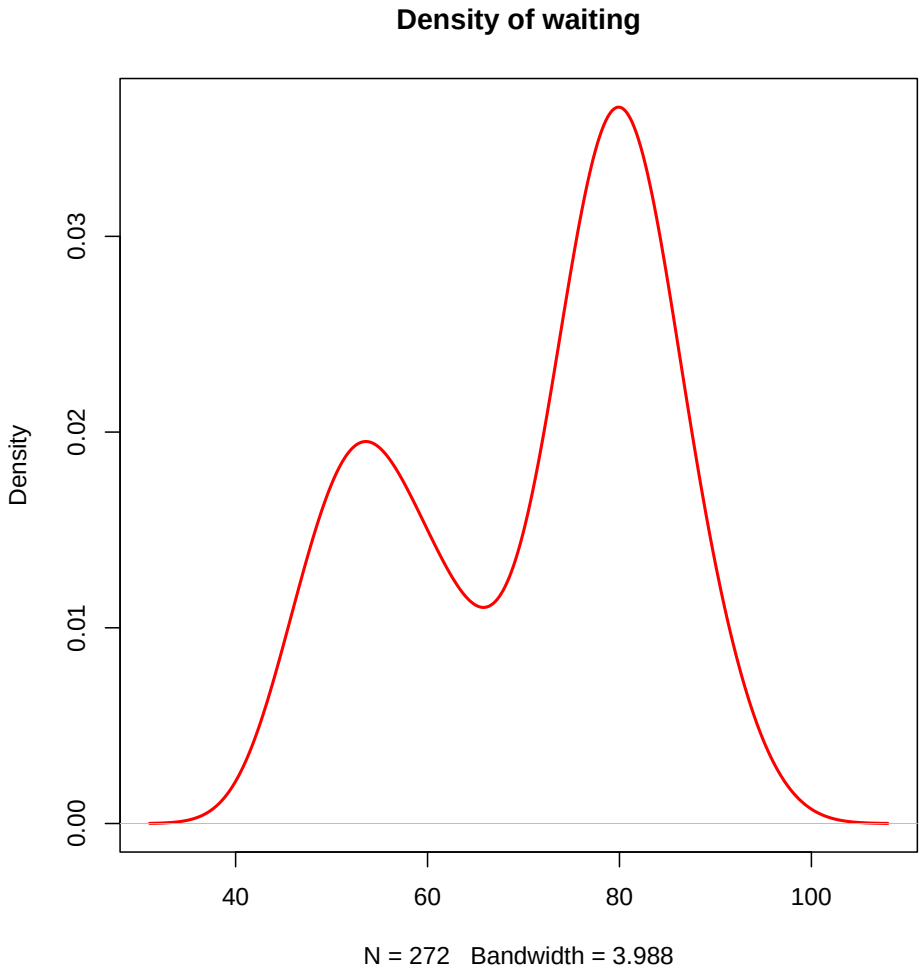
$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{287}$$

where $K(\cdot)$ is known as the kernel function and h is the bandwidth or smoothing parameter. Kernel functions are generally symmetric density functions which, again, must be non-negative and integrate to one. Common kernels are

- `Rectangular` $K(x) = \frac{1}{2},\ |x| < 1$
- `Triangular` $K(x) = 1 - |x|,\ |x| < 1$
- `Gaussian` $K(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2},\ -\infty < x < +\infty$

Density estimation in R





The smooth red curve is the kernel density estimate $\hat{f}(x)$.

Density estimation in R

The kernel density estimate can be obtained using the R-command

density.

The R-code for the kernel density estimate of the rv $X = \textit{waiting}(\textit{time})$ of the data set "faithful" in the figure above reads as follows:

```
> faithful
> waiting=faithful[,2]
> par(mfrow=c(1,2))
> hist(waiting, freq=F, col="grey",
+ main="Histogram of waiting")
> lines(density(waiting), col="red", lwd=2)
> plot(density(waiting), col="red", lwd=2,
+ main="Density of waiting")
```

Details on KDE

Mathematical-statistical details on kernel density estimation, including bias, variance and mse ($= \mathit{bias}^2 + \mathit{variance}$), as well as the choice of an optimal bandwidth parameter h, can be looked up in

`https://towardsdatascience.com/understanding-histograms-and-kernel-density-\estimation-6f9a1f09f960`

Most packages such as *R* and *Scipy* make use of

Scott's rule: $h^* \approx 1.06 \cdot \hat{\sigma} \cdot n^{-1/5}$

Exponential distribution

2.4.3 The exponential distribution

This distribution is the "continuous twin" of the discrete Poisson distribution and is often used to model (random) times. Some examples are

- the lifetime of an individual picked at random from some biological population
- the time until decay of a radioactive atom
- the length of time a patient survives after an operation
- the time it takes a computer to process a job of some kind

Such random times will be regarded as random variables with range $[0, \infty)$. The simplest model for a random time with no upper bound on its range is the exponential distribution.

Definition: The random variable T has **exponential distribution with rate** λ, briefly $T \sim \text{Exp}(\lambda)$, where λ is a positive parameter, $\lambda > 0$, if

Exponential distribution

T has pdf

$$f_T(t) = \lambda \cdot e^{-\lambda t}, \quad t > 0 \tag{288}$$

Equivalently, for $0 \leq a < b < \infty$,

$$\begin{aligned} P(a < T \leq b) &= \int_a^b \lambda e^{-\lambda t} dt \\ &= e^{-\lambda a} - e^{-\lambda b}. \end{aligned} \tag{289}$$

To see that $f_T(t)$ is a probability density on $[0, \infty)$, let $a = 0$, and let $b \to \infty$ to find the total probability of 1 on $[0, \infty)$.
Set $a = 0, b = t$ to obtain the cdf of T as

$$P(T \leq t) = 1 - e^{-\lambda t}, \quad t > 0. \tag{290}$$

Exponential distribution

Setting $a = t$, and letting $b \to \infty$ we get the formula for the so-called **survival function**

$$\bar{F}(t) = P(T > t) = e^{-\lambda t}, \ t \geq 0$$

Note that the rate λ has an inverse relationship with the survival probability $P(T > t)$, with a large rate λ this probability will be small. Remarkably, the exponential distribution has the property that it is **memoryless**.

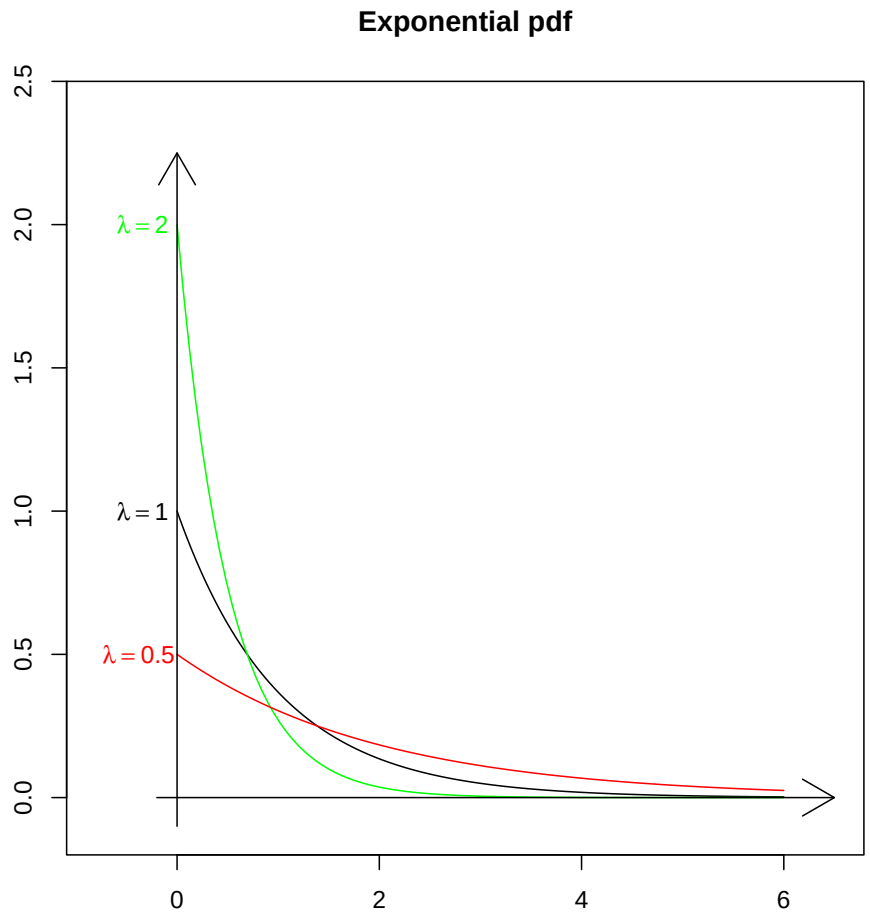
Corollary: If the rv T has an exponential distribution with rate λ, $T \sim \text{Exp}(\lambda)$, then T has the **memoryless property**

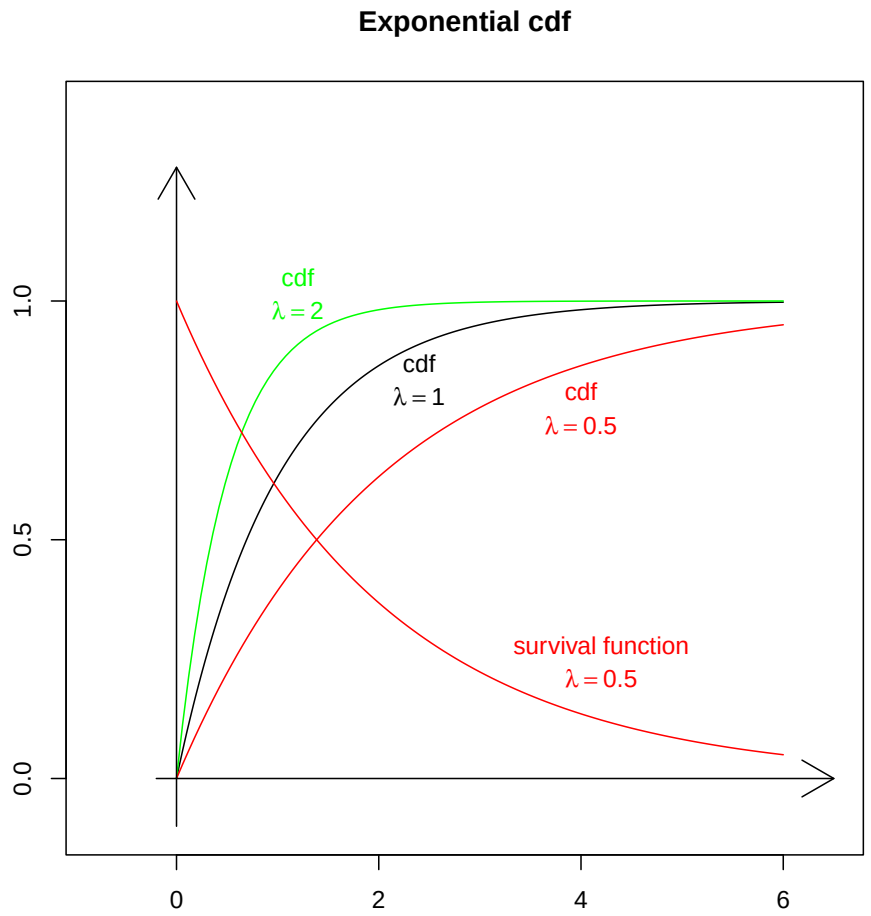
$$P(T > t + s | T > t) \quad = \quad P(T > s) \ \ \forall s, t \geq 0 \tag{291}$$

In words: Given survival to time t, the chance of surviving a further time s is the same as the chance of surviving to time s in the first place, regardless of the time t already survived.

Exponential distribution

This can be checked easily using the definition of the conditional probability, see Section 1.5, eq. (129). Before going on, let us graph the pdf, cdf and the survival function for different values of λ.





Exponential distribution

Even the converse of the above assertion on the memoryless property of the exponential distribution is true: the survival function $\bar{F}_T(t)$ satisfies the functional equation

$$\bar{F}_T(t+s) \quad = \quad \bar{F}_T(t) \cdot \bar{F}_T(s), \quad \forall t > 0, s > 0 \tag{292}$$

with $\bar{F}_T(t)$ decreasing and bounded between 0 and 1. It can be shown that every such function must be of the form $e^{-\lambda t}$ for some λ. So, the memoryless property is characteristic of the exponential distribution, no other distribution has this property.
This property simply means that, whatever the current age of the object, the distribution of the remaining lifetime is the same as the original lifetime distribution. Some objects, such as atoms or electrical components have this property (**no aging!**), hence have exponential distribution. But most forms of life do not have this distribution because they experience an aging process.

Exponential distribution

Interpretation of the rate λ: Consider, for a time t and a further length of time Δ, the probability

$$P(T \leq t + \Delta | T > t) = 1 - P(T > t + \Delta | T > t)$$
$$= 1 - P(T > \Delta) = 1 - e^{-\lambda \Delta}$$

by the memoryless property. Now, using the Taylor series expansion

$$e^{-\lambda \Delta} = 1 - \lambda \Delta + \tfrac{1}{2} \lambda^2 \Delta^2 - \dots$$

we get for small Δ that

$$P(T \leq t + \Delta | T > t) \approx \lambda \cdot \Delta,$$

with error negligible in comparison to Δ as $\Delta \to 0$. This is but just the death rate λ times the length of the time interval within which the object will fail. Therefore, λ is also called the **instantaneous failure rate** or, likewise, the **hazard rate**.
The characteristic feature of exponentially distributed lifetimes is that this rate is **constant**, not depending on t.

Exponential distribution

For other continuous distributions on $(0, \infty)$ the death (hazard) rate is a time-dependent function $\lambda(t)$.

Exponential distribution in R:

All the necessary probabilities involving the exponential distribution can be easily computed in R as follows

- `dexp(x, rate= 2)` computes the pdf $f_T(x)$ of $T \sim \text{Exp}(\lambda = 2)$
- `pexp(x, rate= 1)` computes the cdf $F_T(x)$ of $T \sim \text{Exp}(\lambda = 1)$
- `qexp(prob, rate= 2)` computes the quantile x of $T \sim \text{Exp}(\lambda = 2)$ for a given probability $prob \in (0, 1)$
- For simulation `rexp(k, rate= 3)` generates k realizations of $T \sim \text{Exp}(\lambda = 3)$

Exponential distribution

Example (Reliability)

Some kinds of electrical components, for example, fuses and transistors, have a lifetime distribution well fitted by an exponential distribution. Such a component does not wear out gradually. Rather, it stops functioning suddenly and unpredictably. Roughly speaking, so long as it is still functioning, such a component is **as good as new**. Suppose the average lifetime of a transistor is 100 working hours, and that the lifetime is approximately exponential.

Q1: What is the probability that the transistor will work for at least 50 hours?

Since the average lifetime is $1/\lambda$, we put

$$1/\lambda = 100, \text{ so } \lambda = 0.01$$

and we calculate

$$P(T > 50) = 1 - P(T \leq 50) = 1 - pexp(50, rate = 0.01) = 0.6065.$$

Exponential distribution

Example (cont'd)

Q2: Given that the transistor has functioned for 50 hours, what is the chance that it fails in the next two minutes of use?

From the interpretation of λ as the instantaneous rate of failure per hour given survival so far, the probability is about

$\lambda \cdot \Delta = 0.01 \cdot \frac{2}{60} \approx 0.00033.$

Alternatively, using R, we have, of course, the same result:

$pexp(2/60, rate = 0.01) = 0.00033.$

Example (Radioactive decay)

Atoms of radioactive isotopes like Carbon 14, Uranium 235, or Strontium 90 remain intact up to a random instant in time when they suddenly decay, meaning that they split or turn into some other kind of atom. It is reasonable to assume that the random lifetime, or time until decay, must have the memoryless property.

Exponential distribution

Example (cont'd)

A common way to indicate the rate of decay of a radioactive isotope is by the half-life τ Thus, we have

$$P(T \leq \tau) = 1 - P(T > \tau) = 1 - e^{-\lambda\tau} = 1/2,$$

i.e. $\tau = log(2)/\lambda$ *and* $\lambda = log(2)/\tau$.

Strontium 90 is a particularly dangerous component of fallout from nuclear explosions. The substance is toxic, easily absorbed into bones when eaten, and has a long half-life of about 28 years.

Q1: What is the probability that a Strontium 90 atom survives at least 50 years?

First, observe that the decay rate is

$$\lambda = log(2)/\tau = log(2)/28 = 0.024755 \text{ per year.}$$

Thus, the probability of surviving at least 50 years is

$$P(T > 50) = e^{-50\lambda} = 1 - pexp(50, rate = 0.024755) = 0.29.$$

Exponential distribution

Example (cont'd)

Q2: How many years will it take until 99% of the Strontium 90 produced by the explosion has decayed?

We have to find t such that $P(T \leq t) \geq 0.99$, or, equivalently, $P(T \geq t) \leq 0.01$, i.e. we are looking for the 0.99 quantile t of the Sr 90 life time distribution:

$$e^{-0.024755t} = 0.01, \ so \ \ t = log(100)/0.024755 = 186.03 years.$$

Using R, we verify that $qexp(0.99, rate = 0.024755) = 186.03$

Relation to a Poisson process: We already know that the Poisson distribution comes out as a limit of Bernoulli trials when $n \to \infty$ and $p \to 0$ such that $np = constant = \lambda$, yielding a Poisson (arrival) process with rate λ on the time line $(0, \infty)$. Think of arrivals representing something like telephone calls coming in, particles arriving at a (Geiger) counter, or customers entering a store.

Relation between Poisson and Exponential

We then have two descriptions of such a Poisson arrival process:

- **Counts of arrivals.** The distribution of the number of arrivals $N(I)$ in a given fixed time interval I of length t is $\mathrm{Po}(\lambda t)$, and the numbers of arrivals in disjoint time intervals are independent.
- **Times between arrivals.** The distribution of the waiting time W_1 until the first arrival is $\mathrm{Exp}(\lambda)$, and W_1 and the subsequent waiting times $W_2, W_3, \ldots$ between each arrival and the next are independent, all with the same exponential distribution.

Corollary: These two descriptions of a random arrival process are **equivalent**.

Probabilities can then be calculated from whichever of these two descriptions is more convenient.

Relation between Poisson and Exponential

Example (Telephone calls)

Suppose calls are coming into a telephone exchange at an average rate of $\lambda = 3$ per minute, according to a Poisson arrival process. So, for instance, $N([2,4]) \sim \text{Po}(\lambda(4-2))$, and W_3, the waiting time between the second and third calls, has $\text{Exp}(\lambda = 3)$ distribution.

Q1: What is the probability that no calls arrive between $t = 0$ and $t = 2$?

Since $N((0,2]) \sim \text{Po}(6)$, this is

$$P(N((0,2]) = 0) = e^{-6} = dpois(0,6) = 0.00248.$$

Q2: What is the probability that the first call after $t = 0$ takes more than 2 minutes to arrive?

Since $W_1 \sim \text{Exp}(3)$, this is

$$P(W_1 > 2) = e^{-3*2} = 1 - pexp(2, rate = 3) = 0.00248.$$

This is the same as before, because the events are, in fact, identical.

2.4.4 Gamma distribution

The gamma distribution is often used to model the waiting time until a certain number of events (malfunction in engines, accidents at an intersection, and similar scenarios) in a Poisson process takes place. In the previous section we have seen that the waiting time until the first outcome in a Poisson process follows an exponential distribution.

Now, let W denote the waiting time until the α-th outcome ($\alpha \geq 1$) and let us find the distribution of W. Clearly, $F_W(w) = 0$ for $w < 0$. For $w \geq 0$, the event "$W > w$" means "fewer than α outcomes in $[0, w]$", and thus,

$$\begin{aligned} F_W(w) &= 1 - P(W > w) \\ &= 1 - \sum_{k=0}^{\alpha-1} \frac{(\lambda w)^k}{k!} e^{-\lambda w} \end{aligned} \tag{293}$$

Gamma Distribution

Consequently, when $w > 0$, the pdf of W is $f_W(w) = \frac{d}{dw} F_W(w)$. Then it follows that

$$f_W(w) = -\sum_{k=0}^{\alpha-1} \frac{(\lambda w)^k}{k!} e^{-\lambda w}(-\lambda) - \sum_{k=0}^{\alpha-1} \frac{(\lambda w)^{k-1}}{k!} \lambda k e^{-\lambda w} \quad (294)$$

which, after some simplifications reduces to

$$f_W(w) = \frac{\lambda^\alpha}{(\alpha-1)!} w^{\alpha-1} e^{-\lambda w}.$$

Before we go on, let us review the definition of the **gamma function** from mathematics, which generalizes the factorial to non-integer values of $\alpha > 0$:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \ \alpha > 0 \quad (295)$$

Gamma Distribution

Some of the more important properties of the gamma function include

- For $\alpha > 0$ it holds $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$.
- If α is a positive integer, then $\Gamma(\alpha) = (\alpha - 1)!$
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Definition: Gamma distribution

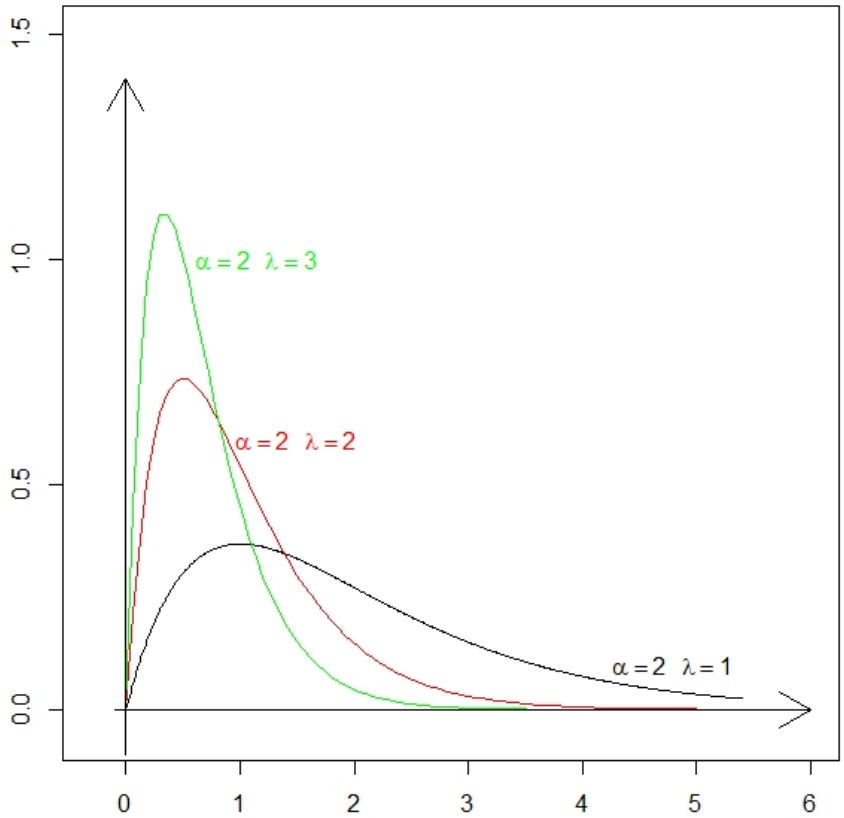
The random variable X has **gamma distribution with shape parameter** α and **rate parameter** λ, where α, λ are positive parameters, $\alpha, \lambda > 0$, if X has pdf

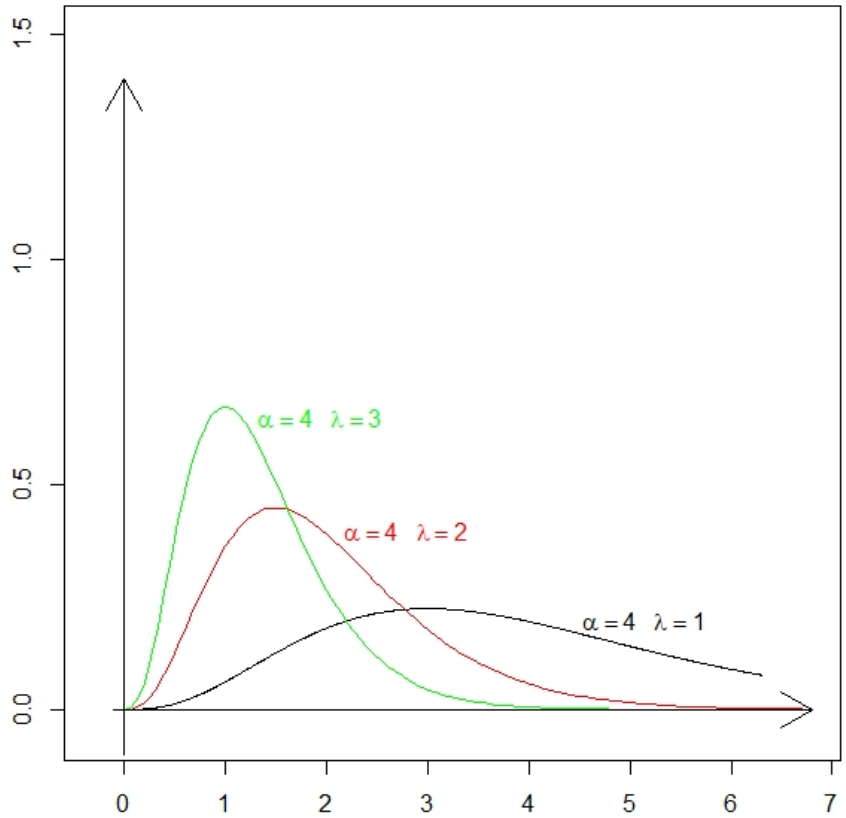
$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} e^{-\lambda x}, \quad x > 0 \tag{296}$$

We then briefly write

$$X \sim \mathrm{Ga}(\alpha, \lambda).$$

Gamma pdf's with different rates and shapes





Note: Varying λ changes the scale (units of measurement), but does not affect the shape of the density.

Gamma pdf's with different rates and shapes

Important relationships:

- For $\alpha = 1$, the resulting rv is $X \sim \text{Exp}(\lambda)$, i.e. the exponential distribution is a special case of the gamma distribution: $\text{Ga}(\alpha = 1, \lambda) = \text{Exp}(\lambda)$.
- When $\alpha = \frac{n}{2}$ and $\lambda = \frac{1}{2}$, the resulting variable has a Chi-square distribution with n degrees of freedom (this plays an important role in statistical estimation and testing of variances).
- For positive integer values α, the $\text{Ga}(\alpha, \lambda)$ distribution is also known as **Erlang** distribution. We already know that this distribution gives the waiting time until the α-th occurence when the number of outcomes in an interval of length t follows a Poisson distribution with parameter λt.

The gamma distribution has a lot of further applications, e.g. in meteorology (modelling the amount of (heavy) rainfall, the length and return period of droughts,...) and in banking/accounting (total assets, balance sheet totals,...).

Gamma distribution in R

In R, we can work with gamma distributions as follows

- `dgamma(x, shape= 2, rate= 4.5)` computes the pdf $f_X(x)$ of $X \sim \text{Ga}(\alpha = 2, \lambda = 4.5)$
- `pgamma(x, shape= 0.8, rate= 1)` computes the cdf $F_X(x)$ of $X \sim \text{Ga}(\alpha = 0.8, \lambda = 1)$
- `qgamma(prob, shape=3.5, rate= 2.4)` computes the quantile x of $X \sim \text{Ga}(\alpha = 3.5, \lambda = 2.4)$ for a given probability $prob \in (0, 1)$
- For simulation `rgamma(k, shape= 1.5, rate= 3.4)` generates k realizations of $X \sim \text{Ga}(\alpha = 1.5, \lambda = 3.4)$

Note: The cdf-values can also be computed using the **integrate()** function in R: e.g. for $X \sim \text{Ga}(4, 2)$ we have

$$P(X < 3.1) = pgamma(3.1, 4, 2) = 0.86577.$$

Gamma distribution in R

Using the integrate function, we arrive, up to five decimal places, at the same value:

```
> gam42= function(x) {dgamma(x, 4, 2)}
> integrate(gam42, 0, 3.1)$value
 [1] 0.86577.
```

Example

Suppose that the average arrival rate at a local fast food drive-through window is three cars per minute.

Q1: What is the probability that at least five cars arrive within two minutes?

If the number of car arrivals follows a Poisson distribution with a rate of $\lambda = 3$ per minute, then the average rate of arrival for 2 minutes is six cars. Now, with $X \sim \text{Po}(\lambda = 6)$,

$$P(X \geq 5) = 1 - P(X \leq 4) = 1 - \sum_{k=0}^{4} \frac{6^k}{k!} e^{-6} = 0.71494.$$

Gamma distribution in R

Example (cont'd)

To solve the problem with R, we use `ppois()`:

```
> 1- ppois(4,6)
 [1] 0.71494
> # or, alternatively,
> ppois(4, 6, lower = FALSE)
 [1] 0.71494
```

Q2: What is the probability that more than one minute passes by before the next car arrives?

Let W represent the waiting time until the next (second) arrival. Then, $W \sim \text{Ga}(\alpha = 2, \lambda = 3)$, consequently,

$$P(W > 1) = 1 - P(W \leq 1) = 1 - \int_0^1 \frac{3^2}{\Gamma(2)} x\, e^{-3x}\, dx$$

Gamma distribution in R

Example (cont'd)

This can be easily computed, using integration by parts, with $u = 3x$, and $dv = 3e^{-3x}$, to yield $P(W > 1) = 1 - (1 - 4e^{-3}) = 0.1991$.
To solve the problem with R, use the commands `pgamma()` or `integrate`:

```
> 1 - pgamma(1, 2, 3)
 [1] 0.1991483
> gam23= function(x) {9*x*exp(-3*x)}
> integrate(gam23, 1, Inf)$value
 [1] 0.1991483
```

2.4.5 Hazard Function

When dealing with lifetime data, it is helpful at times to study other functions related to the pdf such as the reliability (survival) function, or

Hazard Function

the hazard function, which is also often called the failure rate. Suppose the rv T represents the useful life of some component with pdf and cdf given by $f_T(t)$ and $F_T(t)$, respectively.

Definition: (i) The **reliability (survival) function** $R(t)$ is defined as

$$R(t) = P(T > t) = 1 - F_T(t), \ t > 0$$

and represents the probability that the lifetime of the component exceeds the time t.

(ii) The **hazard function** $h(t)$ is defined as

$$h(t) = f_T(t)/(1 - F_T(t)) = f_T(t)/R(t), \ t > 0, F_T(t) < 1.$$

The functions $h(t), f_T(t)$ and $F_T(t)$ provide mathematically equivalent specifications of the distribution of T. In fact, it can be shown that

$$f_T(t) = h(t) \cdot exp\left(-\int_0^t h(x)dx\right) \tag{297}$$

Hazard Function

Corollary: The hazard function $h(t)$ represents the instantaneous rate of death or failure at time t, given the individual or component has survived to time t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \,|\, T > t)}{\Delta t} \tag{298}$$

In other words, $h(t) \cdot \Delta t$ represents the approximate probability of having a breakdown during the interval $(t, t + \Delta t)$ given that this component has lasted up to to time t. Note, however, that the hazard function itself is a rate rather than a probability. We already know that the failure rate for an exponential rv is a constant λ:

$$h(t) = \frac{f_T(t)}{1 - F_T(t)} = \frac{\lambda e^{-\lambda t}}{1 - [1 - e^{-\lambda t}]} = \lambda$$

due to the (unique) memoryless property. Not many components have a constant failure rate. For most manufactured items as well as human populations, the failure rate increases with age, after some initial time

Weibull Distribution

period. The gamma distribution provides an adequate model for a more flexible lifetime distribution. However, since the hazard function for the gamma does not have a closed form expression, and moreover, it approaches λ from both above (when $\alpha < 1$) and below (when $\alpha > 1$) as t gets large, distributions with closed form expressions for $h(t)$ such as the Weibull tend to be favoured by practitioners. The hazard rate for this distribution is

$$h(t) = \alpha t^{\alpha-1}/\beta^{\alpha},\ \alpha, \beta > 0$$

Using eq. (297) we derive the pdf of the Weibull distribution as follows.

Definition: The random variable X is said to have a **Weibull** distribution with parameters $\alpha > 0$ and $\beta > 0$ if it has the pdf

$$f_X(x) = \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1}\ exp\left(-(x/\beta)^{\alpha}\right),\ x \geq 0 \tag{299}$$

We briefly write

$$X \sim \text{Weib}(\alpha, \beta).$$

Weibull Distribution

In the Weibull distribution defined above, the parameter α is the shape parameter, and the second one, β, is the scale parameter. In the special case, where $\alpha = 1$, the Weibull distribution reduces to the exponential one:

$$X \sim \text{Weib}(\alpha = 1, \beta) \Rightarrow X \sim \text{Exp}(\lambda = \frac{1}{\beta}) \tag{300}$$

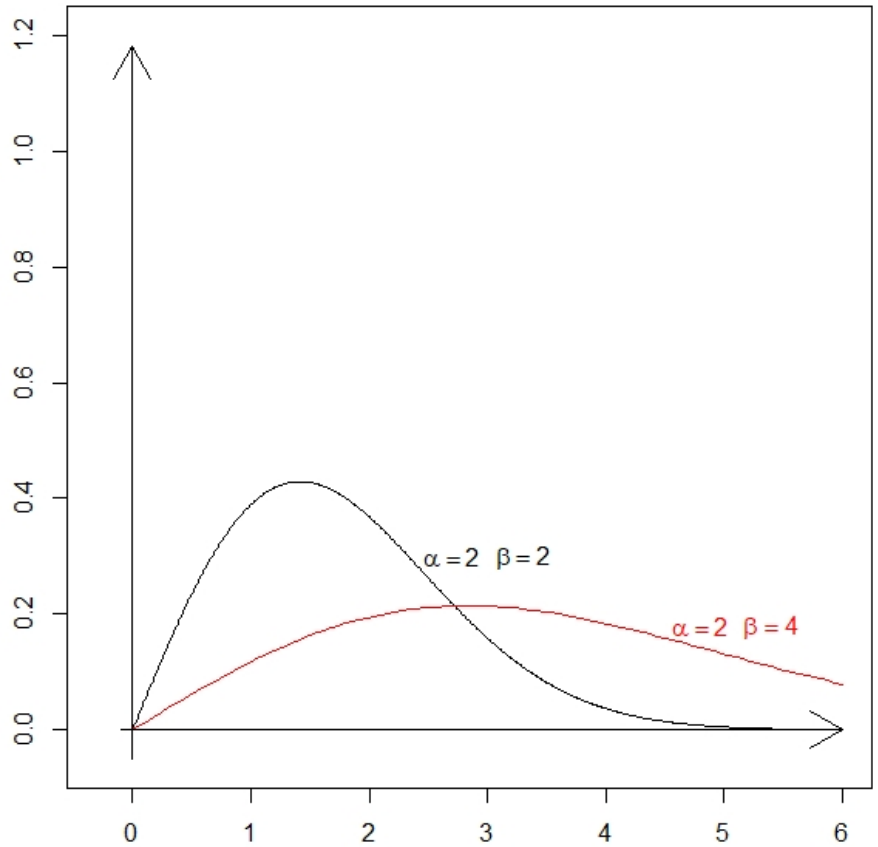
The cdf of the Weibull distribution can be easily obtained by integration, for $X \sim \text{Weib}(\alpha, \beta)$ we have

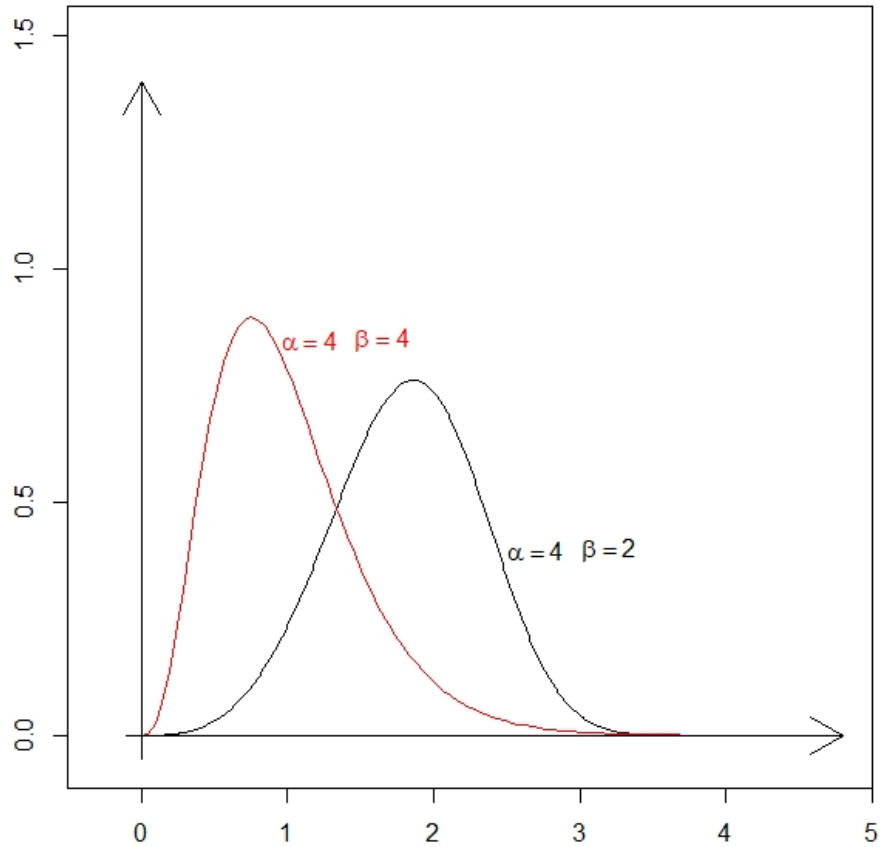
$$F_X(x) = \int_0^x \alpha\beta^{-\alpha} t^{\alpha-1} e^{-(t/\beta)^\alpha} dt = [-exp(-(t/\beta)^\alpha)]_0^x$$

$$= 1 - exp(-(x/\beta)^\alpha).$$

The following graphs give an illustration of the pdf's of the Weibull distribution for various choices of α and β.

Weibull pdf's with different shapes and scales





Note that with increasing shape values the distribution becomes more concentrated.

Weibull distribution in R

In R, we can work with Weibull distributions as follows

- `dweibull(x, shape= 2, scale= 4.5)` computes the pdf $f_X(x)$ of $X \sim \text{Weib}(\alpha = 2, \beta = 4.5)$
- `pweibull(x, shape= 0.8, scale= 1)` computes the cdf $F_X(x)$ of $X \sim \text{Weib}(\alpha = 0.8, \beta = 1)$
- `qweibull(prob, shape=3.5, scale= 1.4)` computes the quantile x of $X \sim \text{Weib}(\alpha = 3.5, \beta = 1.4)$ for a given probability $prob \in (0, 1)$
- For simulation `rweibull(k, shape= 1.5, scale= 3.4)` generates k realizations of $X \sim \text{Weib}(\alpha = 1.5, \beta = 3.4)$

Example

The useful life (in thousands of hours) of a certain type of transistor follows a Weibull distribution with $\alpha = 2.2$ and $\beta = 8.4$.

Q: What is the probability that a randomly selected transistor lasts more than 8000 hours?

Weibull distribution in R

Example (cont'd)

Using the cdf of the Weibull, we immediately have that

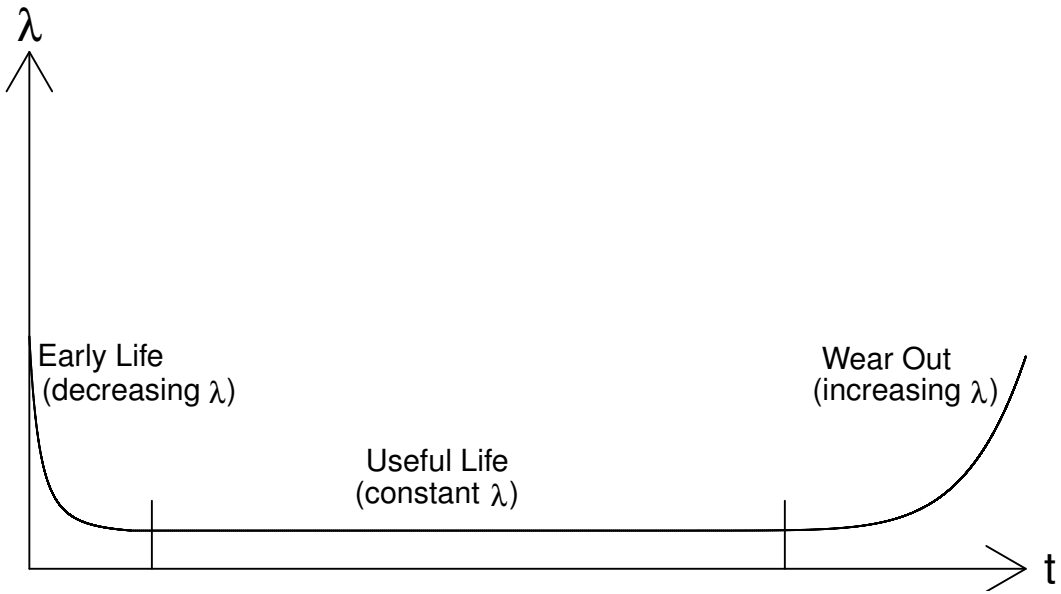
$$P(X > 8) = 1 - F_X(8) = 1 - \left[1 - e^{-(8/8.4)^{2.2}}\right]$$
$$= e^{-0.8982217} = 0.4072933.$$

With R, we get

```
> 1 - pweibull(8, 2.2, 8.4)
 [1] 0.4072933
> #or, equivalently
> pweibull(8, 2.2, 8.4, lower=F)
 [1] 0.4072933
```

Closing Remarks on Reliability Modeling

Some final remarks on lifetime modeling are in order. In practice, one usually has sufficient data to model the failure rate (hazard rate) for components that have been in use for some time t, and you can easily estimate this rate, for example, by the number of failures per hour among similar components. Often it is found that these empirically estimated failure rates tend to follow a smooth curve, which, more or less, has a **bathtub** shape:



Closing Remarks on Reliability Modeling

It is then reasonable to fit an ideal model in which the hazard rate $h(t)$ would be some continuous function of t. The exponential distribution is the simplest possible model corresponding to constant failure rate λ (middle part of the curve above). Other distributions (Gamma, Weibull,...) correspond to time-varying failure rates. Eq. (297) shows the connection between the hazard rate and the pdf of a lifetime variable. Another useful relationship is that the reliability (survival) function can be obtained as

$$R(t) = exp\left(-\int_0^t h(x)\, dx\right) \tag{301}$$

which can be proven using logarithmic integration.

2.5 Change of Variable

Many problems require finding the distribution of some function of X, say $Y = g(X)$. The distribution of Y can then be found provided the function $g(\cdot)$ has a derivative dy/dx which does not equal zero on any interval in the range of X.

Linear functions

Consider first a linear change of variable,

$$Y = aX + b, \ \textit{with constants } a, b \in \mathbb{R} \tag{302}$$

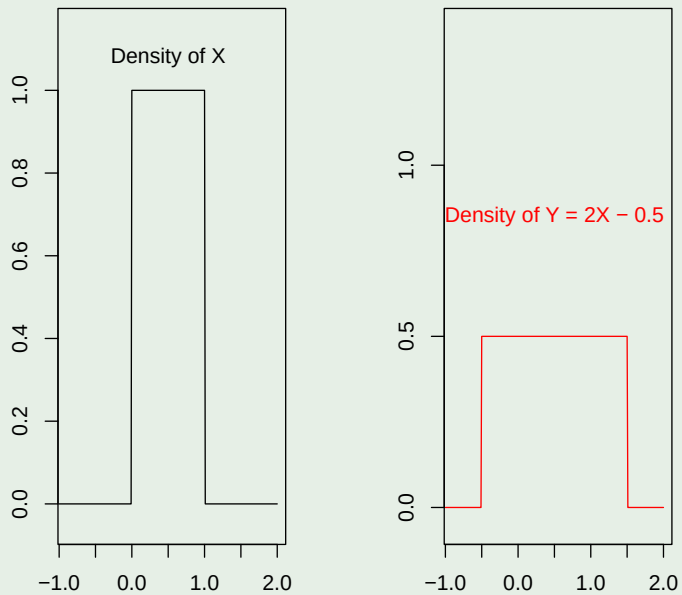
which describes e.g. the change from temperatures recorded in degrees of Celsius to those in degrees of Fahrenheit, where $a = \frac{9}{5}, b = 32$. Then, of course, the pdf (pmf in the discrete case) of $Y = aX + b$ at y is the density of X at the corresponding point $x = (y - b)/a$, the transformation from x to $ax + b$ multiplies units by a factor of $|a|$. This implies

Linear Transformation

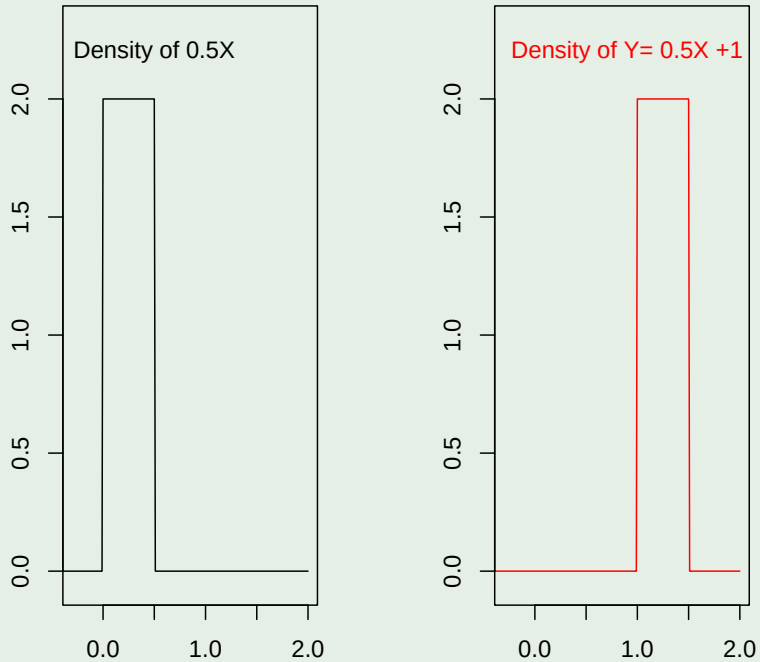
$$f_Y(y) = \frac{1}{|a|} \cdot f_X\left(\frac{y-b}{a}\right) \tag{303}$$

Example (Linear change of uniform rv's)

Let $X \sim \mathrm{U}(0,1)$. If $a > 1$ the range is spread out and the density decreases. For $0 < a < 1$ the range is shrunken and the density increases. Adding $b > 0$ shifts to the right by b, and adding $b < 0$ shifts to the left by $-b$.



Example (cont'd)



One-to-one Differentiable Functions

Let X be a rv with pdf $f_X(x)$ on the range $(a, b) \subset \mathbb{R}$. Let $Y = g(X)$ where $g(\cdot)$ is either strictly increasing or strictly decreasing on (a, b).

One-to-one Transform

For example, we might have

$$X \sim \text{Exp}(\lambda) \text{ and } Y = X^2, Y = \sqrt{X}, Y = 1/X, Y = log(X), ...$$

along with $(a, b) = (0, \infty)$. The range of Y is then an interval with endpoints $g(a)$ and $g(b)$.

The aim now is to calculate the pdf $f_Y(y)$ for y in the range of Y. Consider infinitesimal intervals dy and dx near y and the unique x such that $y = g(x)$, respectively. Then we have the identities

$$P(Y \in dy) = P(X \in dx), \text{ where } y = g(x) \tag{304}$$

i.e.

$$f_Y(y)dy = f_X(x)dx \tag{305}$$

and thus

$$f_Y(y) = f_X(x) \cdot \left|\frac{dx}{dy}\right| = f_X(x) / \left|\frac{dy}{dx}\right|$$

One-to-one Change of Variable for Densities

Theorem : Let X be a rv with pdf $f_X(x)$ on the range $(a.b)$ and $Y = g(X)$, where $g(\cdot)$ is either strictly increasing or strictly decreasing on (a, b). Then the pdf of Y is

$$f_Y(y) = f_X(x) / \left|\frac{dy}{dx}\right| \quad \textit{where } y = g(x),\ x = g^{-1}(y) \tag{306}$$

Example (Square root of an exponential rv)

Let be $X \sim \text{Exp}(\lambda = 1)$, *i.e.* $f_X(x) = e^{-x}, x \in (0, \infty)$.

Q: What is the pdf of $Y = \sqrt{X}$?

The range of X is $(a, b) = (0, \infty)$ and that of Y is $(g(a), g(b)) = (0, \infty)$, too. The function $g(x) = \sqrt{x}$ is one-to-one and

$$\frac{dy}{dx} = \frac{1}{2\sqrt{x}},\ x = g^{-1}(y) = y^2,\ \frac{dx}{dy} = 2y.$$

According to the theorem above we thus get

$$f_Y(y) = e^{-y^2} / \frac{1}{2\sqrt{y^2}} = 2y \cdot e^{-y^2},\ y > 0 \ \ (\textit{Rayleigh pdf}).$$

One-to-one Change of Variable for Densities

Example (Log of uniform)

Let X have uniform distribution on $(0,1)$, $X \sim \mathrm{U}(0,1)$.

Q: What is the pdf of $Y = -\frac{1}{\lambda} log(X),\ where\ \lambda > 0$?

Here $y = -\frac{1}{\lambda}\ log(x)$ has derivative

$$\frac{dy}{dx} = -\frac{1}{\lambda x}\ for\ 0 < x < 1,$$

so y decreases from ∞ to 0 as x increases from 0 to1. The pdf of Y is then

$$f_Y(y) = f_X(x)/|\frac{dy}{dx}| = 1/\frac{1}{\lambda x} = \lambda x,$$

where $-\frac{1}{\lambda}\ log(x) = y,\ or\ x = e^{-\lambda y}$, so

$$f_Y(y) = \lambda e^{-\lambda y},\ \ y > 0.$$

Conclusion: Y is exponentially distributed with rate λ. This way of obtaining an exponential rv as a function of a uniform $\mathrm{U}(0,1)$ rv is a standard method of simulating exponential rv's by computer.

Probability Integral Transform, PIT

This idea is based on a more general concept known as Probability Integral Transform (PIT): Suppose that an rv X has a continuous distribution with cdf F_X. Then the rv Y has a standard uniform distribution, i.e.

$$F_X(X) \sim \mathrm{U}[0,1]$$

Proof: Given $y \in [0,1]$ such that $F_X^{-1}(y)$ exists, then

$$F_Y(y) = P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F_X^{-1}(y))$$
$$= F_X(F_X^{-1}(y)) = y.$$

Applications:

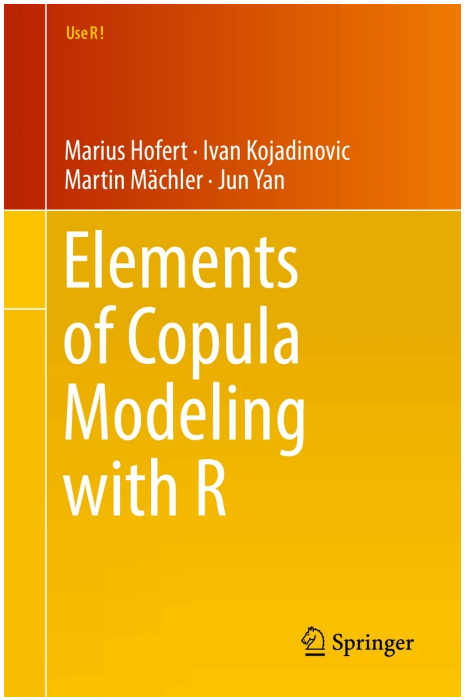
- PIT provides the basis for testing whether a set of observations can reasonably be modelled as arising from a specified distribution. This is used e.g. for constructing P-P-Plots and KS-tests.

PIT

- Inverse transform sampling: Converting an rv having a uniform distribution to follow a selected distribution, i.e.

$$Y \sim \mathrm{U}[0,1] \longrightarrow X = F_X^{-1}(Y) \sim F_X$$

- Copula theory: defining distributions for statistically dependent multivariate data based on standard uniform marginals.



3. Characteristic measures of probability distributions

What are the features of a random variable that describe its distribution best? In the sequel, we will consider the most important quantities summarizing the distribution efficiently.

3.1 Expectation of random variables

Consider a random variable X with cdf $F_X(x)$. This is a step function when X is discrete with values x_i and probabilities p_i; $i = 1, ..., n, ...$
When X is continuous then the cdf is generated by its underlying pdf $f_X(x)$.
The most important measure of the distribution of X is the expected value of its distribution, i.e. the point which is taken on the average.

Definition: Expected value of a discrete rv

Let be X a discrete random variable with

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n & \dots \\ p_1 & p_2 & \dots & p_n & \dots \end{pmatrix}. \tag{307}$$

Expected Value

Then we call

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i \cdot p_i \tag{308}$$

the **expected value (Erwartungswert)** of X.

In a similar way we define the expected value of a continuous rv, replacing the sum by an integral.

Definition: Expected value of a continuous rv

Let X be a continuous random variable with pdf f_X. Then we call

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx \tag{309}$$

the **expected value** of the continuous rv X.

Physically, the expected value is just the "center of mass" ("Schwerpunkt").

Expected Value of discrete rv's

Example (Rolling a die)

Let X be the number of points obtained when rolling a die.

$$X \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix} \tag{310}$$

For the expected value we then have

$$\begin{aligned} \mathbb{E}(X) = \sum_{i=1}^{6} x_i \cdot p_i &= \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \cdots + \frac{1}{6} \cdot 6 \\ &= \frac{1}{6} \cdot (1 + 2 + \cdots + 6) \\ &= 3.5 \end{aligned} \tag{311}$$

This means that in a long series of rolling a single die the average number of points scored is 3.5.

Expected Value of discrete rv's

Example (Expected value of Poisson distribution)

Let $X \sim \text{Po}(\lambda)$, then we have

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{\infty} k \cdot P(X = k) \qquad (312)\\
&= \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda}\\
&= 0 + \sum_{k=1}^{\infty} k \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda}\\
&= e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}\\
&= \lambda \cdot e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}
\end{aligned}$$

Expected Value of Poisson rv

Example (cont'd)

With an index transformation $m = k - 1$, we finally obtain

$$\mathbb{E}(X) = \lambda \cdot e^{-\lambda} \cdot \underbrace{\sum_{m=0}^{\infty} \frac{\lambda^m}{m!}}_{=e^{\lambda}} = \lambda. \tag{313}$$

Example (Expected Value of a Binomial rv)

If $X \sim \text{Bi}(n, p)$, then

$$\begin{aligned} \mathbb{E}(X) &= \sum_{k=0}^{n} k \cdot P(X = k) \\ &= \sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot p^k (1-p)^{n-k} \end{aligned} \tag{314}$$

Expected Value of Binomial rv

Example (cont'd)

$$
\begin{aligned}
\mathbb{E}(X) &= np \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} \cdot p^{k-1}(1-p)^{n-k} \qquad (315)\\
&= np \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!\,[(n-1)-(k-1)]!} \cdot p^{k-1}(1-p)^{(n-1)-(k-1)}\\
&= np \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} \cdot p^{k-1}(1-p)^{(n-1)-(k-1)}\\
&= np \cdot \underbrace{\sum_{m=0}^{n-1} \binom{n-1}{m} \cdot p^{m}(1-p)^{(n-1)-m}}_{=[p+(1-p)]^{n-1}=1}\\
&= n \cdot p
\end{aligned}
$$

Expected Value of Gaussian rv

Example (Expectation of normal distribution)

Let $X \sim \mathrm{N}(\mu, \sigma^2)$, then we have for the expected value

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx \\ &= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\, dx \end{aligned} \tag{316}$$

Using the substitution (normalization!) $t = \frac{x-\mu}{\sigma}$, $\frac{dt}{dx} = \frac{1}{\sigma}$, we have $dx = \sigma \cdot dt$ and further obtain

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} (\mu + \sigma t) \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{t^2}{2}\right)\, \sigma\, dt$$

Expected Value of Gaussian rv

Example (cont'd)

$$\mathbb{E}(X) = \mu \cdot \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt}_{=1} + \frac{\sigma}{\sqrt{2\pi}} \cdot \underbrace{\int_{-\infty}^{\infty} t \cdot \exp\left(-\frac{t^2}{2}\right) dt}_{=0}$$

$$= \mu.$$

The first integral yields one since we are integrating over the pdf of the standard normal $\mathrm{N}(0,1)$. The second integral vanishes since the function $g(t) = t \cdot e^{-t^2/2}$ is an odd function (plot it!):

```
> integrate(dnorm,-Inf,Inf)$value
 [1] 1
> g= function(t) {t*exp(-t*t/2)}
> integrate(g,-Inf,Inf)$value
 [1] 0
```

Expected Value of $X \sim \text{Exp}(\lambda)$

Example (Expectation of Exponential Distribution)

Let $X \sim \text{Exp}(\lambda)$, then it has expectation

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx \\ &= \int_{0}^{\infty} x \cdot \lambda \cdot e^{-\lambda \cdot x}\, dx \end{aligned} \tag{317}$$

Using the substitution $t = \lambda x$, $\frac{dt}{dx} = \lambda$, we have $dx = \frac{1}{\lambda} \cdot dt$ and then obtain

$$\mathbb{E}(X) = \frac{1}{\lambda} \cdot \int_{0}^{\infty} t \cdot e^{-t}\, dt = \frac{1}{\lambda} \cdot \Gamma(2) = \frac{1}{\lambda},$$

where we have made use of the Gamma function $\Gamma(\cdot)$, see eq. (295).

Expected Value of transformed rv's

Consider a random variable X, either being discrete,

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n \dots \\ p_1 & p_2 & \dots & p_n \dots \end{pmatrix}$$

or continuous with pdf $f_X(x)$. When we do a change of variable from X to $Y = g(X)$ with a measurable function (see Section 2.1) $g(\cdot)$, not necessarily being one-to-one, then we define the expectation accordingly:

$$\mathbb{E}\left[g(X)\right] = \begin{cases} \sum_{i=1}^{\infty} g(x_i) \cdot p_i & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx & \text{if } X \text{ is continuous} \end{cases} \tag{318}$$

This can be used to define the variance of a random variable through expectation of the squared deviation of the rv. from its mean.

Definition of Variance

3.2 Variance of random variables

Let be X a random variable with expectation $\mathbb{E}(X)$ and $g(\cdot)$ a (measurable) transformation given by

$$g(X) = [X - \mathbb{E}(X)]^2 \tag{319}$$

Definition: The **variance** of X, briefly denoted as $Var(X)$, is defined as

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}\left[X - \mathbb{E}(X)\right]^2 \\ &= \int_{-\infty}^{\infty} [x - \mathbb{E}(X)]^2 \; dF_X(x) \end{aligned} \tag{320}$$

The variance represents the **mean quadratic dispersion around the expected value**. Actually, the integral in (320) is a so-called **Riemann-Stieltjes integral**, for a continuous rv X with pdf $f_X(x)$ this

Definition of Variance

just means

$$\mathrm{Var}(X) \quad = \quad \int_{-\infty}^{\infty} [x - \mathbb{E}(X)]^2 \, f_X(x) \, dx \tag{321}$$

and for a discrete rv X with values $x_1, \ldots, x_n, \ldots$ and corresponding probabilities $p_i = P(X = x_i),\ i = 1, 2, \ldots$ we have to take the sum

$$\mathrm{Var}(X) \quad = \quad \sum_{i=1}^{\infty} [x_i - \mathbb{E}(X)]^2 \cdot p_i \tag{322}$$

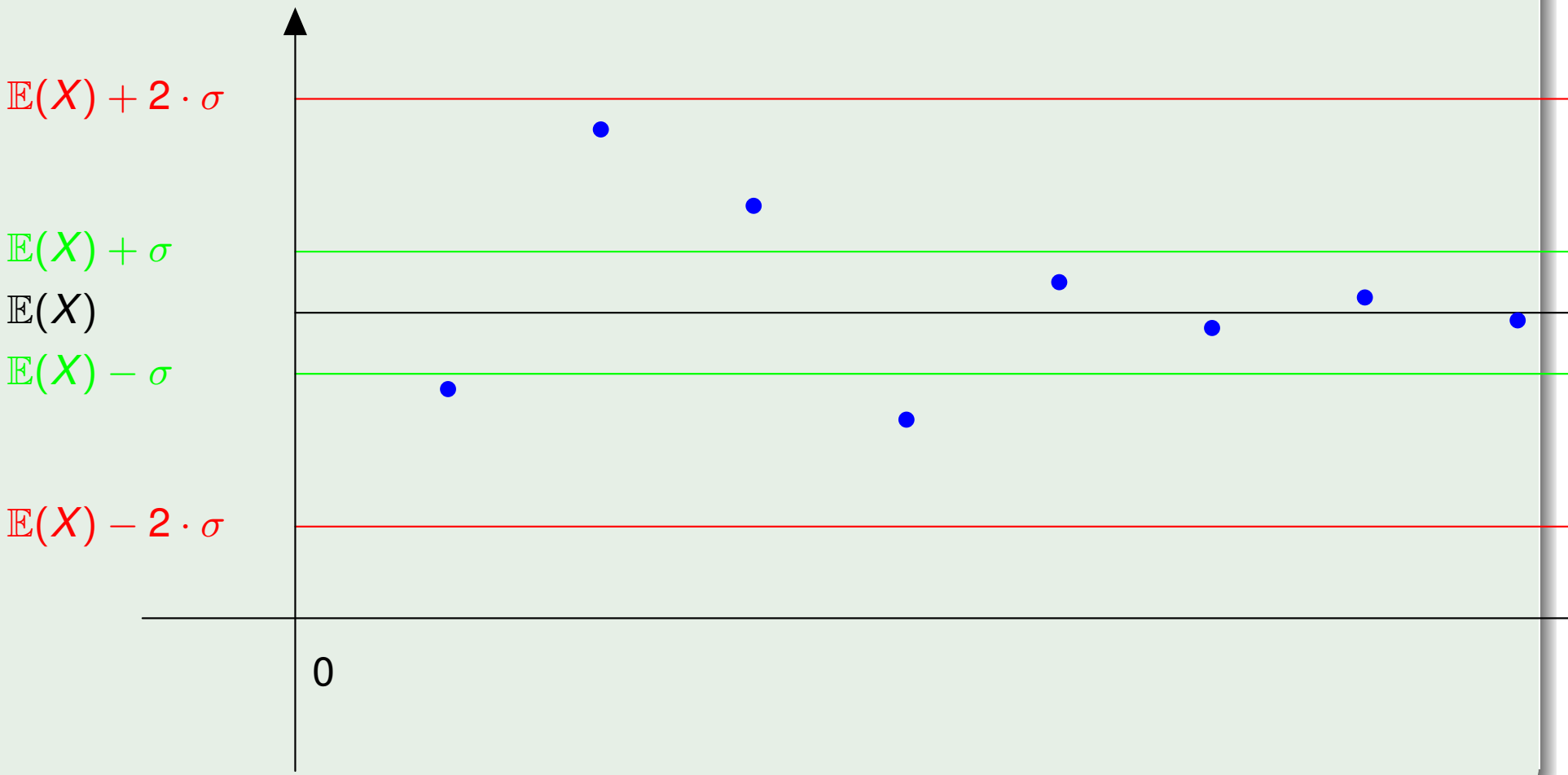
To better compare the degree of dispersion of different rv's we take the square root of the variance, which has the same measurement unit as the observed values $x_1, x_2, \ldots$

Definition: The **standard deviation** of the rv X is defined as

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} \tag{323}$$

Tolerance intervals in statistical quality control

Example (Control Charts in SQC)



Beyond control charts, the idea of tolerance bounds finds broad application in statistics for the construction of confidence intervals.

Variance and coefficient of variation

Clearly, the variance of a random variable depends on the measurement unit and usually increases with unit size. To circumvent this, a normalization could be useful.

Definition: For a non-negative rv X we call

$$v_X = \frac{\mathrm{SD}(X)}{\mathbb{E}(X)} \tag{324}$$

the **coefficient of variation** of X.

For example, consider two stock values X and Y with means $\mathbb{E}(X) = 40\$, \mathbb{E}(Y) = 80\$$ and standard deviations $\mathrm{SD}(X) = 8\$, \mathrm{SD}(Y) = 12\$$, respectively. Then,

$$v_X = \tfrac{8}{40} = 0.2, \quad v_Y = \tfrac{12}{80} = 0.15,$$

i.e. the stock value Y has smaller (coefficient of) variation than stock value X.

Variance of Poisson distribution

Example (Variance of $X \sim \text{Po}(\lambda)$)

Let $X \sim \text{Po}(\lambda)$, then

$$\begin{aligned} \text{Var}(X) &= \sum_{k=0}^{\infty}(k-\lambda)^2 \cdot p_k, \quad p_k = \frac{\lambda^k}{k!}e^{-\lambda} \qquad (325) \\ &= \sum_{k=0}^{\infty}(k^2 - 2\lambda k + \lambda^2)p_k \\ &= \sum_{k=0}^{\infty} k^2 p_k - 2\lambda \sum_{k=0}^{\infty} k p_k + \lambda^2 \sum_{k=0}^{\infty} p_k \\ &= \sum_{k=0}^{\infty} k^2 p_k - \lambda^2 \end{aligned}$$

observing that the second sum is just $\mathbb{E}(X) = \lambda$ (see eq. (313)) and the third sum yields 1. For the first sum we obtain

Variance of Poisson distribution

Example (cont'd)

$$\begin{aligned}
\sum_{k=0}^{\infty} k^2 p_k &= \sum_{k=0}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} \qquad (326)\\
&= e^{-\lambda} \cdot \left[\sum_{k=1}^{\infty} (k-1+1) \cdot \frac{\lambda^k}{(k-1)!} \right] \\
&= e^{-\lambda} \cdot \left[\left(\sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} \right) + \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \right] \\
&= e^{-\lambda} \cdot \left[\left(\lambda^2 \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \right) + \left(\lambda \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \right) \right]
\end{aligned}$$

Thus we have $\sum_{k=0}^{\infty} k^2 p_k = e^{-\lambda} \cdot \left(\lambda^2 \cdot e^{\lambda} + \lambda \cdot e^{\lambda} \right) = \lambda^2 + \lambda$, hence

$$\mathrm{Var}(X) = \sum_{k=0}^{\infty} k^2 p_k - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

3.3 Important Properties of Expectation and Variance

Before we go on we will list up some general properties of the expectation and variance operators.

(i) Linearity of Expectation:

$\mathbb{E}$ is a linear operator, i.e. for any linear transformation $Y = aX + b$ with constants $a, b \in \mathbb{R}$ it holds

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}(aX + b) \\ &= \int_{-\infty}^{\infty} (ax + b)\, dF_X(x) \\ &= a \cdot \int_{-\infty}^{\infty} x\, dF_X(x) + b \cdot \int_{-\infty}^{\infty} dF_X(x) \\ &= a \cdot \mathbb{E}(X) + b. \end{aligned} \tag{327}$$

In particular: $\mathbb{E}(b) = b$ for any constant $b \in \mathbb{R}$.

(ii) Non- negativity of the variance:
The variance of rv's is non-negative, $\mathrm{Var}(X) \geq 0$, since

$$\int_{-\infty}^{\infty} \underbrace{[x - \mathbb{E}(X)]^2}_{\geq 0} \, dF_X(x) \quad \geq \quad 0 \tag{328}$$

Equality only holds if $X = \mathbb{E}(X) = const$:

$$\mathrm{Var}(X) = 0 \quad \Longleftrightarrow \quad \exists c \in \mathbb{R} : P(X = c) = 1 \tag{329}$$

Clearly, if $P(X = c) = 1$ then $\mathbb{E}(X) = c$. Such a random variable X is said to be **degenerate**, it is just a deterministic constant.

Properties of Variance

(iii) Variance of linear transformations

For a linear change of variable from X to $Y = aX + b$ we have

$$\begin{aligned} \mathrm{Var}(Y) &= \mathbb{E}\left[Y - \mathbb{E}(Y)\right]^2 \\ &= \mathbb{E}\left[(aX + b) - \mathbb{E}(aX + b)\right]^2 \\ &= \mathbb{E}\left[aX + b - a \cdot \mathbb{E}(X) - b\right]^2 \\ &= \mathbb{E}\left[a \cdot X - a \cdot \mathbb{E}(X)\right]^2 \\ &= \mathbb{E}\left[a \cdot (X - \mathbb{E}(X))\right]^2 \\ &= a^2 \cdot \mathbb{E}\left[X - \mathbb{E}(X)\right]^2 \\ &= a^2 \cdot \mathrm{Var}(X) \end{aligned} \tag{330}$$

Scale factors enter the variance quadratically.
Moreover, constants have zero variance:

$$\mathrm{Var}(b) = 0.$$

3.4 Moments of a random variable

Definition: Let X be a random variable with cdf $F_X(x)$. Then

$$\mathbb{E}\left(X^k\right) = \int_{-\infty}^{\infty} x^k \cdot dF_X(x) \tag{331}$$

is called the k**-th moment** of X.

Note: For $k = 1$ we get the expected value $\mathbb{E}(X)$.

The second moment is related to the variance.

Theorem: (Steiner's translation theorem)
For any rv X it holds

$$\mathrm{Var}(X) = \mathbb{E}\left(X^2\right) - [\mathbb{E}(X)]^2 \tag{332}$$

Steiner's translation theorem

Proof:

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 \, dF_X(x) \qquad (333)\\
&= \int_{-\infty}^{\infty} \left(x^2 - 2x \cdot \mathbb{E}(X) + [\mathbb{E}(X)]^2 \right) \, dF_X(x)\\
&= \int_{-\infty}^{\infty} x^2 \, dF_X(x) - 2\,\mathbb{E}(X) \underbrace{\int_{-\infty}^{\infty} x \, dF_X(x)}_{\mathbb{E}(X)} + [\mathbb{E}(X)]^2 \underbrace{\int_{-\infty}^{\infty} dF_X(x)}_{=1}\\
&= \mathbb{E}\left(X^2\right) - 2\,\mathbb{E}(X) \cdot \mathbb{E}(X) + [\mathbb{E}(X)]^2\\
&= \mathbb{E}\left(X^2\right) - 2\,[\mathbb{E}(X)]^2 + [\mathbb{E}(X)]^2\\
&= \mathbb{E}\left(X^2\right) - [\mathbb{E}(X)]^2 .
\end{aligned}$$

Expectation and variance of Gamma Distribution

Example (Moments of Gamma distribution)

Let $X \sim \mathrm{Ga}(\alpha, \lambda)$, then it follows

$$\begin{aligned}
\mathbb{E}\left(X^k\right) &= \int_{-\infty}^{\infty} x^k \cdot f_X(x)\, dx \qquad (334)\\
&= \int_{0}^{\infty} x^k \cdot \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} e^{-\lambda x}\, dx\\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \int_{0}^{\infty} x^{k+\alpha-1} \cdot e^{-\lambda \cdot x}\, dx\\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+k)}{\lambda^{\alpha+k}}\\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha) \cdot \lambda^k}
\end{aligned}$$

Expectation and variance of gamma distribution

Example (cont'd)

Setting $k = 1$, we get the expected value as

$$\mathbb{E}(X) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\cdot\lambda} = \frac{\alpha}{\lambda}$$

and, using Steiner's theorem we obtain for the variance

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left(X^2\right) - [\mathbb{E}(X)]^2 \\ &= \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)\cdot\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 \\ &= \frac{(\alpha+1)\cdot\alpha\cdot\Gamma(\alpha)}{\Gamma(\alpha)\cdot\lambda^2} - \frac{\alpha^2}{\lambda^2} \end{aligned} \tag{335}$$

Thus, we finally have

$$\mathrm{Var}(X) = \frac{\alpha^2+\alpha}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}.$$

3.4.1 Existence/Non-existence of moments

Note: The moments of a distribution do not necessarily exist! This is especially true for so-called **heavy-tailed distributions.**

Example (Cauchy Distribution)

The random variable X is said to have a **Cauchy** distribution if it has the pdf

$$f_X(x) = \frac{1}{\pi \cdot (1 + x^2)}, \quad \infty < x < +\infty \tag{336}$$

We now show that for this distribution the expectation

$$\mathbb{E}(X) \quad = \quad \int_{-\infty}^{\infty} \frac{x}{\pi \cdot (1 + x^2)} \, dx \tag{337}$$

does not exist. This can be shown using logarithmic integration:

Non-existence of Cauchy moments

Example (cont'd)

$$\begin{aligned} \mathbb{E}(X) &= \frac{1}{2\pi} \cdot \int_{-\infty}^{\infty} \frac{2x}{1+x^2}\, dx \\ &= \frac{1}{2\pi} \cdot \ln(1+x^2) \Big|_{-\infty}^{\infty} \end{aligned} \tag{338}$$

This is, however, an undefined limit of the form $\infty - \infty$. Thus, the expectation does not exist.
In the above integration we have used the fact that the function $g(\cdot)$ defined by $g(x) = \ln(1+x^2)$ has derivative $g'(x) = 2x/(1+x^2)$, so the antiderivative (Stammfunktion) of $g'(x)$ is just $g(x)$.

Note: The fact that the first moment $\mathbb{E}(X)$ does not exist implies that all other moments $\mathbb{E}(X^k),\ k \geq 2$ do not exist, either. This can be concluded from the following, more general statement.

Existence of higher order moments

Corollary: Suppose that $\mathbb{E}\left(X^k\right)$ exists for some $k > 1$. Then there exist all moments $\mathbb{E}\left(X^r\right)$ with $1 \leq r \leq k$.

Proof: The assumption says that

$$I = \int_{-\infty}^{\infty} |x|^k \, dF_X(x) < \infty \tag{339}$$

Now let $r \leq k$, then it holds

$$\int_{-\infty}^{\infty} |x|^r \, dF_X(x) = \underbrace{\int_{-1}^{1} |x|^r \cdot dF_X(x)}_{\leq 1} + \underbrace{\int_{|x|>1} |x|^r \, dF_X(x)}_{\leq I} \tag{340}$$

observing that $|x|^r \leq 1$ for $x \in [-1, +1]$ and $|x|^r \leq |x|^k$ for $|x| > 1$.

Existence of higher order moments

Conversely, the non-existence of $\mathbb{E}(X^r)$ for a given $r \geq 1$ implies the non-existence of all higher order moments $\mathbb{E}(X^k)$ with $k > r$. This was the line of reasoning for the Cauchy distribution, for which neither the expected value nor the variance, not to mention moments of order $k > 2$, exist.

3.4.2 Method of Moments (MoM)

A well-known statistical method for estimating the parameters of a probability distribution F_X is the so-called **method of moments**, which proceeds by equating the empirical moments with the theoretical moments of X.

Definition

For a given sample $X_1, X_2, ..., X_n$ we define the **k-th empirical moment** as

$$E_{k,n} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i^k; \; k = 1, 2, ...$$

Method of Moments, MoM

In particular, $E_{1,n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ coincides with the sample average (arithmetic mean).

Now, equating the empirical moments with the theoretical moments of X and solving the resulting (nonlinear) equation system

$$E_{k,n} = \mathbb{E}\left(X^k\right);\ k = 1, ..., r$$

we get the MoM-estimates, where r denotes the number of unknown parameters.

Example

Let $X \sim \mathrm{Ga}(\alpha, \lambda)$, i.e. we have $r = 2$ unknown parameters, which can be estimated according to the method of moments by equating

$$\frac{\alpha}{\lambda} = E_{1,n} \ \textit{and} \ \frac{\alpha(\alpha+1)}{\lambda^2} = E_{2,n}$$

Example (cont'd)

Equivalently, the second equation reads: $\hat{\sigma}^2 = \alpha/\lambda^2$ (see (335)). Solving this equation system for α and λ leads us to the estimates

$$\hat{\alpha} = \bar{X}^2/\hat{\sigma}^2 \quad \textit{and} \quad \hat{\lambda} = \bar{X}/\hat{\sigma}^2$$

where $\bar{X}$ and $\hat{\sigma}^2$ denote the mean and variance of the data, resp.

3.5 Central Moments, Skewness and Kurtosis

Definition: Let X be a random variable with cdf $F_X(\cdot)$. Then

$$\mu_k(X) = \mathbb{E}\left[X - \mathbb{E}(X)\right]^k = \int_{-\infty}^{\infty} \left[x - \mathbb{E}(X)\right]^k \, dF_X(x) \tag{341}$$

is called the k**-th central moment** of X , $k = 1, 2, \ldots$. In particular, for $k = 2$ we have

$$\mu_2(X) = \mathrm{Var}(X). \tag{342}$$

Central moments

Note: The third and fourth central moments are often used to describe the shape of a probability distribution.

Example (Third central moment of a Gaussian rv)

For $X \sim N(\mu, \sigma^2)$ we have

$$\begin{aligned} \mu_3(X) &= \mathbb{E}\left[X - \mathbb{E}(X)\right]^3 \\ &= \int_{-\infty}^{\infty} (x-\mu)^3 \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\, dx \\ &= \int_{-\infty}^{\infty} (\sigma t)^3 \cdot \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} \cdot \sigma\, dt \end{aligned} \tag{343}$$

where, in the last step, we have made the substitution

$$t = \tfrac{x-\mu}{\sigma};\ \tfrac{dt}{dx} = \tfrac{1}{\sigma},\ dx = \sigma \cdot dt$$

Third central moment of a Gaussian rv

Example (cont'd)

Simplifying further, we get

$$\begin{aligned} \mu_3(X) &= \frac{\sigma^3}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} \underbrace{t^3 \cdot e^{-\frac{t^2}{2}}}_{\text{odd function}} \, dt \\ &= 0 \end{aligned} \tag{344}$$

This is also confirmed when using R with the `integrate` function. Likewise, since any function of the type

$$g(t) = t^{2k-1} * exp(-t^2/2)$$

is odd for $k = 1, 2, \ldots$, it follows that all odd central moments of normally distributed rv's are vanishing:

$$X \sim \mathrm{N}(\mu, \sigma^2) \implies \mu_{2k-1}(X) = 0 \ \ \forall k = 1, 2, \ldots \tag{345}$$

Fourth central moment

Example (4-th central moment of a Gaussian rv)

Let $X \sim N(\mu, \sigma^2)$, then the 4-th central moment reads

$$\begin{aligned} \mu_4(X) &= \mathbb{E}\left[X - \mathbb{E}(X)\right]^4 \\ &= 2 \cdot \int_0^\infty (x-\mu)^4 \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \, dx \end{aligned} \tag{346}$$

due to the symmetry of the function under the integral sign around $x = \mu$. We apply the following substitutions

$$t = \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2, \quad \frac{dt}{dx} = \frac{x-\mu}{\sigma^2}$$

$$dx = \frac{\sigma^2}{x-\mu}\, dt = \frac{\sigma}{\sqrt{2t}}\, dt$$

and then further obtain

Fourth central moment of a Gaussian rv

Example (cont'd)

$$\begin{aligned}\mu_4(X) &= \frac{2}{\sigma\sqrt{2\pi}} \int_0^\infty 4\sigma^4 t^2 \cdot e^{-t} \frac{\sigma}{\sqrt{2t}}\, dt \qquad (347)\\ &= 4\,\frac{\sigma^4}{\sqrt{\pi}} \int_0^\infty t^{3/2} e^{-t}\, dt \\ &= 4\,\frac{\sigma^4}{\sqrt{\pi}} \cdot \Gamma\left(\frac{5}{2}\right)\end{aligned}$$

Observing that

$$\Gamma\left(\tfrac{5}{2}\right) = \tfrac{3}{2}\,\Gamma\left(\tfrac{3}{2}\right) = \tfrac{3}{2}\,\tfrac{1}{2}\,\Gamma\left(\tfrac{1}{2}\right) = \tfrac{3}{4}\cdot\sqrt{\pi},$$

we finally have:

$$\mu_4(X) = 3\sigma^4.$$

Skewness

For distributions which are symmetric around the mean $\mathbb{E}(X)$, as it is the case for Gaussian rv's or Student-t-distributed rv's, it holds

$$\mu_3(X) = 0 \tag{348}$$

This gives rise to introduce a measure of skewness of distributions to quantify the degree of deviations from symmetry, based on $\mu_3(X)$.

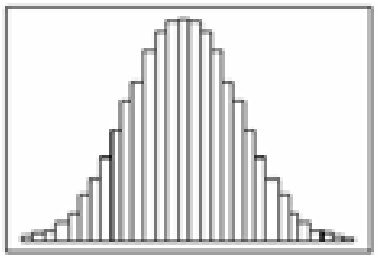
Definition: Skewness of a distribution

The **skewness** (= Schiefe) of the distribution of the rv X is measured by the coefficient

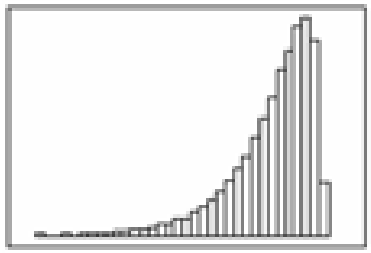
$$\gamma(X) = \frac{\mu_3(X)}{\mathrm{SD}(X)^3} = \frac{\mathbb{E}\left[X - \mathbb{E}(X)\right]^3}{\left[\sqrt{\mathrm{Var}(X)}\right]^3} \tag{349}$$

Skewness

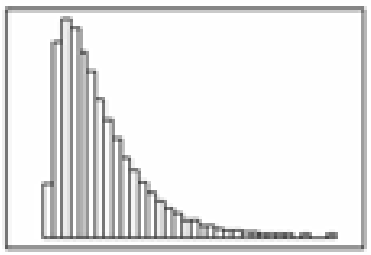
For nearly symmetric distributions we have $\gamma(X) \approx 0$, left-skewed distributions have negative skewness $\gamma(X) < 0$ and for right-skewed distributions we have $\gamma(X) > 0$.



Symmetric
Bell shaped



Skewed to
the Left



Skewed to
the Right

Kurtosis

The fourth central moment is used to characterize the degree of kurtosis of a distribution w.r.t. the Gaussian distribution.

Definition: Kurtosis (Steilheit, Wölbung)

The **kurtosis(excess)** of the distribution of the rv X is measured by the coefficient

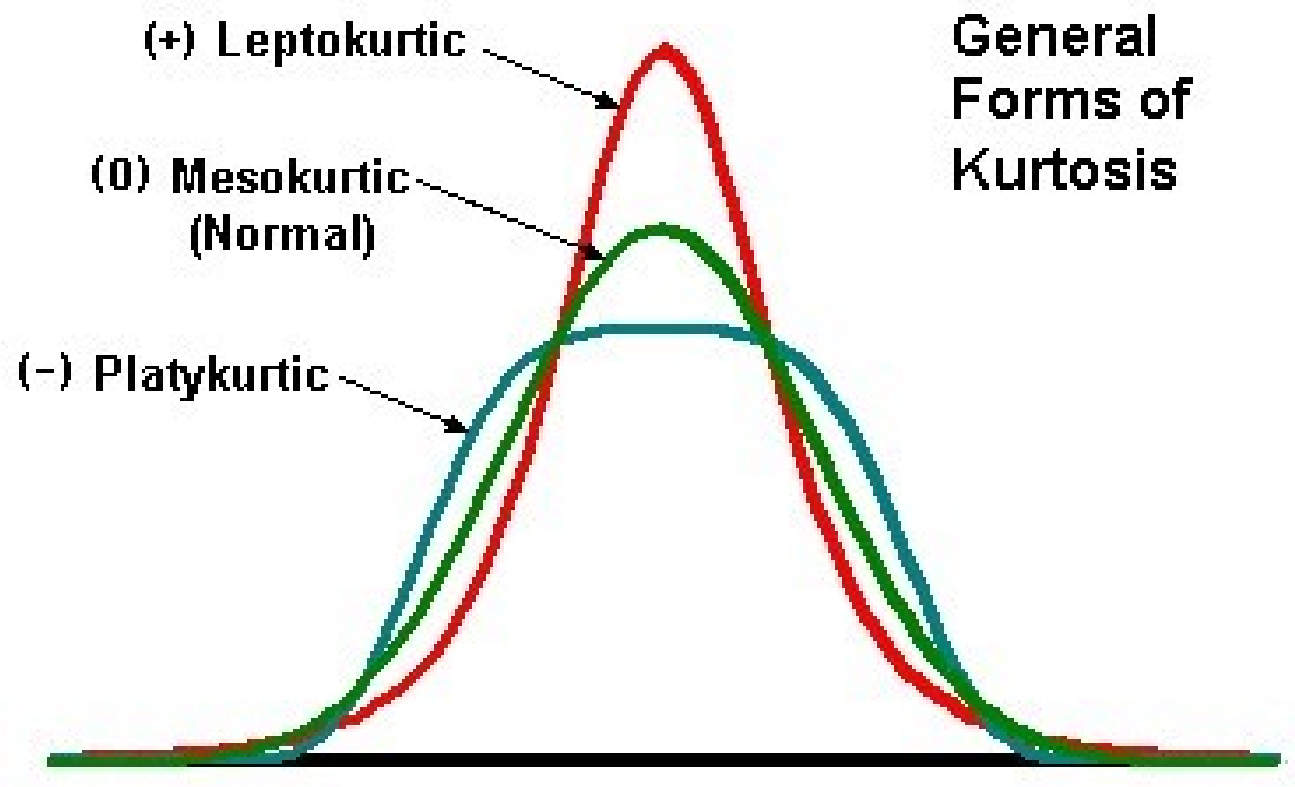
$$\epsilon(X) = \frac{\mu_4(X)}{\mu_2(X)^2} - 3 = \frac{\mu_4(X)}{\mathrm{SD}(X)^4} - 3 \tag{350}$$

We already know that a Gaussian rv $X \sim \mathrm{N}(\mu, \sigma^2)$ has 4-th central moment $\mu_4(X) = 3\sigma^4$, thus it has kurtosis

$$\epsilon(X) = 3\sigma^4/\sigma^4 - 3 = 0.$$

Distributions with positive kurtosis are also called **leptocurtic (steilgipflig)**, those with negative kurtosis are said to be **platykurtic (flachgipflig)**.

Different types of kurtosis



3.6 Mode and Median

Definition: The **Mode (Modalwert)** x_{mod} of the distribution of X is defined as the value for which it holds

$$f_X(x_{\text{mod}}) = \max_{x \in \mathbb{R}} f_X(x) \tag{351}$$

in case of a continuous X or

$$P(X = x_{\text{mod}}) = \max_{x \in \mathbb{R}} P(X = x) \tag{352}$$

for a discrete X. It is the value which occurs most often in the distribution.

The distribution of X is said to be **unimodal** if there exists only one mode. Should there exist more than one mode, then the distribution is said to be **multimodal**.

Mode of Distributions

Example (Mode of Bi(8, 0.5))

Consider the rv $X \sim \text{Bi}\left(n = 8, p = \frac{1}{2}\right)$, which has pmf

$$P(X = k) = \binom{8}{k} \cdot \left(\frac{1}{2}\right)^k \cdot \left(\frac{1}{2}\right)^{8-k} = \binom{8}{k} \cdot \frac{1}{256} \tag{353}$$

Tabulating the probabilities for the various values k of this distribution,

k	0	1	2	3	4	5	6	7	8
$\binom{8}{k}$	1	8	28	56	70	56	28	8	1

we see that $x_{\text{mod}} = 4$, since it has highest probability $P(X = 4) = \frac{70}{256}$.
This value coincides with the expected value:

$$x_{\text{mod}} = \mathbb{E}(X) = n \cdot p = 8 \cdot 0.5 = 4.$$

Mode of Binomial

Obviously, the mode of a binomial distribution is unique if the number of the values in the range of X is odd (nine, in the above example). The following general result is not hard to prove.

Corollary: Let X have a binomial distribution, $X \sim \text{Bi}(n, p)$. Then the mode is given by

$$x_{\text{mod}} = \texttt{floor}\,(p \cdot (n+1)) \tag{354}$$

In case that $p \cdot (n+1)$ is an integer, then there are two modes: $p \cdot (n+1)$ and $p \cdot (n+1) - 1$.

Remark: For a non-negative, real number x we mean by $\texttt{floor}(x)$ the largest integer part of x. For the above example, we have with R:

```
> n=8; p=0.5
> floor((n+1)*p)
 [1] 4
```

Mode of continuous distributions

Example (Mode of LogNormal)

Let be $Z \sim \mathrm{N}(\mu, \sigma^2)$ and $X = exp(Z)$. Then, X follows the so-called **Lognormal distribution** with parameters μ and σ^2, briefly written as $X \sim \mathrm{LN}(\mu, \sigma^2)$. Using the change-of-variable theorem we derive the pdf of X as follows:

$$f_X(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{1}{\sigma\sqrt{2\pi}}\, x^{-1} \exp\left[-\frac{1}{2\sigma^2}\,(\ln(x)-\mu)^2\right] & \text{for } x > 0 \end{cases} \tag{355}$$

Q: What is the mode of the lognormal distribution?

First, setting $g(x) = exp\left(-\frac{1}{2\sigma^2}(\ln(x)-\mu)^2\right),\ x > 0$, we write the pdf of X as

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\ \frac{g(x)}{x}\ I_{(0,\infty)}(x)$$

The necessary condition for obtaining the mode of X is just to require

Mode of Lognormal distribution

Example (cont'd)

$f'_X(x) = 0$. Now, differentiating $f_X(x)$ w.r.t. x leads to

$$f'_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, x^{-2}\, [x\, g'(x) \;-\; g(x)]\,,$$

where

$$g'(x) = -\frac{1}{\sigma^2}\, x^{-1}\, g(x)\, (\ln(x) - \mu).$$

Thus,

$$f'_X(x) = 0 \iff x\, g'(x) = g(x) \iff \ln(x) - \mu = -\sigma^2 \tag{356}$$

The solution to the latter equation finally yields the unique mode

$$x_{\text{mod}} \;=\; \exp(\mu - \sigma^2) \;=\; e^{\mu}/e^{\sigma^2} \tag{357}$$

Remark: It can easily be checked that $f''_X(x_{\text{mod}}) < 0$. Therefore, x_{mod} is really the (unique, global) maximum of $f_X(\cdot)$.

Median of the distribution

For a continuous distribution, which is symmetric (around its mean), the mode divides the distribution into equal halves containing just 50% of the total probability mass. We will now focus on such points in more generality.

Definition: The point $x_{0.5}$ is said to be the **median** of the distribution of the random variable X, if it holds

$$P(X \leq x_{0.5}) \geq 0.5 \tag{358}$$

$$\wedge \; P(X \geq x_{0.5}) \geq 0.5 \tag{359}$$

The median splits the probability mass of X into equal halves.

Corollary: For a continuous rv X, the median is determined as solution of the nonlinear equation

$$F_X(x_{0.5}) = 0.5 \tag{360}$$

Median of the distribution

Example (Median of Binomial)

Let $X \sim \text{Bi}\left(n = 8, p = \frac{1}{2}\right)$. The median is easily determined as $x_{0.5} = 4$, since

$$P(X \leq 4) = \texttt{pbinom(4,8,0.5)} = 0.6367 \quad (361)$$

$$P(X \geq 4) = \texttt{1 - pbinom(3,8,0.5)} = 0.6367 \quad (362)$$

Note: Due to the symmetry of the binomial distribution, we have $x_{0.5} = x_{\text{mod}} = 4$.

Example (Median of Lognormal)

Let $X \sim \text{LN}(\mu, \sigma^2)$, with pdf as given in (355). Equivalently, we then have by the change-of-variable theorem:

$$Y = \ln(X) \sim \text{N}(\mu, \sigma^2) \quad (363)$$

Median of the Lognormal Distribution

Example (cont'd)

Now, for $x > 0$ we have

$$F_X(x) = P(X \leq x) = P\left(e^Y \leq x\right) = P\left(Y \leq \ln(x)\right) = F_Y(\ln(x)) \quad (364)$$

After the normalization, $Z = (Y - \mu)/\sigma$ follows the standard normal distribution: $Z \sim \mathrm{N}(0, 1)$, which has median $z_{0.5} = 0$. Observing that

$$F_Y(\ln(x)) \quad = \quad F_Z\left(\frac{\ln(x) - \mu}{\sigma}\right) \quad (365)$$

we then have

$$F_X(x_{0.5}) = 0.5 \iff F_Z\left(\frac{\ln(x_{0.5}) - \mu}{\sigma}\right) = 0.5 \iff \ln(x_{0.5}) = \mu \quad (366)$$

Thus, we finally arrive at the median of X: $x_{0.5} = e^{\mu}$.

3.7 Quantile Function and Symmetry

We will now generalize the notion of the median, which divides the distribution into equal halves, to arbitrary partitionings of the distribution mass.

Definition: Let $0 < p < 1$, then we call x_p the p**-quantile** of the distribution of X, if it holds

$$P(X \leq x_p) \geq p \tag{367}$$

$$\wedge \quad P(X \geq x_p) \geq 1 - p \tag{368}$$

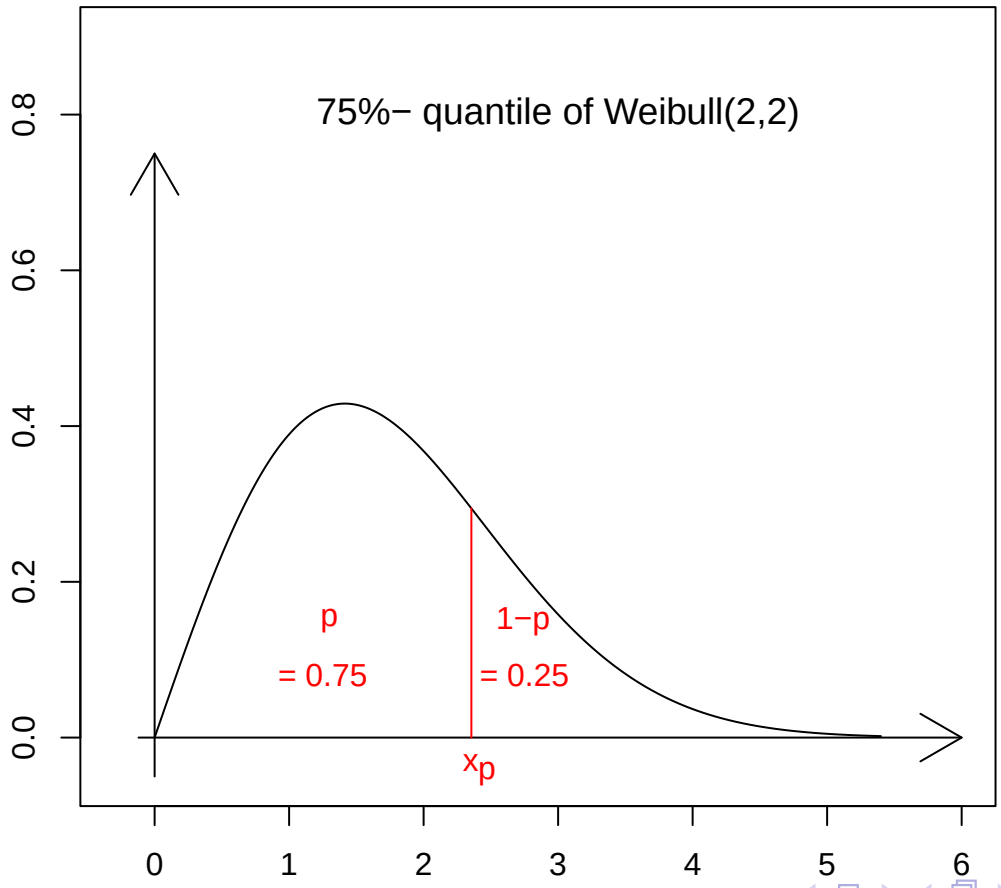
Corollary: If X is a continuous rv, then x_p is the solution of the (non-linear) equation

$$F_X(x_p) = p. \tag{369}$$

Quantiles

Interpretation: x_p splits the distribution mass of X according to the ratio

$$p \ : \ (1-p)$$



Quantiles

Definition: The function

$$Q_X : (0,1) \to \mathbb{R} \quad \textit{with} \quad Q_X(p) = x_p \; \forall p \in (0,1) \tag{370}$$

is called the **quantile function** of the distribution of X.

Corollary: Relationship between Q_X and F_X
For a continuous rv X it holds

$$F_X(x_p) = p \tag{371}$$

$$\Rightarrow x_p = F_X^{-1}(p) = Q_X(p) \tag{372}$$

If F_X is strictly monotone increasing then x_p is uniquely determined.

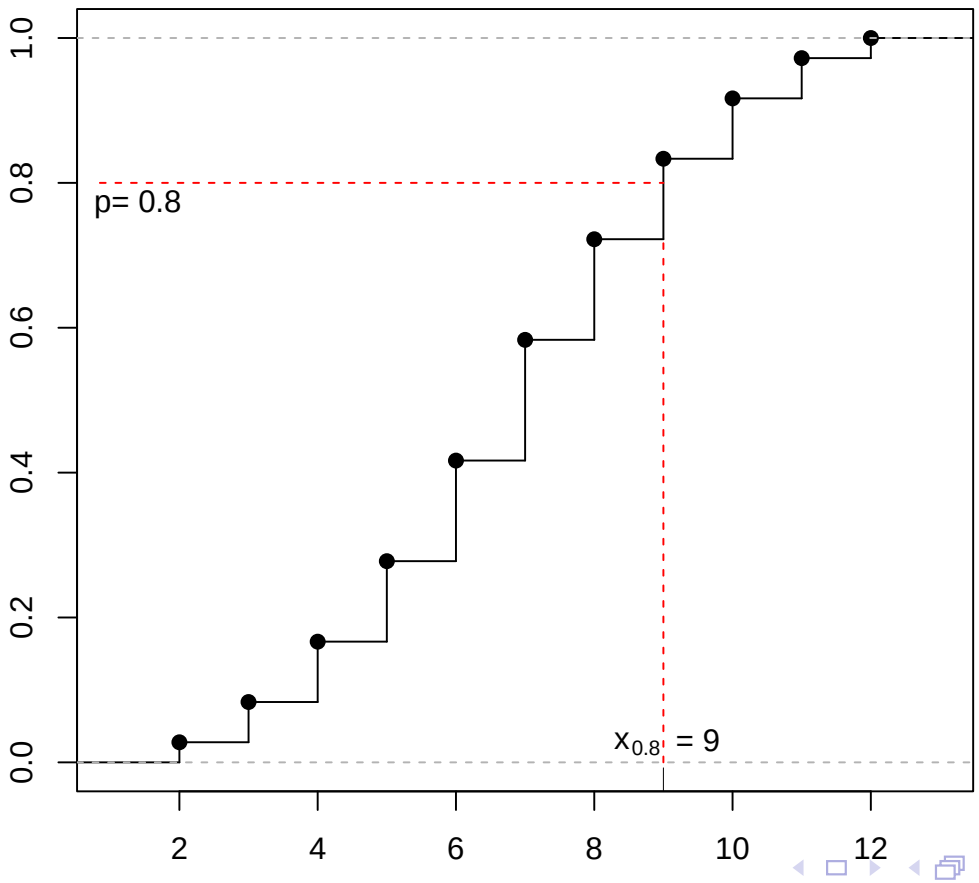
If the distribution function $F_X(\cdot)$ is not strictly monotone increasing then it is piecewise constant (stückweise konstant). This is, in particular, the case for all discrete rv's.

Quantile Function

In such cases we define

$$x_p = inf\{x \in \mathbb{R} : F_X(x) \geq p\} \tag{373}$$

as the smallest value for which the required probability is attained.



Quantile Function

The quantile function Q_X thus represents the generalized inverse of the distribution function F_X.

$$Q_X(p) = F_X^-(p) \tag{374}$$

As discussed for various distributions already before, the computation of quantiles with R is easily accomplished using the `qname` function, where "name" stands for the name of the distribution, e.g. "binom" and "norm" for the binomial and normal distribution, respectively.
For example, the 75%- quantile of the Weibull (2,2) distribution illustrated before, can be obtained as follows:

```
> qweibull(0.75, shape=2, scale=2)
 [1] 2.35482
```

With the help of the quantile function we are now also in a position to say what symmetry of the distribution really means.

Relationship between mean, median and mode

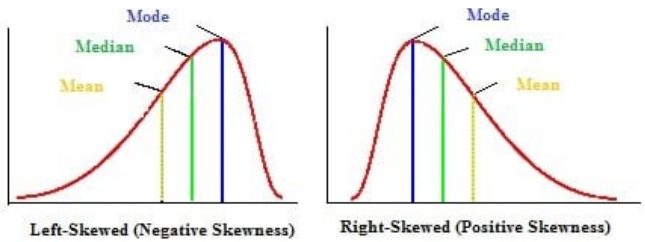
Definition: The distribution of a continuous rv X is said to be **symmetric**, if it holds

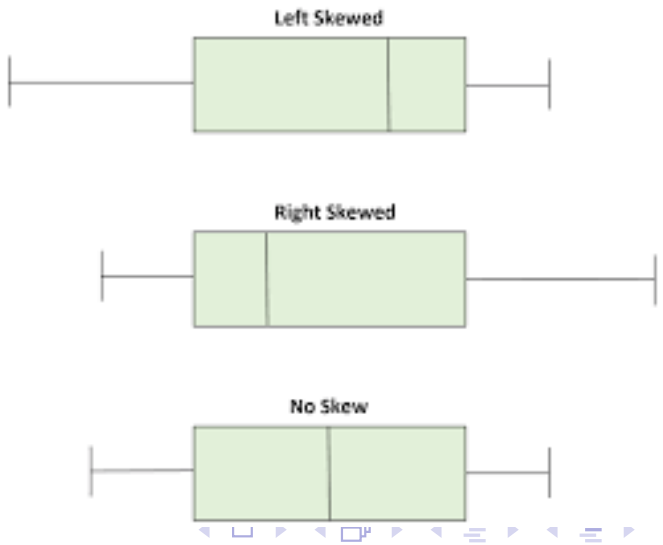
$$x_{1-p} - x_{0.5} = x_{0.5} - x_p \qquad \forall p \in (0, 0.5) \tag{375}$$

Moreover, for a symmetric distribution, we will always have that

$$\mathbb{E}(X) \;=\; x_{0.5} \;=\; x_{\text{mod}}$$

i.e. mean $=$ median $=$ mode. For non-symmetric distributions, however, these three characteristic values of F_X fall apart.





4. Law of large numbers, Central limit theorem

Consider a sequence $\{X_i\}$ of rv's defined on the same (Kolmogorov) Probability Space $[\Omega, \mathscr{E}, P]$.

Problem: How do the arithmetic means and the normalized sums, respectively,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{and} \quad \frac{\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i))}{\sqrt{\sum_{i=1}^{n} Var(X_i)}} \tag{376}$$

behave when $n \to \infty$?

Here we will only consider so-called iid-sequences of rv's $X_i, i = 1, 2, \ldots$.

Definition: The rv's $X_1, X_2, \ldots$ are said to be **independently and identically (iid) distributed** if

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} P(X_i \leq x_i) \tag{377}$$

for all $n \geq 1$ and all n-tuples $(x_1, \ldots, x_n) \in \mathbb{R}^n$, and, additionally,

$$F_{X_i}(x) = F_X(x) \quad \forall i = 1, 2, \ldots \tag{378}$$

4.1 Chebyshev inequality and concepts of convergence

The following inequality provides a quite general upper bound on the deviation of an rv from its mean.

Theorem: Chebyshev inequality

Let X be a random variable with existing variance $\mathrm{Var}(X)$, then it holds

$$P(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2} \quad \forall \epsilon > 0 \tag{379}$$

Chebyshev inequality

Proof:

$$
\begin{aligned}
P\left(|X - \mathbb{E}(X)| \geq \epsilon\right) &= P(\overbrace{(X - \mathbb{E}(X))^2}^{Y} \geq \epsilon^2) \qquad (380)\\
&= \int_{\epsilon^2}^{\infty} 1 \cdot dF_Y(y) \leq \int_{\epsilon^2}^{\infty} \frac{y}{\epsilon^2} \cdot dF_Y(y)\\
&= \frac{1}{\epsilon^2} \cdot \int_{\epsilon^2}^{\infty} y \cdot F_Y(y) \leq \frac{1}{\epsilon^2} \int_{0}^{\infty} y \cdot dF_Y(y) = \frac{1}{\epsilon^2} \cdot \mathbb{E}(Y)\\
&= \frac{1}{\epsilon^2} \cdot \mathbb{E}(X - \mathbb{E}(X))^2 = \frac{1}{\epsilon^2} \cdot \mathrm{Var}(X)
\end{aligned}
$$

4.2 Convergence of sequences of rv's

The classical notion of convergence of sequences of numbers or functions as known from calculus is rather different from the convergence concepts employed in probability theory.

Weak convergence

This is due to the fact that the relation $\subseteq$ only provides a partial ordering on the underlying algebra $\mathscr{E}$, see Theorem 1 in Section 1.2. The following two notions of convergence are the most commonly used in probability theory.

Definition: The sequence of rv's $\{X_i\}$ on $[\Omega, \mathscr{E}, P]$

- converges **almost sure** (with probability 1) towards X, if

$$P(\{\omega \in \Omega : \lim_{n\to\infty} : X_n(\omega) = X(\omega)\}) = 1 \tag{381}$$

  Briefly: $X_n \xrightarrow[\text{a.s.}]{} X$

- converges **in probability** towards X, if

$$\lim_{n\to\infty} P(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \epsilon) = 0 \quad \forall \epsilon > 0 \tag{382}$$

  Briefly: $X_n \xrightarrow[\text{p}]{} X$

Convergence of sequences of rv's

Remark: Almost sure convergence means a "pointwise" convergence of the values $X_n(\omega)$ towards $X(\omega)$.

Note: Almost sure convergence implies convergence in probability:

$$X_n \xrightarrow[\text{a.s.}]{} X \implies X_n \xrightarrow[\text{p}]{} X \tag{383}$$

Corollary: Let $X_i, i = 1, 2, \ldots$ be a sequence of iid random variables with finite variance and denote

$$\mathbb{E}(X_i) = \mathbb{E}(X) = \mu \text{ and } \operatorname{Var}(X_i) = \operatorname{Var}(X) = \sigma^2.$$

Then it holds for the arithmetic means

$$\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

of this sequence

$$\mathbb{E}(\overline{X}_n) = \mu \; ; \; \operatorname{Var}(\overline{X}_n) = \frac{\sigma^2}{n} \tag{384}$$

Convergence of sequences of rv's

Proof: The proof of (385) is elementary:

$$\begin{aligned} \mathbb{E}(\overline{X}_n) &= \mathbb{E}\left(\frac{1}{n}\cdot\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\cdot\sum_{i=1}^{n}\mathbb{E}(X) \qquad (385)\\ &= \frac{1}{n}\cdot n\cdot\mathbb{E}(X) = \mathbb{E}(X) = \mu \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(\overline{X}_n) &= \mathrm{Var}\left(\frac{1}{n}\cdot\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\cdot\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \qquad (386)\\ &= \frac{1}{n^2}\cdot\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\cdot n\cdot\mathrm{Var}(X) \qquad (387)\\ &= \frac{\mathrm{Var}(X)}{n} = \frac{\sigma^2}{n} \end{aligned}$$

observing that $\mathrm{Var}\left(\sum X_i\right) = \sum \mathrm{Var}(X_i)$ due to independence of the X_i's.

4.3 Strong and Weak Laws of Large Numbers

Definition: The sequence of rv's $\{X_i\}_{i=1,2,..}$ follows the

- **weak law of large numbers**, if

$$\overline{X}_n \xrightarrow[\text{p}]{} \mathbb{E}(\overline{X}_n) \tag{388}$$

- **strong law of large numbers**, if

$$\overline{X}_n \xrightarrow[\text{a.s.}]{} \mathbb{E}(\overline{X}_n) \tag{389}$$

Theorem: Any sequence $\{X_i\}_{i=1,2,..}$ of iid variables with existing variance follows the weak law of large numbers, i.e.

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| \geq \epsilon\right) = 0 \quad \forall \epsilon > 0 \tag{390}$$

Law of large numbers

Proof: Applying Chebychev's inequality, we easily get

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| \geq \epsilon\right) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n \cdot \epsilon^2} \to 0 \;\; \textit{as } n \to \infty \tag{391}$$

The iid-assumption in the above theorem can be drastically relaxed, as stated in the following result.

Theorem (Markov): Let $\{X_i\}_{i=1,2..}$ be a sequence of rv's, not necessarily iid, but with finite variances, $\mathrm{Var}(X_i) < \infty \;\; \forall i$. Then the sequence follows the weak law of large numbers if

$$\lim_{n\to\infty} \mathrm{Var}(\overline{X}_n) = 0. \tag{392}$$

For the strong law of large numbers to hold, we need stronger assumptions, e.g. as stated in the following theorem (again, without proof).

Law of large numbers

Theorem (Kolmogorov): Let $\{X_i\}_{i=1,2..}$ be a sequence of totally independent rv's, not necessarily identically distributed. If the series

$$\sum_{i=1}^{\infty} \frac{\mathrm{Var}(X_i)}{i^2} \tag{393}$$

converges to a finite limit, then the series of rv's $\{X_i\}_{i=1,2..}$ follows the strong law of large numbers.

4.4 Important applications

4.4.1 Bernoulli Scheme

Suppose we are conducting n independent Bernoulli trials X_i; $i = 1, ..., n$; in which we record whether or not an event A has occurred whose probability $P(A) = p$ is constant throughout the trials. The distribution of the rv's X_i s is given by

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix} \tag{394}$$

Bernoulli Scheme

Such rv's X_i are called **Bernoulli variables**. Clearly, the rv's X_i are (totally) independent and identically distributed as

$$X_i \sim \text{Bi}(1, p) \tag{395}$$

with mean and variance given by

$$\mathbb{E}(X_i) = 0 \cdot (1-p) + 1 \cdot p = p \tag{396}$$

$$\text{Var}(X_i) = \mathbb{E}\left(X_i^2\right) - [\mathbb{E}(X_i)]^2 = p \cdot (1-p) \tag{397}$$

The sum of all X_i yields the absolute frequency of the occurrence of event A. We already know that

$$\sum_{i=1}^{n} X_i \sim \text{Bi}(n, p) \tag{398}$$

The relative frequency of the occurrence of A,

$$\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i, \tag{399}$$

then follows the weak law of large numbers (cp. eq. (391)), i.e.

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - p\right| \geq \epsilon\right) = 0 \quad \forall \epsilon > 0 \tag{400}$$

Thus, the relative frequency of successes in a Bernoulli scheme converges to the probability $p = P(A)$ of the occurrence of A. According to Kolmogorov's theorem above, even the strong law of large numbers holds for the relative frequency, i.e.

$$\overline{X}_n \xrightarrow[\text{a.s.}]{} p \tag{401}$$

since the series defined in (394) converges to a finite limit:

$$\sum_{i=1}^{\infty} \frac{\text{Var}(X_i)}{i^2} = p(1-p) \cdot \sum_{i=1}^{\infty} \frac{1}{i^2} = p(1-p) \cdot \frac{\pi^2}{6} \tag{402}$$

Applications of LLN: Histogram and ecdf

4.4.2 Histogram and Empirical cdf

The empirical cumulative distribution function (ecdf) of the data provides an estimate of the true underlying cdf F_X of the data.

Definition: (ecdf) Let $X_1, \ldots, X_n$ be i.i.d random variables. The empirical cdf is then given by

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i),\ x \in \mathbb{R}.$$

Thus it follows that the ecdf $\hat{F}_n$ almost surely converges to the cdf F_X, i.e.:

$$\forall x \in \mathbb{R} : \hat{F}_n(x) \xrightarrow[\text{a.s.}]{} F_X(x).$$

Accordingly, it holds: For continuously distributed (iid) data $X_1, ..., X_n$ the histogram converges a.s. to the pdf $f_X(x) = \frac{d}{dx} F_X(x)$.

4.4.3 Monte-Carlo-Integration

The computation of integrals of the form

$$I = \int_a^b g(x)\, dx, \qquad -\infty < a < b < \infty \tag{403}$$

can be easily done using realizations of random variables, so-called random numbers, following suitable probability distributions. Often, the uniform distribution on some interval $(a, b) \in \mathbb{R}$ with density

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{else} \end{cases} \tag{404}$$

is used for this purpose. With $X \sim \mathrm{U}(a, b)$ we may rewrite the above integral as

$$I = \int_a^b g(x)\, dx = (b-a) \cdot \int_a^b g(x) \cdot f_X(x)\, dx \tag{405}$$

Applications of LLN

The last integral, however, can be interpreted as an expectation

$$\int_a^b g(x)\, dx = (b-a) \cdot \mathbb{E}\left[g(X)\right] \tag{406}$$

with respect to $X \sim U(a,b)$. This gives rise to the following

Algorithm for Monte-Carlo-Integration

1. Generate n random numbers X_i with $X_i \sim U[a,b]$; i=1,...,n.
2. By the (weak) LLN we then have

$$\frac{(b-a)}{n} \cdot \sum_{i=1}^{n} g(X_i) \quad \underset{n\to\infty}{\longrightarrow} \quad \int_a^b g(x)\, dx \tag{407}$$

3. increase n to achieve a required level of accuracy of approximation.

Applications of LLN

Example (MC integration)

Let us compute the integral

$$I = \int_0^3 x\, e^{-x^2}\, dx \tag{408}$$

using Monte-Carlo-Integration on the basis of uniform random numbers.

Implementation in R:

```
> n = 1000000
> x = runif(n, 0, 3)
> I = 3*mean(x*exp(-x*x))
> I
[1] 0.49993 #approximate value of I
```

Importance Sampling

Improved accuracy can be achieved by adapting the integrand to a convenient pdf which better follows the curve than the uniform density: Importance Sampling

Example (MC Integration with non-uniform RNs)

Let us re-compute the previous integral using the much better adapted Weibull pdf with shape $\alpha = 2$ and scale $\beta = 1$:

$$f_X(x) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} \; exp\left(-(x/\beta)^\alpha\right) = 2x \; e^{-x^2}, \; x \geq 0 \tag{409}$$

Then we may rewrite the integral as

$$I = \int_0^3 x \; e^{-x^2} \; dx = 0.5 \int_0^3 f_X(x) \; dx = 0.5 \int_0^\infty g(x) \cdot f_X(x) \; dx \tag{410}$$

where $g(x) = I_{[0,3]}(x)$ is the indicator function.

Importance Sampling

Example (cont'd)

Observing that $I = 0.5 * \mathbb{E}\, g(X)$ with $X \sim \text{Weibull}(\alpha = 2, \beta = 1)$, we proceed as follows:

Implementation in R:

```
> n = 1000000
> x = rweibull(n, shape=2, scale=1)
> g= function(x){ ifelse(x<=3, 1, 0) }
> I = 0.5*mean(g(x))
> I
[1] 0.4999465 #approximate value of I
```

Success and effectiveness of Monte Carlo Methods rests on LLN!
For more applications of MC sampling see
https://towardsdatascience.com/monte-carlo-methods-decoded-d63301bde7ee

4.5 Central Limit Theorem

So far, we have stated results of the type

$$\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i \qquad \underset{n\to\infty}{\longrightarrow} \qquad \mathbb{E}(\overline{X}_n) \tag{411}$$

leading to convergence to a degenerate rv. What happens in case of weaker normalizations, e.g. with $\frac{1}{\sqrt{n}}$ instead of $\frac{1}{n}$? Will we then observe convergence to a "proper" (non-degenerate) distribution?

Central limit theorems indicate conditions under which a convergence towards the Gaussian distribution will occur. Thereby, **normalized sums** of the form

$$Z_n = \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}(X_i)}{\sqrt{\sum_{i=1}^{n} \mathrm{Var}(X_i)}} \tag{412}$$

are of interest. For these sums we indeed have:

$$\mathbb{E}(Z_n) = 0 \tag{413}$$
$$\mathrm{Var}(Z_n) = 1 \tag{414}$$

Let us start with a sequence of Gaussian rv's,

$$X_k \underset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2),\ k = 1, 2, \ldots \tag{415}$$

This immediately implies

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \tag{416}$$
$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{417}$$

After standardization we obtain

$$\frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \tag{418}$$

Equivalently, we have

$$Z_n = \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}(X_i)}{\sqrt{\sum_{i=1}^{n} \mathrm{Var}(X_i)}} = \frac{n \cdot (\overline{X}_n - \mu)}{n \cdot \sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \tag{419}$$

For sufficiently large n, the normalized sums Z_n remain to be normally distributed, even when the rv's X_k are no longer Gaussian.

Central Limit Theorem: For any sequence $\{X_k\}_{k=1,2,..}$ of iid variables with existing variance it holds

$$\lim_{n\to\infty} P(Z_n \leq z) = pnorm(z) \quad \forall z \in \mathbb{R}^1 \tag{420}$$

We briefly write:

$$Z_n \quad \underset{d}{\longrightarrow} \quad N(0,1) \tag{421}$$

and say: Z_n converges in distribution to the standard normal.

For iid variables we have from (418): $Z_n = \frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - \mu\right) =$ standardized sample mean.

Summarizing, for sequences of iid variables with existing mean μ and variance σ^2 we can restate the LLN and CLT as follows:

$$\overline{X}_n \quad \underset{p}{\longrightarrow} \quad \mu$$

LLN: Sample mean converges in probability to population mean

$$\frac{\sqrt{n}}{\sigma} \cdot \left(\overline{X}_n - \mu\right) \quad \underset{d}{\longrightarrow} \quad N(0,1)$$

CLT: Standardized sample mean converges to standard normal.

Remark: Similar limit theorems hold in case of weaker assumptions (e.g. for non-identical distributions or weakly dependent rv's X_k). Different types of limit theorems emerge for distributions with non-existing variances, where the limiting distribution is not Gaussian but some extreme-value-distribution.

We will now consider some important applications of the CLT.

4.5.1 Approximation of the binomial distribution

Consider a sequence of Bernoulli variables with (success) parameter $p \in (0,1)$ and their sum:

$$X_k \underset{\text{i.i.d.}}{\sim} \text{Bi}(1,p);\ k = 1,\ldots,n \longrightarrow \sum_{i=1}^{n} X_i \sim \text{Bi}(n,p) \tag{422}$$

According to the CLT above we then have

$$Z_n = \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \underset{\text{approx.}}{\sim} N(0,1) \tag{423}$$

which, in turn, implies

$$\sum_{i=1}^{n} X_i \underset{\text{approx.}}{\sim} N(np, np(1-p)) \tag{424}$$

Thus, it holds

$$P\left(\sum_{i=1}^{n} X_i \leq b\right) \approx pnorm\left(\frac{b-np}{\sqrt{np(1-p)}}\right), \ i.e. \tag{425}$$

$$\text{Bi}(n,p) \underset{\text{approx.}}{\sim} N(np, np(1-p)) \tag{426}$$

Binomial Approximation

The accuracy of the approximation can be improved by a continuity correction of 0.5:

$$P\left(a \leq \sum_{i=1}^{n} X_i \leq b\right) \approx pnorm\left(\frac{b+0.5-np}{\sqrt{np(1-p)}}\right) - pnorm\left(\frac{a-0.5-np}{\sqrt{np(1-p)}}\right) \tag{427}$$

For this approximation to apply properly it is recommended to observe the following "rule of thumb":

$$n \cdot p \cdot (1-p) > 9 \tag{428}$$

Apply this to compute (Covid) disease probabilities in a town of about 50000 inhabitants assuming a disease percentage of about 5%.

4.5.2 Approximation of the Poisson distribution

This approximation can be accomplished in a similar way as for the binomial distribution, which then yields:

$$X \sim \text{Po}(\lambda) \underset{\text{approx.}}{\sim} N(\lambda, \lambda) \tag{429}$$

observing that the Poisson distribution has the unique property

$$\mathbb{E}(X) = \text{Var}(X) = \lambda.$$

Again, the approximation can be improved by a continuity correction as follows:

$$P\left(a \leq \sum_{i=1}^{n} X_i \leq b\right) \approx pnorm\left(\frac{b+0.5-\lambda}{\sqrt{\lambda}}\right) - pnorm\left(\frac{a-0.5-\lambda}{\sqrt{\lambda}}\right)$$

Poisson Approximation

For this approximation to work properly, a similar rule of thumb should be obeyed: $\lambda > 9$.

Remark: We applied such an approximation in our recent paper on predicting food and beverage demands in staff canteens and restaurants, see
https://www.sciencedirect.com/science/article/pii/S0169207021001011?ref=pdf_download&fr=RR-2&rr=82e2ec010ab2c270

General approximation principle: For cdf's F_X with existing variance the approximation by a Gaussian distribution proceeds such that

$$F_X \approx N(\mathbb{E}(X), \mathrm{Var}(X)).$$

5. Multivariate Distributions

5.1 Bivariate Distributions

Let be given two random variables X_1 und X_2. We consider two-dimensional random vectors $\underline{X} = (X_1, X_2)^T$, defined as measurable function

$$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} : \Omega \to \mathbb{R}^2 \tag{430}$$

Example (Examination Results)

We are given the results of exams in Linear Algebra (X_1) and in Calculus (X_2). The following table displays the number of students along with their respective marks. The marks in Calculus appear horizontally and those in Linear Algebra vertically.
The values (marks) are denoted by $x_{11}, \ldots, x_{15}$ and $x_{21}, \ldots, x_{25}$, respectively.

Example (cont'd)

	$x_{21}=1$	$x_{22}=2$	$x_{23}=3$	$x_{24}=4$	$x_{25}=5$
$x_{11}=1$	3	4	2	1	0
$x_{12}=2$	6	9	17	3	1
$x_{13}=3$	2	7	11	5	2
$x_{14}=4$	1	3	8	3	1
$x_{15}=5$	0	0	4	6	1

We define probabilities as relative frequencies

$$p_{ij} = \frac{H_{ij}}{n} \tag{431}$$

where $n = 100$ means the total number of examinees and H_{ij} the number of students with marks i and j in the respective subjects; $i, j = 1, ..., 5$.

Bivariate Distributions

Again, we distinguish between **discrete** and **continuous** random vectors. Discrete random vectors (drv's) are defined on a discrete (finite or countably infinite) set and can be represented as a table

	x_{21}	x_{22}	$\dots$	x_{2m}
x_{11}	p_{11}	p_{12}	$\dots$	p_{1m}
x_{12}	p_{21}	p_{22}	$\dots$	p_{2m}
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$
x_{1k}	p_{k1}	p_{k2}	$\dots$	p_{km}

in which the pairs of values (x_{1i}, x_{2j}) are arranged jointly with their respective probabilities p_{ij}. The probabilities are non-negative and normalized such that

$$p_{ij} \geq 0 \quad ; \quad \sum_i \sum_j p_{ij} = 1. \tag{432}$$

where $m, k = \infty$ in case of countably infinite sets of values.

Bivariate Distributions

Continuous random vectors (crv's) $\underline{X}$ have a continuous range of values (Wertebereich) and a pdf $f_{\underline{X}}$ with the properties

- Non-negativity: $f_{\underline{X}}(x_1, x_2) \geq 0$ for all $(x_1, x_2) \in \mathbb{R}^2$
- Normalization: $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\underline{X}}(x_1, x_2)\, dx_2\, dx_1 = 1$

Example (Testing cars)

In a test of cars, their fuel consumption X_1 and acceleration X_2 are recorded. Suppose that $\underline{X} = (X_1, X_2)^T$ has a continuous distribution, with realizations

$$x_1 \in [3, 8], x_2 \in [4, 11]. \tag{433}$$

Assuming that the realizations $\underline{X}$ are equiprobable, corresponding to a uniform distribution over the rectangle $[3, 8] \times [4, 11]$ with area $5x7 = 35$, we thus have the pdf

Example (cont'd)

$$f_{\underline{X}}(x_1, x_2) = \begin{cases} \frac{1}{35} & \text{if} 3 \leq x_1 \leq 8, \quad 4 \leq x_2 \leq 11 \\ 0 & \text{else} \end{cases} \tag{434}$$

Definition: The **cdf of a bivariate random vector** $\underline{X}$ is defined as

$$F_{\underline{X}}(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2) \tag{435}$$

where it holds:

$$F_{\underline{X}}(x_1, x_2) = \begin{cases} \sum\limits_{x_{1i} \leq x_1} \sum\limits_{x_{2j} \leq x_2} P(X_1 = x_{1i}, X_2 = x_{2j}) & \text{if } \underline{X} \text{ is discrete} \\ \int\limits_{-\infty}^{x_1} \int\limits_{-\infty}^{x_2} f_{\underline{X}}(t_1, t_2)\, dt_2\, dt_1 & \text{if } \underline{X} \text{ is continuous} \end{cases} \tag{436}$$

Example (Results of exams, cont'd)

First, we divide the absolute frequencies by $n = 100$, so that we now the table displays the probabilities:

	1	2	3	4	5
1	0.03	0.04	0.02	0.01	0.00
2	0.06	0.09	0.17	0.03	0.01
3	0.02	0.07	0.11	0.05	0.02
4	0.01	0.03	0.08	0.03	0.01
5	0.00	0.00	0.04	0.06	0.01

- The probability that mark 2, or even better, in Linear Algebra has been achieved is given by

$$P(X_1 \leq 2, X_2 \leq 5) = \sum_{X_{1i} \in \{1,2\}} \sum_{X_{2j} \in \{1,\ldots,5\}} p_{ij} \qquad (437)$$

$$= 0.46$$

Example (cont'd)

$$= \quad F_{\underline{X}}(2,5)$$

This result corresponds to the sum of probabilities in the upper two rows of the table, which have been marked in blue.

- The probability of achieving mark 2 or better in Calculus reads

$$\begin{aligned} P(X_1 \leq 5, X_2 \leq 2) &= \sum_{x_{1i} \in \{1,\ldots,5\}} \sum_{x_{2j} \in \{1,2\}} p_{ij} \qquad (438) \\ &= 0.35 \\ &= F_{\underline{X}}(5,2). \end{aligned}$$

- The probability of having at least mark 2 in both subjects reads

$$P(X_1 \leq 2, X_2 \leq 2) = \sum_{x_{1i} \in \{1,2\}} \sum_{x_{2j} \in \{1,2\}} p_{ij} \qquad (439)$$

Example (cont'd)

$$\begin{aligned} &= 0.22 \\ &= F_{\underline{X}}(2,2). \end{aligned}$$

Example (Test of cars, cont'd)

With the chosen pdf (density of uniform distribution)

$$f_{\underline{X}}(x_1,x_2) = \begin{cases} \frac{1}{35} & \text{if } 3 \le x_1 \le 8,\ 4 \le x_2 \le 11 \\ 0 & \text{else} \end{cases} \tag{440}$$

we now calculate the probability that a car (chosen arbitrarily) does not need more than 6 seconds to accelerate from 0 to 100 km/h and still needs less than 6 liters of fuel per 100 km:

$$P(X_1 \le 6, X_2 \le 6) = F_{\underline{X}}(6,6) \tag{441}$$

Example (cont'd)

$$= \int_3^6 \int_4^6 \frac{1}{35}\, dx_2\, dx_1 = \int_3^6 \frac{x_2}{35}\Big|_4^6\, dx_1$$

$$= \int_3^6 \frac{2}{35}\, dx_1 = \frac{2x_1}{35}\Big|_3^6$$

$$= \frac{6}{35} = 0.171.$$

Since we have based the probability calculations on a uniform distribution, this result is identical with the geometric probability

$$\text{favourable area} \,/\, \text{possible area}$$

$$= \lambda([3,6] \times [4,6]) / \lambda([3,8] \times [4,11]) = 6/35.$$

Marginal Distributions

Corollary: Relationship between pdf, probabilities and cdf

$$f_{\underline{X}}(x_1, x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F_{\underline{X}}(x_1, x_2), \ \forall (x_1, x_2) \in \mathbb{R}^2 \quad (442)$$

$$P((X_1, X_2) \in A) = \iint_A f_{\underline{X}}(x_1, x_2) \ dx_2 \ dx_1, \ \forall A \subseteq \mathbb{R}^2 \quad (443)$$

i.e. the pdf f_X is the mixed partial derivative of the cdf F_X, which, in turn, represents the antiderivative (Stammfunktion) of the pdf.

5.2 Marginal Distributions

The unconditional (total) distributions of X_1 and X_2, respectively, are called the marginal distributions.

Definition: The function

$$\begin{aligned} F_{X_1}(x_1) &= P(X_1 \leq x_1, \ -\infty < X_2 < \infty) \quad (444) \\ &= F_{\underline{X}}(x_1, \infty), \ \forall x_1 \in \mathbb{R}^1 \end{aligned}$$

Marginal Distributions

is called the cdf of the **marginal distribution of** X_1. Analogously, the function

$$\begin{aligned} F_{X_2}(x_2) &= P(-\infty < X_1 < \infty,\ X_2 \leq x_2) \\ &= F_{\underline{X}}(\infty, x_2),\ \forall x_2 \in \mathbb{R}^1 \end{aligned} \tag{445}$$

is called the cdf of the **marginal distribution of** X_2.

Corollary: For the marginal distributions of a discrete random vector $\underline{X}$ it holds

$$X_1 \sim \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ \sum_{j=1}^{m} p_{1j} & \sum_{j=1}^{m} p_{2j} & \cdots & \sum_{j=1}^{m} p_{kj} \end{pmatrix} \tag{446}$$

$$X_2 \sim \begin{pmatrix} x_{21} & x_{22} & \cdots & x_{2m} \\ \sum_{i=1}^{k} p_{i1} & \sum_{i=1}^{k} p_{i2} & \cdots & \sum_{i=1}^{k} p_{im} \end{pmatrix} \tag{447}$$

Marginal Distributions

The **marginal probabilities** of X_1 are the **row sums (Zeilensummen)** and those of X_2 are the **column sums (Spaltensummen)** in the table of the joint probabilities p_{ij} of $\underline{X}$.

Example (Results of Exams, Testing cars)

In our example concerning the results of exams in Linear Algebra (X_1) and Calculus (X_2) the marginal distributions are given by

$$X_1 \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0.10 & 0.36 & 0.27 & 0.16 & 0.11 \end{pmatrix} \tag{448}$$

$$X_2 \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0.12 & 0.23 & 0.42 & 0.18 & 0.05 \end{pmatrix} \tag{449}$$

In the example of testing cars we obtain the marginal distribution of X_1 as

$$F_{X_1}(x_1) = F_{\underline{X}}(x_1, \infty), \text{ i.e.}$$

Example (cont'd)

$$F_{X_1}(x_1) = \int_{-\infty}^{x_1} \underbrace{\left(\int_{-\infty}^{\infty} f_{\underline{X}}(t_1, t_2)\, dt_2 \right)}_{\text{marginal density of of } X_1} dt_1 \tag{450}$$

Analogously, we get the marginal distribution of X_2 from the corresponding marginal density. The marginal densities read

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{\underline{X}}(x_1, x_2)\, dx_2 \tag{451}$$

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{\underline{X}}(x_1, x_2)\, dx_1 \tag{452}$$

Conditional Distributions

Note: Marginal densities are obtained by "integrating out" the (nuisance) variable which is not of current interest in the joint pdf.

5.3 Conditional distributions

5.3.1 Discrete conditional distributions

Let $\underline{X} = (X_1, X_2)^T$ be a discrete rv and assume

$$\text{Condition } C: \; X_1 = x_{1i}$$

to hold. This fixes the variable X_1; we then focus on the distribution of the variable X_2. We call this the **conditional distribution** of X_2 under the condition $C: \; X_1 = x_{1i}$, and briefly write this as

$$X_2 | X_1 = x_{1i} \sim \begin{pmatrix} x_{21} & x_{22} & x_{23} & \dots & x_{2m} \\ q_1 & q_2 & q_3 & \dots & q_m \end{pmatrix} \tag{453}$$

with conditional probabilities defined in the usual way,

$$q_j = \frac{P\left(X_1 = x_{1i}\,,\; X_2 = x_{2j}\right)}{P\left(X_1 = x_{1i}\right)} \tag{454}$$

Conditional Distributions

This, however, means that

$$q_j = \frac{p_{ij}}{p_{i.}}, \quad j = 1, \ldots, m, \text{ where}$$

$$p_{i.} = \sum_{l=1}^{m} p_{il} \text{ denotes the } i\text{-th row sum, } i = 1, \ldots, k.$$

Correspondingly, we arrive at the distribution of X_1 for given fixed value $X_2 = x_{2j}$,

$$X_1 | X_2 = x_{2j} \sim \begin{pmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1k} \\ s_1 & s_2 & s_3 & \ldots & s_k \end{pmatrix} \tag{455}$$

with conditional probabilities

$$\begin{aligned} s_i &= \frac{P\left(X_1 = x_{1i}\ ,\ X_2 = x_{2j}\right)}{P\left(X_2 = x_{2j}\right)} \\ &= \frac{p_{ij}}{p_{.j}}, \quad i = 1, \ldots, k \end{aligned} \tag{456}$$

Here, $p_{\cdot j} = \sum_{l=1}^{k} p_{lj}$ denotes the j-th column sum, $j = 1, \ldots, m$.

Example (Results of Exams in Linear Algebra and Calculus)

We obtain with

$$X_1 | X_2 = 2 \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ \frac{0.04}{0.23} & \frac{0.09}{0.23} & \frac{0.07}{0.23} & \frac{0.03}{0.23} & \frac{0}{0.23} \end{pmatrix} \tag{457}$$

the distribution of the marks in Linear Algebra among all those students whose exam in Calculus ended with mark 2:

$$X_1 | X_2 = 2 \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0.1739 & 0.3913 & 0.3044 & 0.1304 & 0 \end{pmatrix} \tag{458}$$

(rounded up to four decimal places). For the calculation of these probabilities we followed the rule: Divide the entries in the 2nd column by the column sum 0.23.

Example (cont'd)

In an analogous way we obtain the distribution of the marks in Calculus among all those students whose exam in Linear Algebra was evaluate with mark 3:

$$X_2|X_1 = 3 \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ \frac{0.02}{0.27} & \frac{0.07}{0.27} & \frac{0.11}{0.27} & \frac{0.05}{0.27} & \frac{0.02}{0.27} \end{pmatrix} \tag{459}$$

according to the computation rule: Divide the entries in the 3rd row by the row sum 0.27. After rounding to four decimal places we get the distribution

$$X_2|X_1 = 3 \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0.0741 & 0.2592 & 0.4074 & 0.1852 & 0.0741 \end{pmatrix} \tag{460}$$

5.3.2 Continuous conditional distributions

In case of a continuous random vector $\underline{X} = (X_1, X_2)^T$ with pdf $f_{\underline{X}}$ we replace the above condition C by the interval

$$B = (x_1, x_1 + h], \ h > 0 \tag{461}$$

and define the conditional distribution via the limiting approach $h \to 0^+$.

Definition: Let be $x_1 \in \mathbb{R}^1$ such that $f_{X_1}(x_1) > 0$.Then the limiting distribution,

$$F_{X_2|X_1=x_1}(x_2) = \lim_{h \to 0^+} P\left(X_2 \leq x_2 | x_1 < X_1 \leq x_1 + h\right), \forall x_2 \in \mathbb{R}^1 \tag{462}$$

is called the **conditional cdf** of X_2 for given realization $X_1 = x_1$.

Taking the above limit, we arrive at the following result

Conditional Distributions

Corollary: The conditional cdf of X_2 for given value $X_1 = x_1$ reads

$$F_{X_2|X_1=x_1}(x_2) = \int_{-\infty}^{x_2} \frac{f_{\underline{X}}(x_1, t_2)}{f_{X_1}(x_1)}\, dt_2 \tag{463}$$

The function under the integral sign is thus the derivative of the conditional cdf $F_{X_2|X_1=x_1}(\cdot)$ and is called the conditional pdf corresponding to this cdf. It is easily recognizable that this density is the ratio of the joint density of X_1 and X_2 and the marginal density of X_1, in full analogy to the situation with conditional probabilities.

Definition: The **conditional pdf's** of the continuous random vector $\underline{X} = (X_1, X_2)^T$ are defined as

$$f_{X_2|X_1=x_1}(x_2) = \frac{f_{(X_1,X_2)}(x_1, x_2)}{f_{X_1}(x_1)} \tag{464}$$

$$f_{X_1|X_2=x_2}(x_1) = \frac{f_{(X_1,X_2)}(x_1, x_2)}{f_{X_2}(x_2)} \tag{465}$$

Conditional Densities

Remarks: These relations are used for simulating bivariate distributions via conditional and marginal densities. In an analogous manner, we can represent three-dimensional distributions on the basis of unidimensional conditional and marginal distributions:

$$\begin{aligned} f_{(X_1,X_2,X_3)}(x_1,x_2,x_3) &= f_{(X_1,X_2)|X_3=x_3}(x_1,x_2)\, f_{X_3}(x_3) \\ &= f_{X_1|X_2=x_2,X_3=x_3}(x_1)\, f_{X_2|X_3=x_3}(x_2)\, f_{X_3}(x_3) \end{aligned} \tag{466}$$

This is used e.g. in 3D image processing (reconstruction on the basis of two-dimensional projections). Further, even high-dimensional distributions can be generated from unidimensional conditional and marginal densities. This method has become known as **Gibbs-Sampling**; it is widely used in **Bayesian Statistics**.

Conditional Densities

From (463) and (464) we immediately get

$$f_{X_1|X_2=x_2}(x_1) = \frac{f_{X_2|X_1=x_1}(x_2)\, f_{X_1}(x_1)}{f_{X_2}(x_2)} \tag{467}$$

Furtheron, we arrive at the continuous analogue to Bayes' Theorem.

Corollary: Bayes's formula for pdf's

$$f_{X_1|X_2=x_2}(x_1) = \frac{f_{X_2|X_1=x_1}(x_2)\, f_{X_1}(x_1)}{\int_{-\infty}^{\infty} f_{X_2|X_1=x_1}(x_2)\, f_{X_1}(x_1)\, dx_1} \tag{468}$$

One out of many applications:
Variational Autoencoder, see e.g.
`https://python.plainenglish.io/variational-autoencoder-1eb543f5f055`

Bayes's theorem and posterior densities

Bayesian Statistics: Data X and parameters θ of F_X are both considered as random!
Of interest is then the so-called **posterior density** of θ given the data $X = x$, which, according to Bayes's theorem, reads

$$f(\theta|x) = \frac{f(x|\theta)\, p(\theta)}{\int_{-\infty}^{\infty} f(x|\theta)\, p(\theta)\, d\theta} \tag{469}$$

where $p(\theta)$ stands for the **prior density** of the parameter summarizing our knowledge about θ before having seen the data x. The density $f(x|\theta)$ represents the sample density, usually called the **likelihood function**. The posterior density combines our prior knowledge about the parameter and the data. The change from the prior to the posterior reflects the degree of **Bayesian learning** about the parameter.

Beta-Binomial Conjugacy

Example (Modeling rare events)

Imagine you want to test whether the production cycle of a highly reliable item still meets the strict reliability requirements demanded by the customers (e.g. 21 ppm for most of Infineon's chips). Assume, during a burn-in-test you have observed no failure among 70000 chips.

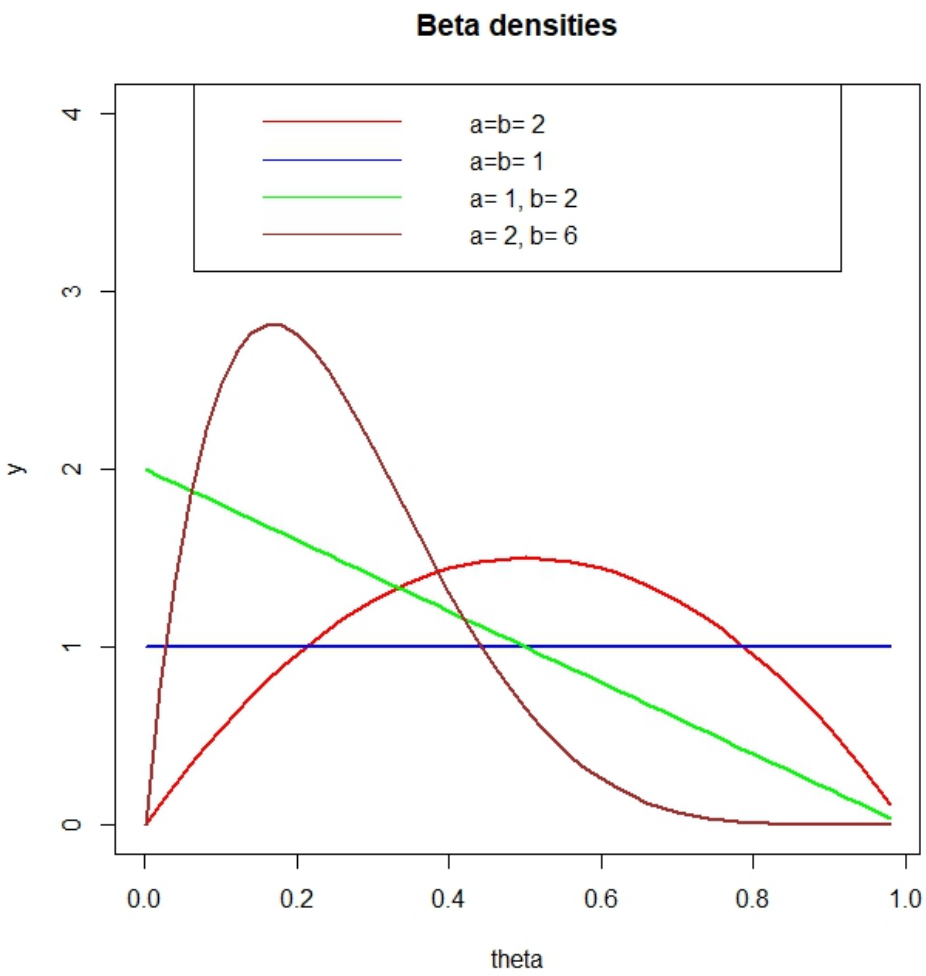
Would you dare claiming that the complete population has zero-failure probability?
More general: how to cope with rare event modelling?

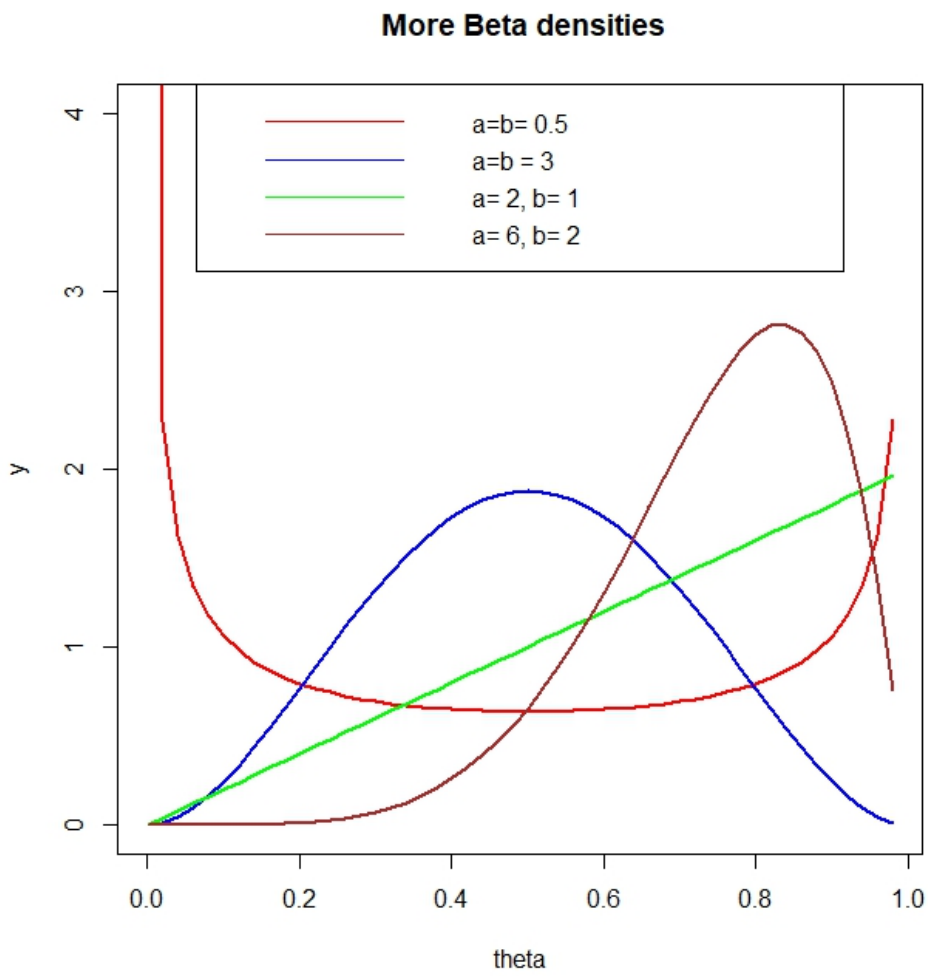
In our case we have: $X = \sum_{i=1}^{n} X_i \sim \text{Bi}(n, \theta = p)$ with $n = 70000$ and unknown failure probability $\theta = p$. The sample estimate is just $\hat{\theta} = \hat{p} = 0$ with a nonsensical confidence interval of length zero!

Bayesian remedy: Model the uncertainty about $\theta = p$ via a continuous **Beta distribution** on the interval [0, 1] with pdf

$$f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, \theta^{a-1}(1-\theta)^{b-1}\, I_{[0,1]}(\theta)$$

Densities of Beta Distribution





Posterior distribution

Example (cont'd)

Prior information: failure rates in more than 80% of comparable samples range from $1ppm \ldots 25ppm$ with an average failure rate of $10ppm$.

Encoding of prior information: $\theta \sim \text{Beta}(a, b)$; $a = 1, b = 100000$

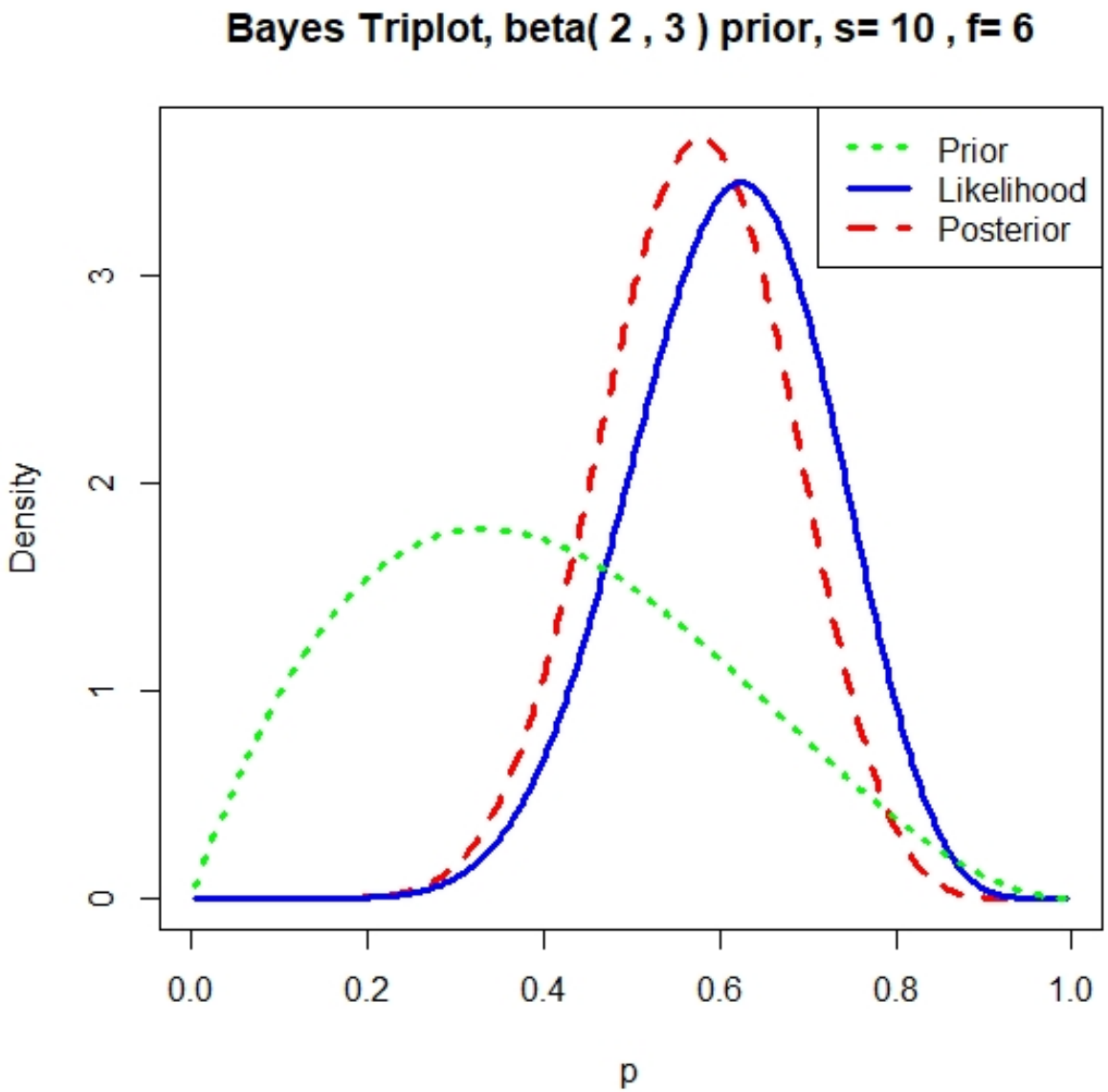
$$E\theta = \frac{a}{a+b} = \frac{1}{100001} = 0.00001$$

$$P(0.000001 < \theta < 0.000025) = 0.823$$

$$P(\theta < 0.000021) = 0.88$$

```
> library(LearnBayes)
> ?triplot
```

Triplot in LearnBayes



Example (cont'd)

Posterior distribution?

The posterior distribution of θ is also Beta (since Beta prior is conjugate to the binomial distribution).

$$\Rightarrow \ \theta|x=0 \sim \text{Beta}(a+\sum x_i, n+b-\sum x_i) = \text{Beta}(1, 170000)$$

The estimated failure rate thus becomes about $6ppm$:

$$\mathbb{E}(\theta|x=0) = \frac{1}{170001} = 0.000006$$

Moreover,

$$P(\theta < 0.000021|x=0) = pbeta(0.000021, 1, 170000) = 0.972$$

5.3.3 Conditional Expectation, Conditional Variance

Definition: The expectation vector $\mathbb{E}(\underline{X})$ of $\underline{X} = (X_1, X_2)^T$ is defined component-wise as:

$$\mathbb{E}(\underline{X}) := (\mathbb{E}(X_1), \mathbb{E}(X_2))^T \text{ with:}$$

$$\mathbb{E}(X_1) = \begin{cases} \int_{-\infty}^{\infty} x_1 \, f_{X_1}(x_1) \, dx_1 & \text{if } \underline{X} \text{ is continuous} \\ \sum_{i=1}^{k} x_{1i} \, P(X_1 = x_{1i}) & \text{if } \underline{X} \text{ is discrete} \end{cases}$$

$$\mathbb{E}(X_2) = \begin{cases} \int_{-\infty}^{\infty} x_2 \, f_{X_2}(x_2) \, dx_2 & \text{for continuous } \underline{X} \\ \sum_{j=1}^{m} x_{2j} \, P(X_2 = x_{2j}) & \text{for discrete } \underline{X} \end{cases}$$

This means: $\mathbb{E}(X_j) =$ Expectation of the marginal distribution of X_j; $j = 1, 2$

Quite analogously, we define conditional expectations:

Conditional Expectation

Definition: For continuous $\underline{X}$ we define

$$\mathbb{E}(X_2|X_1 = x_1) = \int_{-\infty}^{\infty} x_2 \; f_{X_2|X_1=x_1}(x_2) \; dx_2$$

$$\mathbb{E}(X_1|X_2 = x_2) = \int_{-\infty}^{\infty} x_1 \; f_{X_1|X_2=x_2}(x_1) \; dx_1$$

For discrete $\underline{X}$ we define

$$\mathbb{E}(X_2|X_1 = x_{1i}) = \sum_{j=1}^{m} x_{2j} \; P(X_2 = x_{2j}|X_1 = x_{1i}) = \frac{1}{p_{i.}} \sum_{j=1}^{m} x_{2j} p_{ij}$$

$$\mathbb{E}(X_1|X_2 = x_{2j}) = \sum_{i=1}^{k} x_{1i} \; P(X_1 = x_{1ij}|X_2 = x_{2j}) = \frac{1}{p_{.j}} \sum_{i=1}^{k} x_{1i} p_{ij}$$

Example (Results of exams, cont'd)

	1	2	3	4	5
1	0.03	0.04	0.02	0.01	0.00
2	0.06	0.09	0.17	0.03	0.01
3	0.02	0.07	0.11	0.05	0.02
4	0.01	0.03	0.08	0.03	0.01
5	0.00	0.00	0.04	0.06	0.01

$$X_1 \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0.10 & 0.36 & 0.27 & 0.16 & 0.11 \end{pmatrix}$$

$$X_2 \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0.12 & 0.23 & 0.42 & 0.18 & 0.05 \end{pmatrix}$$

$$\mathbb{E}(X_1|X_2 = 2) = \frac{1}{0.23} \sum_{i=1}^{5} x_{1i}\, P(X_1 = x_{1i}, X_2 = 2)$$

$$= \frac{1}{0.23}(1 \cdot 0.04 + 2 \cdot 0.09 + 3 \cdot 0.07 + 4 \cdot 0.03 + 5 \cdot 0)$$

Example (cont'd)

$$= \frac{0.04+0.18+0.21+0.12}{0.23} = \frac{0.55}{0.23} = 2.39$$

The unconditional (marginal) expectation of X_1, however, reads

$$\mathbb{E}(X_1) = 1 \cdot 0.1 + 2 \cdot 0.36 + 3 \cdot 0.27 + 4 \cdot 0.16 + 5 \cdot 0.11 = 2.82$$

In the same way as we have extended the concept of expectation to marginal and conditional expectations, we can introduce marginal and conditional variances.

Definition: Variances for continuous $\underline{X}$

1. $\mathrm{Var}(X_i) := \mathbb{E}(X_i - \mathbb{E}X_i)^2 = \int_{-\infty}^{\infty} (x_i - \mathbb{E}X_i)^2 \, f_{X_i}(x_i) \, dx_i \; ; i = 1, 2$
2. $\mathrm{Var}(X_2|X_1 = x_1) = \int_{-\infty}^{\infty} (x_2 - \mathbb{E}(X_2|X_1 = x_1))^2 \, f_{X_2|X_1=x_1}(x_2) \, dx_2$

Example (Bivariate normal distribution)

The random vector $\underline{X}$ is said to follow a **bivariate normal distribution**, briefly

$$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\underline{\mu}, \Sigma\right) \tag{470}$$

with parameters

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \end{pmatrix} \tag{471}$$

and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \; ; \; \sigma_i^2 = Var(X_i); i = 1, 2; \tag{472}$$

if $\underline{X}$ has pdf

Bivariate normal distribution

Example (cont'd)

$$f_{\underline{X}}(\underline{x}) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1}(\underline{x}-\underline{\mu})\right), \underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$$

Here, ρ is the so-called correlation coefficient (cp. Section 5.5). Observing that

$$\det(\Sigma) = \sigma_1^2\sigma_2^2(1-\rho^2) \quad , \quad \Sigma^{-1} = \frac{1}{\det(\Sigma)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

we obtain the pdf in a more explicit form

$$f_{\underline{X}}(t_1, t_2) = \frac{1}{2\pi\sigma_1\sigma_2\tau} \exp\left(-\frac{1}{2\tau^2}\left[t_1^2 - 2\rho t_1 t_2 + t_2^2\right]\right)$$

where $\tau^2 = 1-\rho^2$, and $t_i = (x_i - \mu_i)/\sigma_i (i = 1, 2)$.

Bivariate normal distribution

Example (cont'd)

From the analytical form of the pdf we easily recognize that it has elliptical contours. The marginal distributions of $\underline{X}$ are unidimensional normal distributions:

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-(x_i - \mu_i)^2/2\sigma_i^2) \quad i = 1, 2$$

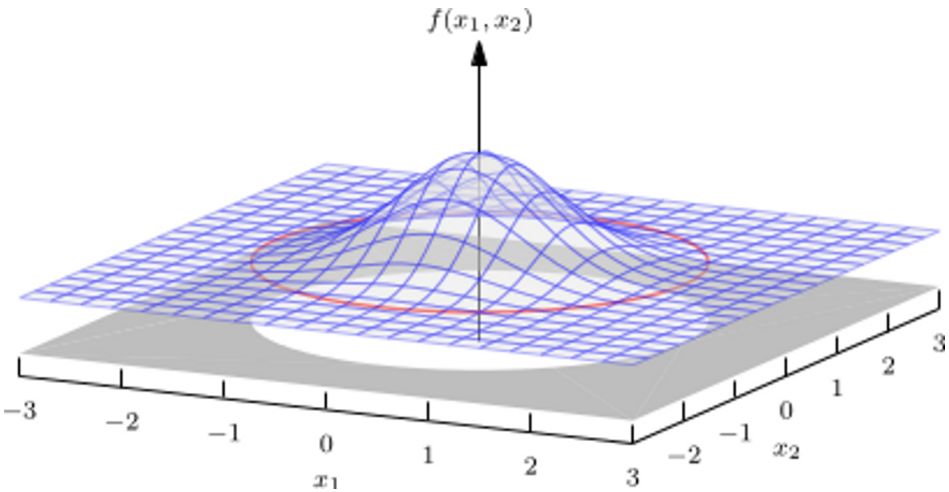
i.e. $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$. For the conditional density $f_{X_2|X_1=x_1}$ it holds:

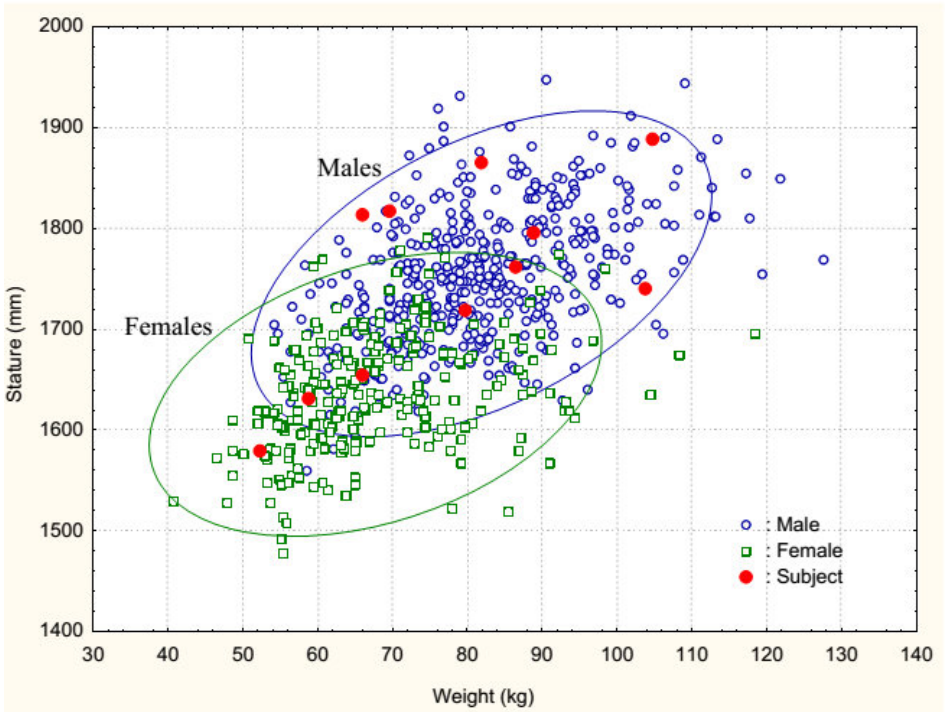
$$f_{X_2|X_1=x_1}(x_2) = \frac{1}{\sqrt{2\pi}\sigma_{2|1}} \exp(-(x_2 - \mu_{2|1})^2/2\sigma_{2|1}^2), \text{ where}$$

$$\mu_{2|1} = \mathbb{E}(X_2|X_1 = x_1) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1)$$

$$\sigma_{2|1}^2 = \mathrm{Var}(X_2|X_1 = x_1) = \sigma_2^2(1 - \rho^2)$$

Bivariate normal distribution

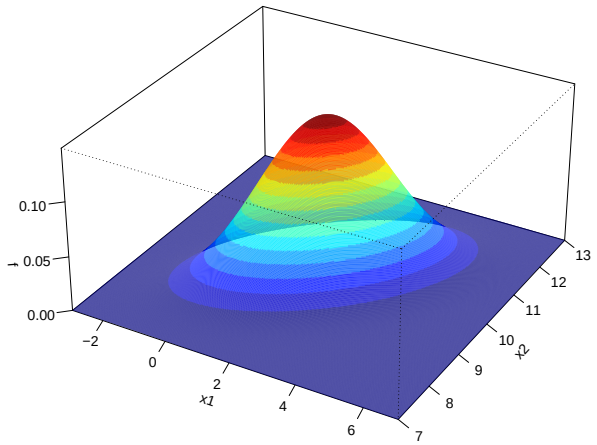


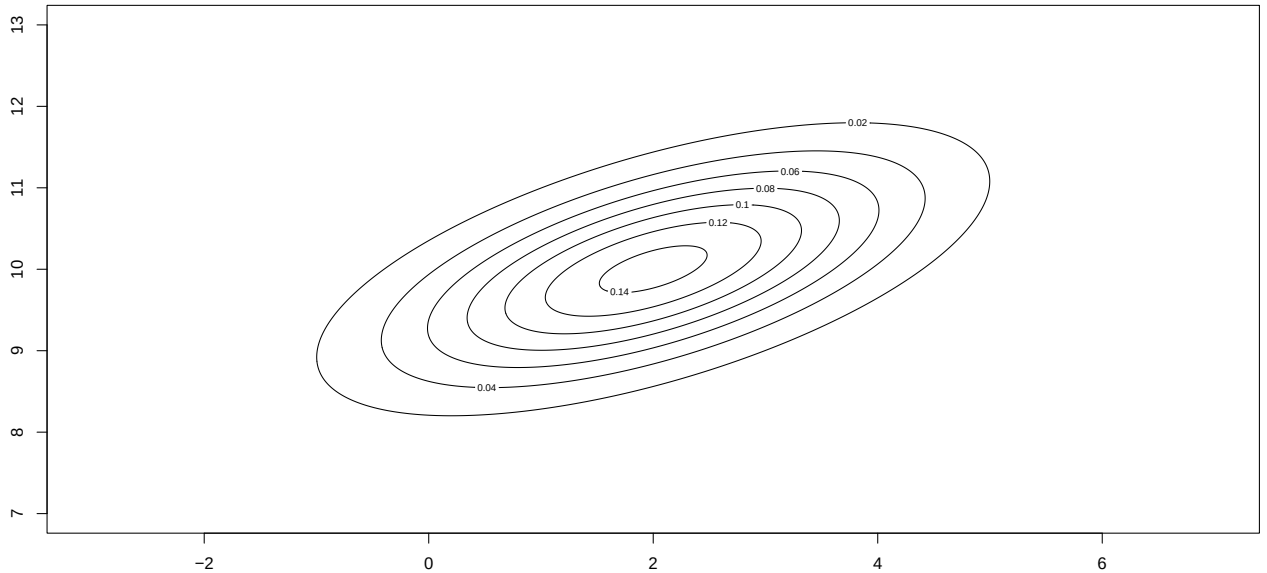


Implementation in R

```
> # Input
> m1 = 2 #mean of X1
> m2 = 10 #mean of X2
> s1 = 1.5 #standard deviation of X1
> s2 = 0.9 #standard deviation of X2
> r = 0.6 #correlation between X1 and X2
> # 3D and contour plot arrangements
> x1 = seq(m1-5,m1+5,length= 500)
> x2 = seq(m2-3,m2+3,length= 500)
> z= function(x1,x2){z=exp(-(((x1-m1)*(x1-m1)/(s1*
s1))+((x2-m2)*(x2-m2)/(s2*s2))-2*r*(x1-m1)*(x2-m2)/
(s1*s2))/(2*(1-r*r)))/(2*pi*s1*s2*sqrt(1-r*r))}
> f = outer(x1,x2,z)
> persp3D(x1,x2,f,theta=30,phi=30,expand=0.5)
> contour(x=x1, y=x2, z=f)
```

Bivariate normal distribution in R





5.4 Independence of random variables

Definition: Let $\underline{X} = (X_1, X_2)^T$ be a random vector, then X_1 and X_2 are said to be **independent** iff

$$P(a \leq X_1 \leq b, c \leq X_2 \leq d) = P(a \leq X_1 \leq b) \cdot P(c \leq X_2 \leq d) \quad (473)$$

$$\forall a, b, c, d \in \mathbb{R}.$$

Corollary: X_1 and X_2 are independent if and only if

$$F_{(X_1,X_2)}(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2) \quad \forall x_1, x_2 \in \mathbb{R} \quad (474)$$

Additionally, it then holds

$$F_{X_2|X_1=x_1}(x_2) = F_{X_2}(x_2) \quad \forall x_1, x_2 \in \mathbb{R} \quad (475)$$

Note: If X is continuous then the assertions of the corollary also hold for the corresponding pdf's:

Independence of rv's

$$f_{(X_1,X_2)}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) \tag{476}$$

where $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ are the corresponding marginal pdf's.

We will now consider measures for various degrees of dependence. A very common measure, which is but often misused, is the so-called **correlation**.

Definition: Let X_1 and X_2 be random variables with joint cdf $F(x_1, x_2)$, then we call

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))\,(X_2 - \mathbb{E}(X_2))\right] \tag{477}$$

the **covariance** between X_1 and X_2.

Corollary: In case of a continuous $\underline{X}$ we thus compute the covariance via a (double) integral and in the discrete case via a (double) sum,

Covariance

respectively, as follows

$$\mathrm{Cov}(X_1, X_2) = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x_1 - \mathbb{E}(X_1)]\,[x_2 - \mathbb{E}(X_2)]\; f_{(X_1,X_2)}(x_1, x_2)\; dx_2 dx_1 \\ \sum_i \sum_j (x_{1i} - \mathbb{E}(X_1))(x_{2j} - \mathbb{E}(X_2))\; p_{ij} \end{cases}$$

For the covariance operator we obviously have

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X) \tag{478}$$

$$\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1) \tag{479}$$

As a generalization of Steiner's translation theorem it holds

Corollary: $\mathrm{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)$.

Definition of Correlation

5.5 Correlation Coefficient

The covariance is scale dependent; dividing it by the product of standard deviations we obtain a scale invariant measure of (linear) dependence.

Definition: The **correlation coefficient** is defined as

$$\text{Corr}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\,\text{Var}(X_2)}} \tag{480}$$

It is easily seen that the correlation is independent of the scale of measurement:

$$\text{Corr}\,(a_1 X_1, a_2 X_2) = \text{Corr}(X_1, X_2), \quad \forall a_1, a_2 > 0 \tag{481}$$

Definition: Two random variables X_1 and X_2 are said to be **uncorrelated**, if

$$\text{Corr}(X_1, X_2) = \text{Cov}(X_1, X_2) = 0. \tag{482}$$

Independence vs. Uncorrelatedness

Theorem: Independent random variables are uncorrelated.

Proof: Let X_1 and X_2 be independent. Then we have, in case of a continuous $\underline{X} = (X_1, X_2)^T$,

$$\begin{aligned}
\mathbb{E}(X_1 X_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_{(X_1,X_2)}(x_1, x_2) dx_2 dx_1 \qquad (483)\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_{X_1}(x_1) f_{X_2}(x_2) dx_2 dx_1\\
&= \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1) dx_1 \int_{-\infty}^{\infty} x_2 f_{X_2}(x_2) dx_2\\
&= \mathbb{E}(X_1)\mathbb{E}(X_2)
\end{aligned}$$

Then, by Steiner's Theorem: $\mathrm{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2) = 0$.

Properties of Correlation

The converse is **not true**, with one exception:

If $\underline{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}(\underline{\mu}, \Sigma)$ with $\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$,

i.e. for the correlation coefficient it holds $\rho = Corr(X_1, X_2) = 0$,
then X_1 and X_2 are independent.

Theorem: The correlation coefficient has the following properties:

$$|\mathrm{Corr}(X_1, X_2)| \leq 1 \tag{484}$$

$$|\mathrm{Corr}(X_1, X_2)| = 1 \iff \exists a, b \in \mathbb{R} : P(X_2 = aX_1 + b) = 1 \tag{485}$$

where $\mathrm{sign}(a) = \mathrm{sign}\,[\mathrm{Corr}(X_1, X_2)]$.

Corollary: The correlation coefficient measures the degree of **linear dependence** between X_1 and X_2.

Positive vs. Negative Correlation

Definition: The random variables X_1 and X_2 are said to be **positively correlated**, if

$$\text{Corr}(X_1, X_2) > 0 \tag{486}$$

Analogously, they are said to be **negatively correlated**, if

$$\text{Corr}(X_1, X_2) < 0. \tag{487}$$

Corollary: For all rv's X_1, X_2 and all real numbers $a_1, a_2 \in \mathbb{R}$ it holds:

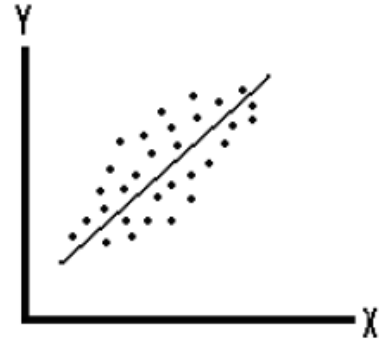
$$\text{Var}\left(a_1 X_1 + a_2 X_2\right) = a_1^2 \,\text{Var}(X_1) + a_2^2 \,\text{Var}(X_2) + 2 a_1 a_2 \,\text{Cov}(X_1, X_2)$$

Remark 1: This result finds application in risk diversification of portfolios of financial assets. Let X_1, X_2 represent stock values and $a_1, a_2 \geq 0$ such that $a_1 + a_2 = 1$, then

$$\text{Var}\left(a_1 X_1 + a_2 X_2\right) < \max\left(\text{Var}(X_1), \text{Var}(X_2)\right) \tag{488}$$

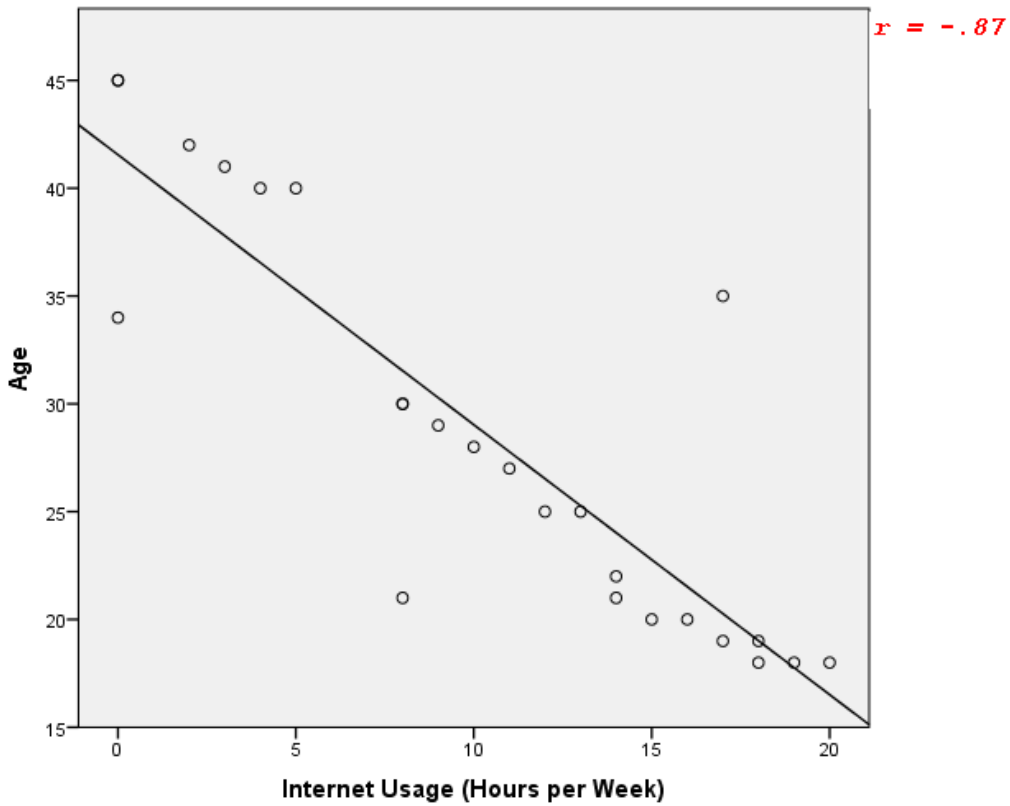
provided that $\text{Cov}(X_1, X_2) < 0$.

Correlation in datasets



Positive Correlation

Negative Correlation



Correlation in datasets

Remark 2: For two-dimensional samples (x_i, y_i) $i = 1, \ldots, n$; we can estimate the correlation as follows:

$$\widehat{\text{Corr}}(X, Y) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{489}$$

Implementation in R: Use the functions `mean`, `var`, `cov`, `cor`.

The concept of correlation plays an essential role in Time Series Analysis, where the (auto)correlation function (acf) is heavily used for discriminating between autoregressive and moving average models.

6. Generalization to multivariate distributions

In the sequel we consider random vectors

$$\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \tag{490}$$

with $n \geq 2$ components.

6.1 Distribution Function and Density

Definition: The function

$$F_{\underline{X}}(x_1, \ldots, x_n) = P\left(X_1 \leq x_1, X_2 \leq x_2, \ldots X_n \leq x_n\right) \tag{491}$$

$$\forall \underline{x} \; = \; (x_1, ..., x_n)^T \in \mathbb{R}^n$$

is called **multivariate distribution function of $\underline{X}$.**

Multivariate Distributions

In the discrete case, the distribution is represented by n-dimensional tables of the values $(x_{1i_1}, x_{2i_2}, \ldots, x_{ni_n})$ and the corresponding probabilities $(p_{1i_1}, p_{2i_2}, \ldots, p_{ni_n})$ where $i_1 \in \{1, \ldots, k_1\}$, $i_2 \in \{1, \ldots, k_2\}$, $\ldots$, $i_n \in \{1, \ldots, k_n\}$.

In the continuous case, $F_{\underline{X}}$ is generated by the pdf $f_{\underline{X}}(x_1, \ldots, x_n)$ with

$$f_{\underline{X}}(x_1, \ldots, x_n) \geq 0, \ \forall \underline{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n \tag{492}$$

Additionally, normalization is required to hold:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\underline{X}}(x_1, \ldots, x_n) \, dx_1 dx_2 \ldots dx_n = 1 \tag{493}$$

Accordingly, we define **marginal distribution functions** and **marginal densities**.

Marginals of Multivariate Distributions

Definition: a) The marginal cdf of the i-th component of $\underline{X}$ reads

$$F_{X_i}(x_i) = P(X_1 < \infty, \ldots, X_{i-1} < \infty, X_i \leq x_i, X_{i+1} < \infty, \ldots, X_n < \infty)$$

$$F_{X_i}(x_i) = F_{\underline{X}}(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty) \tag{494}$$

b) For continuous random vectors, the marginal densities of the components of $\underline{X}$ are given by

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\underline{X}}(x_1, \ldots, x_n)\, dx_1 \ldots dx_{i-1}\, dx_{i+1} \ldots dx_n \tag{495}$$

for $i = 1, \ldots, n$.

c) Accordingly, we define the marginal densities of subvectors of $\underline{X}$, e.g. for the first two components:

Marginals of Multivariate Distributions

$$f_{(X_1,X_2)}(x_1,x_2) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\underline{X}}(x_1,\ldots,x_n)\, dx_3 \ldots dx_n \tag{496}$$

Corollary: Relationship between cdf and pdf

$$F_{\underline{X}}(x_1,\ldots,x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\underline{X}}(t_1,\ldots,t_n) dt_1 \ldots dt_n \tag{497}$$

$$f_{\underline{X}}(x_1,\ldots x_n) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F_{\underline{X}}(x_1,\ldots x_n) \tag{498}$$

The last equation holds for all values $\underline{x} = (x_1,\ldots,x_n)^T \in \mathbb{R}^n$, at which $f_{\underline{X}}(\underline{x})$ is continuous.

Expectation vector, covariance matrix

6.2 Expectation and covariance

Definition: If the expectations of the components X_i; $i = 1, \ldots, n$; exist, then we call the n-dimensional vector

$$\mathbb{E}(\underline{X}) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix} \tag{499}$$

the **expectation vector** of $\underline{X}$.

Definition: The matrix

$$\mathrm{Cov}(\underline{X}) := \Big(\mathrm{Cov}\left(X_i, X_j\right) \Big)_{i,j=1,\ldots,n} \tag{500}$$

is called the **covariance matrix** of the random vector $\underline{X}$.

Expectation vector, covariance matrix

Corollary: The covariance matrix is positive semidefinite, i.e.

$$\underline{a}^T \operatorname{Cov}(\underline{X})\underline{a} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \operatorname{Cov}(X_i, X_j) \geq 0 \ \forall \underline{a} = (a_1, ..., a_n)^T \in \mathbb{R}^n \quad (501)$$

Proof: Let $\underline{a} = (a_1, \ldots, a_n)^T \neq 0$ and $Y = a_1 X_1 + \cdots + a_n X_n$, then

$$\begin{aligned}
0 \ \leq \ \operatorname{Var}(Y) \ &= \ \operatorname{Var}\left(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n\right) \ = \ \mathbb{E}\left[Y - \mathbb{E}(Y)\right]^2 \\
&= \ \mathbb{E}\left[a_1 X_1 + \cdots + a_n X_n - a_1 \mathbb{E}(X_1) - \cdots - a_n \mathbb{E}(X_n)\right]^2 \\
&= \ \mathbb{E}\left[a_1 \left(X_1 - \mathbb{E}(X_1)\right) + a_2 \left(X_2 - \mathbb{E}(X_2)\right) + \ldots a_n \left(X_n - \mathbb{E}(X_n)\right)\right]^2 \\
&= \ \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left(X_i - \mathbb{E}(X_i)\right)\left(X_j - \mathbb{E}(X_j)\right)\right] \\
&= \ \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \underbrace{\mathbb{E}\left[\left(X_i - \mathbb{E}(X_i)\right)\left(X_j - \mathbb{E}(X_j)\right)\right]}_{\operatorname{Cov}(X_i, X_j)}
\end{aligned}$$

observing that variances cannot be negative.

Remark: The proof lets us immediately recognize that the covariance is a bilinear operator.

Example (Portfolio optimization)

Let $X_1, X_2, \ldots X_n$ represent stock values and $a_1, \ldots, a_n$ the weights (percentages) of the corresponding stocks in the portfolio $Y = a_1X_1 + a_2X_2 + \ldots a_nX_n$. Clearly, for the weights we have: $a_i \geq 0$ and $\sum_{i=1}^{n} a_i = 1$. The risk of the portfolio can be measured by its variance, thus we have to solve the following optimization problem:

$$\text{Var}\,(a_1X_1 + \cdots + a_nX_n) \to \min \text{ subject to}$$
$$\sum_{i=1}^{n} a_i\mu_i \geq \mu_0 \text{ and } a_1, \ldots, a_n \geq 0, \sum_{i=1}^{n} a_i = 1,$$

where μ_0 stands for the least revenue to be earned by the portfolio.

6.3 Independence of random variables

Definition: The components $X_1, \ldots, X_n$ of $\underline{X}$ are said to be **totally independent** if it holds

$$F_{\underline{X}}(x_1, \ldots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \ldots F_{X_n}(x_n), \tag{502}$$

i.e. the (joint) cdf of $\underline{X}$ can be written as product of the marginal cdf's of the components of $\underline{X}$.

Corollary: If the components $X_1, \ldots, X_n$ of the random vector $\underline{X}$ are continuous, then it holds, correspondingly:

$$f_{\underline{X}}(x_1, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n), \tag{503}$$

i.e. the pdf of $\underline{X}$ equals the product of the marginal densities of the components.

Independence

The concept of independence plays a central role in the definition of the so-called **likelihood function**, which, in statistics, represents the pdf of the data.

Corollary: If the components $X_1, \ldots, X_n$ of the random vector $\underline{X}$ are totally independent, then it holds:

$$\mathrm{Cov}(X_i, X_j) = 0 \qquad \forall i \neq j = 1, ..., n \tag{504}$$

i.e. the covariance matrix $\mathrm{Cov}(\underline{X})$ is a diagonal matrix:

$$\mathrm{Cov}(\underline{X}) \;=\; \mathit{diag}(\sigma_1^2, \ldots, \sigma_n^2),$$

where

$$\sigma_i^2 = \mathrm{Var}(X_i),\ i = 1, \ldots, n.$$

Maximum Likelihood Estimation (MLE)

The Likelihood function of the data is considered as a function of the parameters θ of the underlying sample distribution; e.g. we have $\theta = (\mu, \sigma^2)$ for (Gaussian) data distributed as $X \sim \mathrm{N}(\mu, \sigma^2)$.

Definition: The **Likelihood function** of iid data $X_1, ..., X_n$ with pdf $f_X(x; \theta)$ and parameter θ is defined as

$$l(\theta; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f_X(x_i; \theta) \tag{505}$$

ML estimation principle: *log* $l(\theta; x_1, ..., x_n) \longrightarrow \underset{\theta}{\mathbf{Max}}$

Example (Gaussian data)

For iid data $X_i \sim \mathrm{N}(\mu, \sigma^2)$ we have $\theta = (\mu, \sigma^2)$ and

$$l(\mu, \sigma; x_1, ..., x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-0.5(x_i - \mu)^2/\sigma^2) \tag{506}$$

Maximum Likelihood Estimation (MLE)

Example (cont'd)

Taking logarithms we get

$$log\ l(\mu, \sigma^2; x_1, ..., x_n) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad (507)$$

Setting the derivatives of *log l* w.r.t. μ and σ^2 equal to zero we obtain the likelihood equations

$$\sum_{i=1}^{n}(x_i - \mu) = 0 \ \ and \ \ n\sigma^2 = \sum_{i=1}^{n}(x_i - \mu)^2$$

which, finally, lead to the mle's:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \ and \ \ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Maximum Likelihood Estimation with R

For computing MLE's in R make use of the library(maxLik):

```
> install.packages("maxLik")
> library(maxLik)
```

Example (Estimating Weibull distribution parameters)

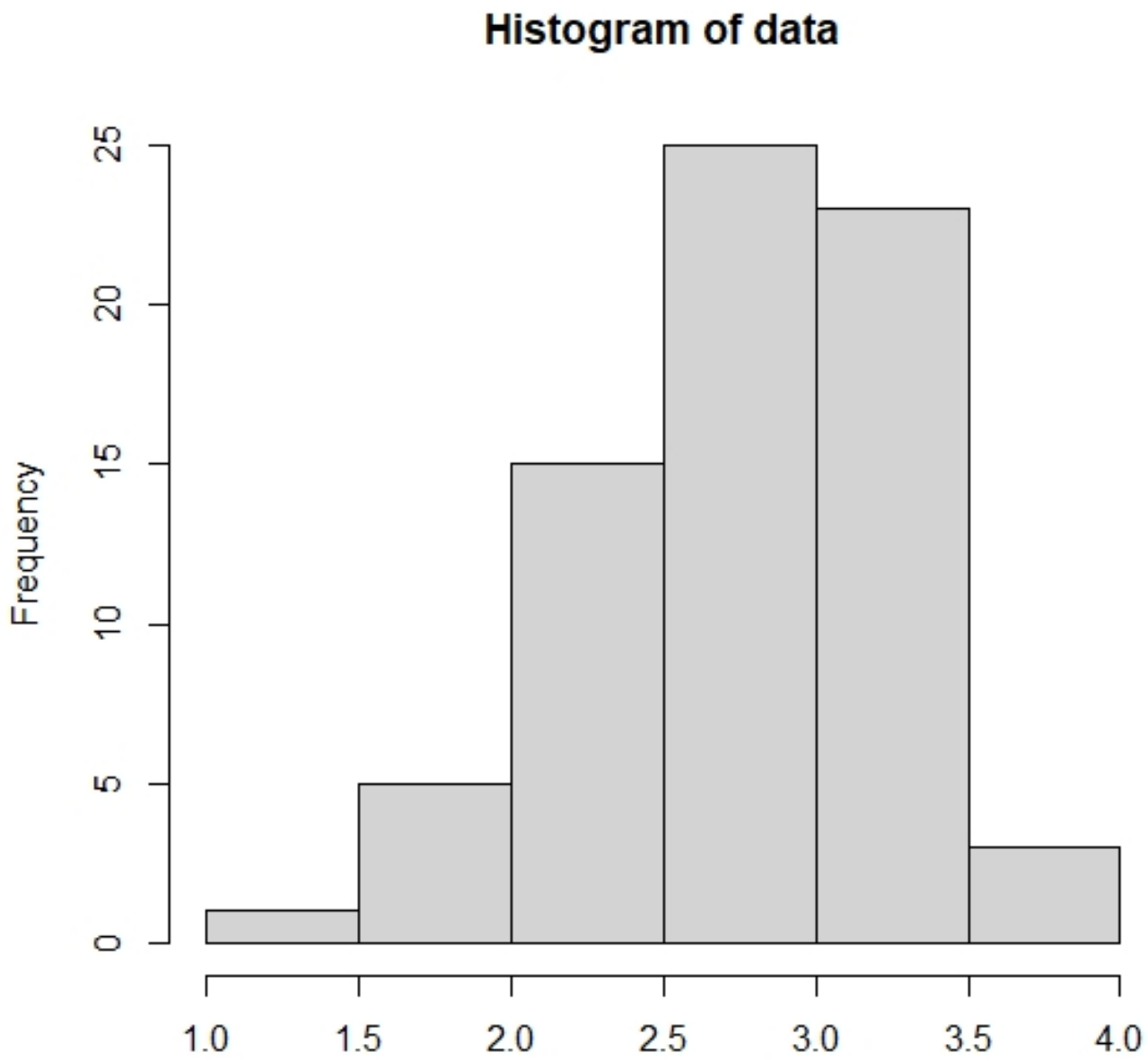
Data: 72 lifetimes of chips (measured in units of 100000 hours)

```
> hist(data)
> range(data)
[1]1.338138  3.736064
```

Data suggests a Weibull (or Gamma) distribution.
We will try a Weibull distribution and estimate its two parameters (shape and scale) by the maximum likelihood method:

```
> llf= function(par){shape= par[1]
+ scale= par[2]
+ llval= dweibull(data, shape=shape, scale=scale, log=TRUE)
+ sum(llval)}
```

Weibull Lifetime data



Maximum Likelihood Estimation with R

Example (cont'd)

```
> mle= maxLik(llf, start= c(shape=6, scale=3))
> print(summary(mle))
Maximum Likelihood estimation
Newton-Raphson maximisation, 4 iterations
Return code 1: gradient close to zero (gradtol)
Log-Likelihood: -50.93736
2 free parameters
Estimates:
        Estimate   Std. error    t value     Pr(> t)
shape   6.53472    0.61773       10.58     <2e-16 ***
scale    2.93704    0.05564       52.78     <2e-16 ***
> ks.test(data, "pweibull", 6.53, 2.94)
One-sample Kolmogorov-Smirnov test
D = 0.063288,   p-value = 0.9176
```

6.4 Examples of multivariate distributions

6.4.1 Multinomial distribution

The multinomial distribution is the most widely used discrete multivariate distribution, it is the multivariate generalization of the binomial distribution. There, the underlying Bernoulli scheme refers to a fixed event with just two outcomes (yes/no or success/failure). The multinomial distribution refers to the multinomial scheme, where experiments with $k \geq 2$ outcomes are considered. The outcomes are denoted as events $A_1, A_2, \ldots, A_k$ with

$$A_i \cap A_j = \emptyset \text{ for } i \neq j \text{ and } \bigcup_{i=1}^{k} A_i = \Omega.$$

Definition: The random vector $\underline{X} = (X_1, \ldots, X_n)^T$ is said to be **multinomially distributed** with parameters $n, p_1, \ldots, p_k$, if it holds

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} \tag{508}$$

Multinomial distribution

Here, n denotes the total number of trials, i.e. $\sum_{i=1}^{k} x_i = n$. Briefly, we write

$$\underline{X} \sim \mathrm{MN}\,(n, p_1, \ldots p_k) \tag{509}$$

Corollary: For $k = 2$ the multinomial distribution reduces to the binomial distribution

$$P\,(X_1 = x_1, X_2 = x_2) = \frac{n!}{x_1! \cdot x_2!} \cdot p_1^{x_1} \cdot p_2^{x_2} \tag{510}$$

Since we have $x_2 = n - x_1$ und $p_2 = 1 - p_1$, we further obtain

$$P\,(X_1 = x_1, X_2 = x_2) \quad = \quad \binom{n}{x_1} \cdot p_1^{x_1} \cdot (1 - p_1)^{n - x_1}, \tag{511}$$

i.e. $X_1 \sim \mathrm{Bi}(n, p_1)$.

Multinomial distribution

Example (Multinomial probabilities in R)

During the municipal elections 2015 it happened for the first time in a small Carinthian village (named Preitenegg) that two candidates received the same number of votes (365) in a second ballot (Stichwahl). Further, there were 92 citizens who had not voted and 5 people had given invalid votes.

Q: What is the probability of this voting result given that the two candidates have an equal percentage (44%) of voters behind them, and that, on the average, there are about 11% non-voters and 1% invalid votes in such duels in small communities?

We have $p_1 = p_2 = 0.44, p_3 = 0.11, p_4 = 0.01$, thus, it follows

$$\begin{aligned} & P(X_1 = 365, X_2 = 365, X_3 = 92, X_4 = 5) \\ &= \frac{827!}{365!365!92!5!} \cdot (0.44)^{365} \cdot (0.44)^{365} \cdot (0.11)^{92} \cdot (0.01)^5 \\ &= 0.0001068. \end{aligned}$$

Multinomial distribution

Example (cont'd)

For computation with R, we use the function "dmultinom":

```
> dmultinom(c(365,365,92,5),
prob=c(0.44,0.44,0.11,0.01))
 [1] 0.0001068
```

Corollary: If $\underline{X} = (X_1, \ldots, X_k)^T$ follows a multinomial distribution with parameters $n, p_1, ..., p_k$, then it holds:

a) The components of $\underline{X}$ are binomially distributed,

$$X_i \sim Bi(n, p_i);\ i = 1, ..., k.$$

b) The covariances are given by

$$\mathrm{Cov}(X_i, X_j) = -np_ip_j, \quad \forall i, j = 1, \ldots, k, i \neq j \tag{512}$$

c) The correlation coefficients read

$$\begin{aligned} \mathrm{Corr}(X_i, X_j) &= \frac{-np_ip_j}{\sqrt{\mathrm{Var}(X_i)\,\mathrm{Var}(X_j)}} \\ &= \frac{-p_ip_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}} \end{aligned} \tag{513}$$

$\forall i, j = 1, \ldots, k, i \neq j.$

6.4.2 Multinormal distribution

Definition: The continuous random vector $\underline{X} = (X_1, \ldots, X_n)^T$ is said to have an n**- dimensional normal distribution** with parameters $\underline{\mu} \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$, Σ positive definite, if it has pdf

$$f_{\underline{X}}(\underline{x}) = (2\pi)^{-\frac{n}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu})\right), \forall \underline{x} \in \mathbb{R}^n$$

The parameters of the normal distribution have the following

Multinormal distribution

interpretation:

$$\underline{\mu} = \mathbb{E}(\underline{X}),\ \Sigma = \mathrm{Cov}(\underline{X}). \tag{514}$$

Example (Directional speeds of gas molecules)

Let denote $\underline{V} = (V_x, V_y, V_z)^T$ the speed vector of gas molecules into the three spatial directions. Maxwell had proved that the components are independently and normally distributed with (identical) parameters:

$$\mathbb{E}(V_x) = \mathbb{E}(V_y) = \mathbb{E}(V_z) = 0$$
$$\mathrm{Var}(V_x) = \mathrm{Var}(V_y) = \mathrm{Var}(V_z) = \sigma^2$$

where $\sigma^2 = \frac{kT}{m}$ and T= temperature, m= mass, k= Boltzmann constant. Then, with $c = \sqrt{2\pi\sigma^2}$ we have

$$f_{\underline{V}}(v_x, v_y, v_z) = \frac{1}{c}\exp\left(-\frac{v_x^2}{2\sigma^2}\right) \cdot \frac{1}{c}\exp\left(-\frac{v_y^2}{2\sigma^2}\right) \cdot \frac{1}{c}\exp\left(-\frac{v_z^2}{2\sigma^2}\right)$$

Multinormal distribution

Example (cont'd)

$$
\begin{aligned}
f_{\underline{V}}(v_x, v_y, v_z) &= (2\pi\sigma^2)^{-\frac{3}{2}} exp\left(-\frac{1}{2\sigma^2}(v_x^2 + v_y^2 + v_z^2)\right) \\
&= (2\pi)^{-\frac{3}{2}} det\,(\Sigma)^{-\frac{1}{2}}\, exp\left(-\frac{1}{2}\underline{v}^T\Sigma^{-1}\underline{v}\right),
\end{aligned}
$$

i.e. $\underline{V} \sim \mathrm{N}(\underline{\mu}, \Sigma)$ with $\underline{\mu} = \underline{0}$ and $\Sigma = \sigma^2 I_3$.

Recall that in case of a normal distribution **uncorrelatedness** is equivalent to (total) **independence**. This is the case if and only if $Cov(\underline{X}) = \Sigma$ is a diagonal matrix (as in the above example, where Σ was a multiple of the identity matrix).

Problem: How can we check for (approximate) multivariate normality of data?

Checking multivariate normality

Cramér-Wold Theorem: The distribution of $\underline{X} = (X_1, ..., X_n)^T$ is completely determined by the set of all projections

$$\left\{\underline{a}^T \underline{X} = a_1 X_1 + ... + a_n X_n : \underline{a} = (a_1, ..., a_n)^T \in \mathbb{R}^n\right\}.$$

Corollary: $\underline{X} = (X_1, ..., X_n)^T$ is normally distributed, $\underline{X} \sim \mathrm{N}(\underline{\mu}, \Sigma)$ if and only if

$$\underline{a}^T \underline{X} = a_1 X_1 + ... + a_n X_n \sim \mathrm{N}\left(\underline{a}^T \underline{\mu}, \underline{a}^T \Sigma \underline{a}\right), \quad \forall \underline{a} \in \mathbb{R}^n.$$

Consequently, all sub-vectors and linear combinations of components of a normal random vector are normally distributed, again. In particular, all histograms of the single components must exhibit bell-curved shapes and contour plots of all pairs of variables must exhibit elliptic surfaces. Pairwise scatter plots may be used to verify this.

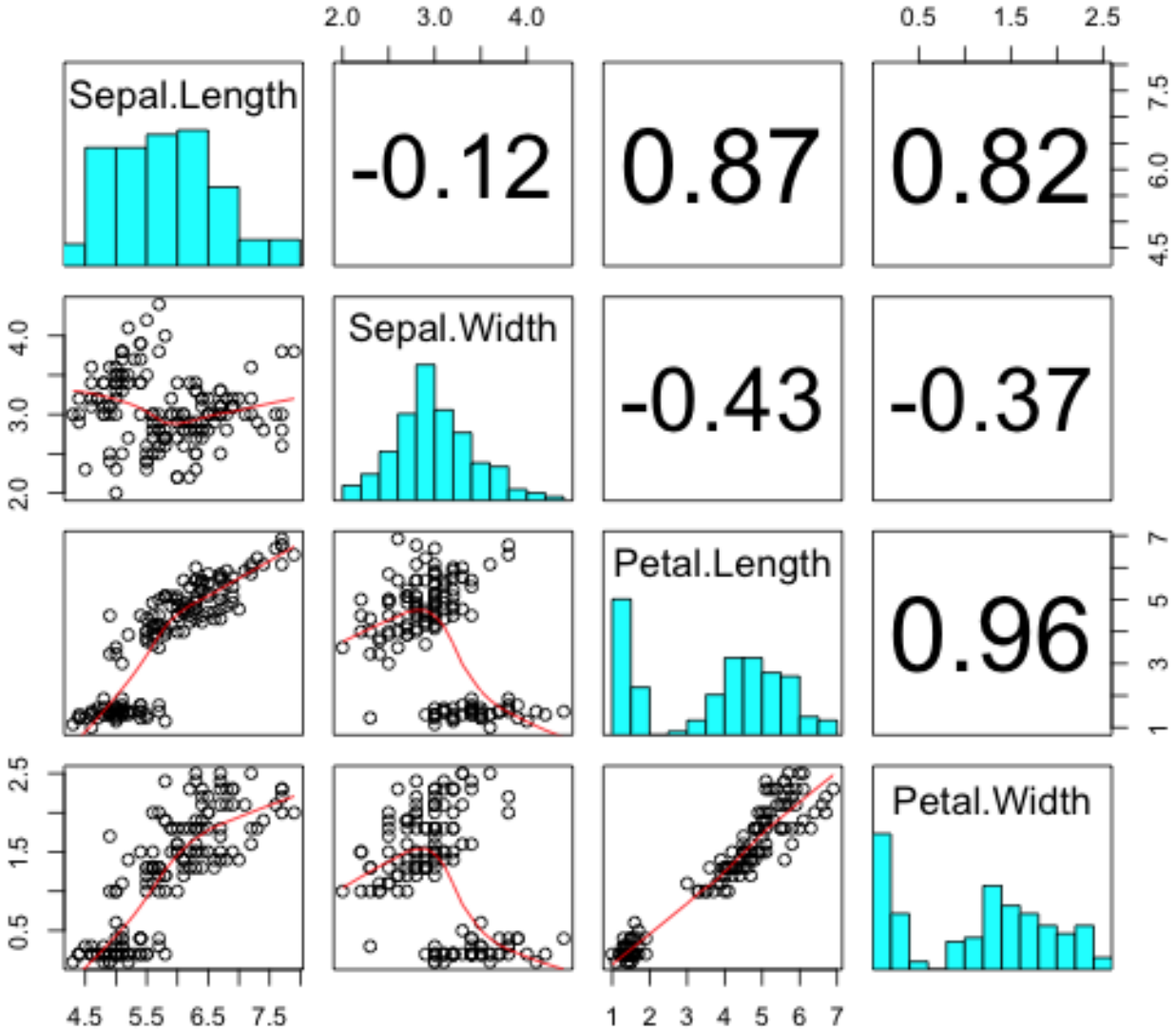
R-Implementation: `> pairs(data frame)`

Pairwise scatter plots

Consider the data set iris from the package datasets. This includes measurements of the Iris flower in centimeters. The 4 variables refer to petal and sepal lengths and widths of the flowers. There are 50 measurements for each of three species (setosa, virginica and versicolor).

```
> data(iris)
> head(iris,5)
  Sepal.Length Sepal.Width Petal.Length Petal.Width
Species
1          5.1         3.5          1.4         0.2
setosa
2          4.9         3.0          1.4         0.2
setosa
3          4.7         3.2          1.3         0.2
setosa
4          4.6         3.1          1.5         0.2
setosa
5          5.0         3.6          1.4         0.2
setosa
>
> pairs(iris[,-5], diag.panel=panel.hist,
+ upper.panel=panel.cor, lower.panel=panel.smooth)
```

Pairwise scatter plots



Computations in R

Mean, covariance and correlation matrix in R:

```
> apply(iris[,-5], 2, mean)
 Sepal.Length Sepal.Width Petal.Length Petal.Width
     5.843333    3.057333    3.758000    1.199333
```

```
> cov(iris[,-5])  # 4x4 matrix of covariances
```

```
> cor(iris[,-5])  # 4x4 matrix of correlations
```

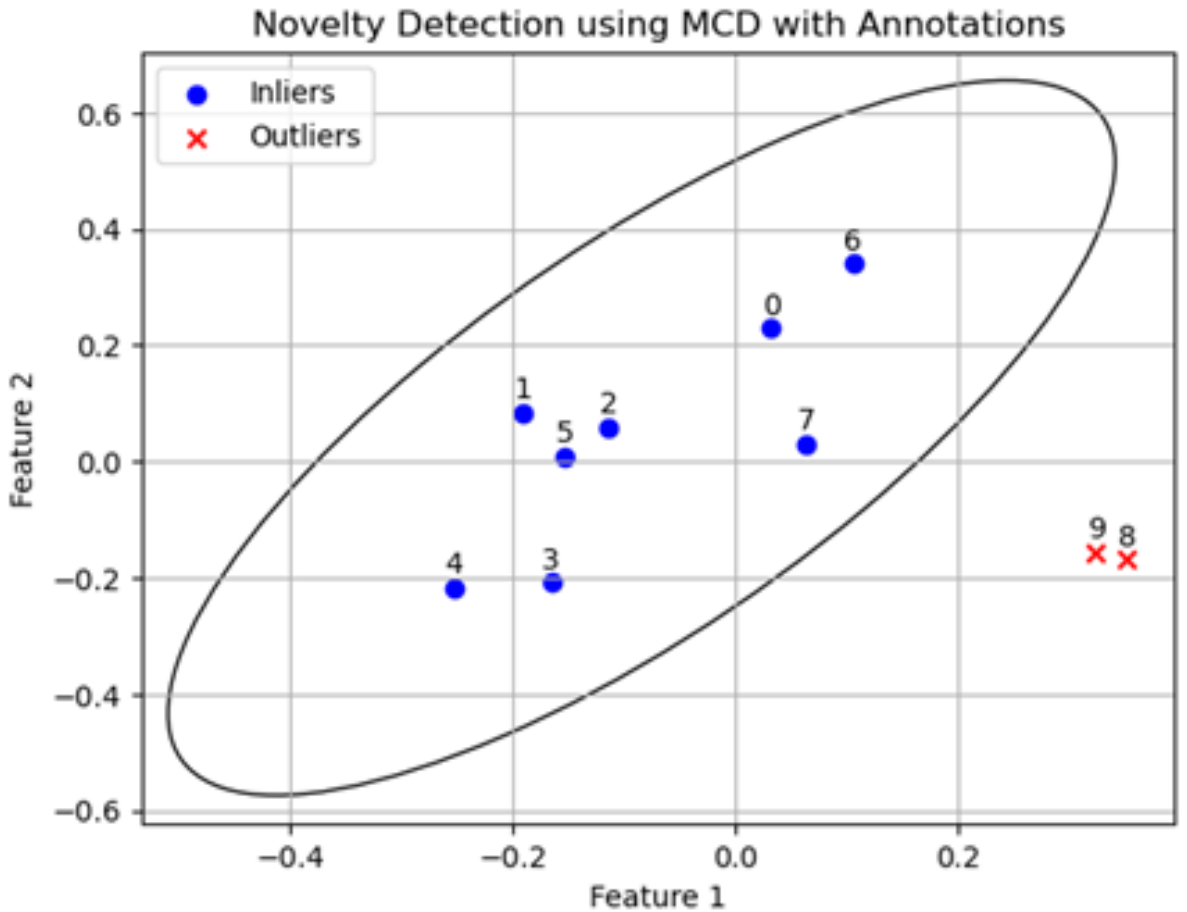
For computation of cdf, pdf, probabilities, quantiles and simulation of realizations of multivariate normal random vectors use

```
> library(mvtnorm)
```

Anomaly Detection via MCD

How to use Minimum Covariance Determinant (MCD) to detect novel news headlines, see e.g.

`https://towardsdatascience.com/textual-novelty-detection-ce81d2e689bf`

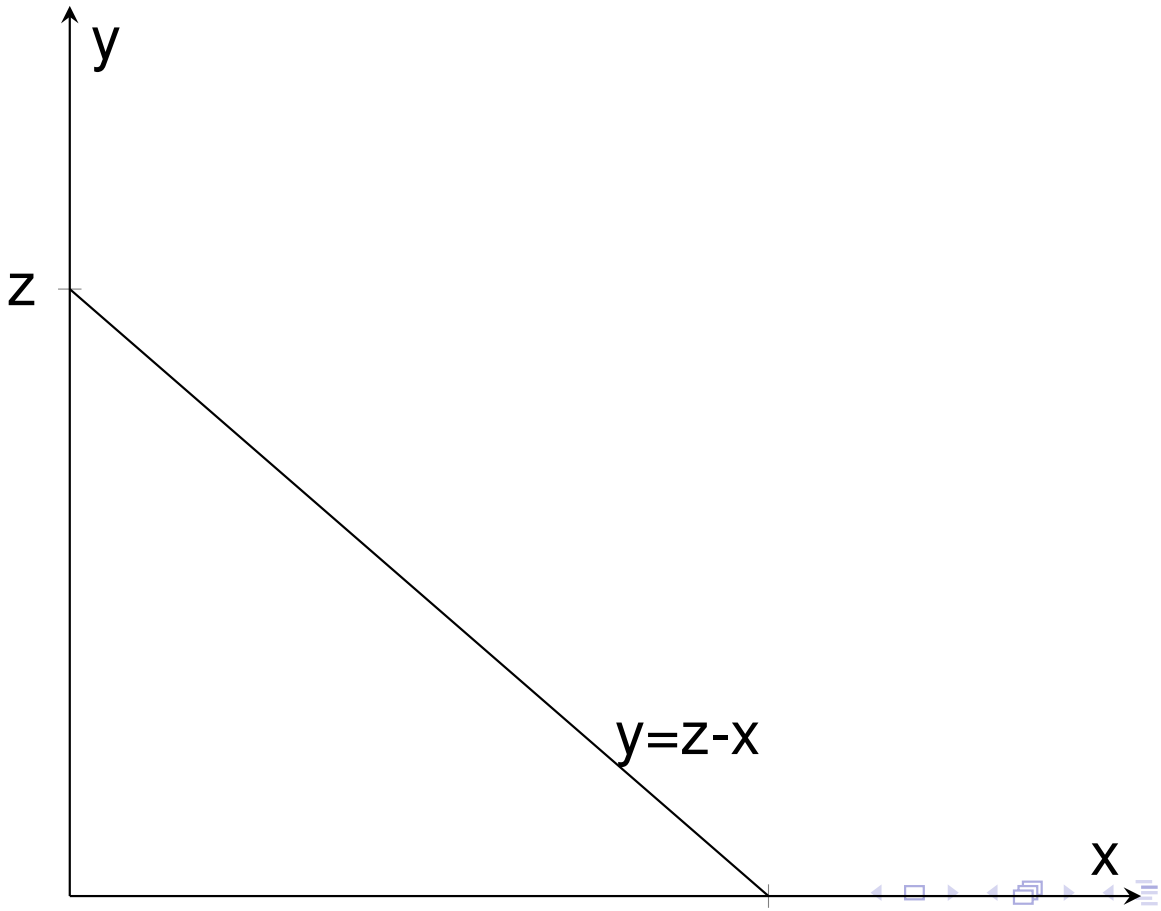


7. Functions of Random Variables

7.1 Sums of rv's

Let X, Y be independent rv's with cdf's F_X, F_Y.

Problem: Find the cdf of $Z = X + Y$.



Convolution Integral

$$
\begin{aligned}
F_Z(z) &= P(X+Y \le z) && (515)\\
&= \iint dF_{(X,Y)}(x,y)\\
&= \iint dF_X(x)\, dF_Y(y) \text{ (independence)}\\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z-y} dF_X(x) \right] dF_Y(y)\\
&= \int_{-\infty}^{\infty} F_X(x)\Big|_{x=-\infty}^{z-y} dF_Y(y)\\
&= \int_{-\infty}^{\infty} F_X(z-y)\, dF_Y(y) = (F_X * F_Y)(z) \text{ convolution of cdf's}
\end{aligned}
$$

Convolution Integral

Definition: The operation $*$ defined by

$$(F_X * F_Y)(z) = \int_{-\infty}^{\infty} F_X(z-y)\, dF_Y(y), \quad \forall z \in \mathbb{R} \tag{516}$$

is called **convolution** (Faltung) of the distributions of X and Y.

Corollary: The sum $Z = X + Y$ of independent rv's X and Y has cdf

$$F_Z(z) = (F_X * F_Y)(z) \tag{517}$$

Note: The convolution operation $*$ is symmetric, i.e.

$$(F_X * F_Y)(z) = (F_Y * F_X)(z) \tag{518}$$

$$\int_{-\infty}^{\infty} F_X(z-y)\, dF_Y(y) = \int_{-\infty}^{\infty} F_Y(z-x)\, dF_X(x) \tag{519}$$

Convolution Integral

Example (Convolution of a discrete rv with a continuous rv)

Let be X a discrete rv,

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_i & \dots \\ p_1 & p_2 & \dots & p_i & \dots \end{pmatrix} \tag{520}$$

and Y a continuous rv with pdf f_Y. Then the sum $Z = X + Y$ has cdf

$$F_Z(z) = \int_{-\infty}^{\infty} F_Y(z-x)\, dF_X(x) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z-x} f_Y(y)\, dy \right] dF_X(x)$$

, which then becomes

$$F_Z(z) = \sum_{i=1}^{\infty} \left[\int_{-\infty}^{z-x_i} f_Y(y)\, dy \right] p_i \tag{521}$$

Convolution Integral

Example (cont'd)

The pdf $f_Z(z) = \frac{\partial}{\partial z} F_Z(z)$ reads

$$f_Z(z) = \sum_{i=1}^{\infty} p_i \, f_Y(z - x_i) \tag{522}$$

Example (Convolution of continuous rv's)

Let X and Y be continuous rv's with pdf's f_X and f_Y, resp. Then,

$$F_Z(z) = \int_{-\infty}^{\infty} F_Y(z - x) \, dF_X(x) = \int_{-\infty}^{\infty} F_Y(z - x) \, f_X(x) dx$$

and the corresponding pdf $f_Z(z) = \frac{\partial}{\partial z} F_Z(z)$ of Z becomes

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z - x) \, f_X(x) \, dx \tag{523}$$

Convolution Integral

Example (cont'd)

This is also termed the **convolution integral for pdf's**. We write

$$f_Z(z) = (f_X * f_Y)(z) \tag{524}$$

Example (Convolution of uniform pdf's)

Suppose that $X \sim U[0, a]$ and $Y \sim U[0, a]$.

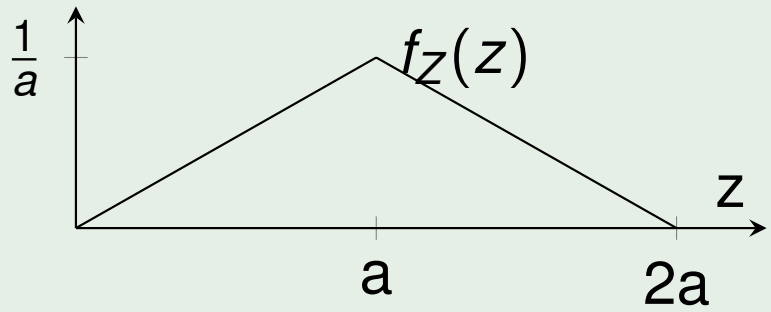
Q: What is the pdf of $Z = X + Y$?

It is easily seen that $0 \leq z \leq 2a$ and $\max(0, z - a) \leq x$. Likewise, $x \leq \min(z, a)$, from eq. (522) we then obtain

$$f_Z(z) = \begin{cases} \int_0^z \frac{1}{a^2} dx = \frac{z}{a^2} & 0 \leq z \leq a \\ \int_{z-a}^a \frac{1}{a^2} dx = \frac{2a-z}{a^2} & a \leq z \leq 2a \end{cases}$$

Convolution of uniform pdf's

Example (cont'd)



It can be easily conjectured that the pdf of the sum of three uniformly i.i.d. rv's has parabolic shape. For an increasing number of summands, the pdf approaches a Gaussian pdf, due to the CLT.

Corollary: The sum $Z = X + Y$ of non-negative, continuously on $(0, \infty)$ distributed rv's X and Y has pdf

$$f_Z(z) = \int_0^z f_X(x)\, f_Y(z-x)\, dx, \quad \forall z \in (0, \infty) \tag{525}$$

Application: Computer Vision

Kernel convolution to generate blurring effects or to effect deblurring, see e.g.

`https://towardsdatascience.com/convolution-explained-introduction-to-convolutional-ne`

`https://medium.com/@koushikkushal95/understanding-convolutional-neural-networks-cnns-in-de`

Computer Vision ultimately leads to classification tasks using CNNs (convolutional neural networks)
Advanced Computer Vision Environments often work with **TensorFlow**

7.2 Additive Distributions

Are there distributions which are additive under convolution, i.e. for which the distribution of the sum of independent rv's is the same as that of their components, with only the parameters getting updated?

We already know that this is true for independent normal rv's:

$$X_1 \sim N(\mu_1, \sigma_1^2),\ X_2 \sim N(\mu_2, \sigma_2^2) \Rightarrow\ X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Such a property of convolution additivity also holds for a few other distributions.
Recall the gamma distribution with parameters α and λ, having pdf

$$X \sim \mathrm{Ga}(\alpha, \lambda) \Rightarrow f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} e^{-\lambda x},\quad x > 0$$

with the special cases of the exponential distribution

$$\mathrm{Ga}(\alpha = 1, \lambda) = \mathrm{Ex}(\lambda),$$

Convolution of gamma rv's

and the so-called **Chi-Square(n) distribution** χ^2_n, where

$$\mathsf{Ga}(\alpha = \tfrac{n}{2}, \lambda = \tfrac{1}{2}) \;=\; \chi^2_n.$$

Theorem: Summation theorem for Gamma distributions

The sum of two independent gamma distributed rv's is gamma distributed again, provided they have the same second parameter λ:

$$\mathsf{Ga}(\alpha_1, \lambda) * \mathsf{Ga}(\alpha_2, \lambda) = \mathsf{Ga}(\alpha_1 + \alpha_2, \lambda) \tag{526}$$

Corollary: For χ^2 distributions the following convolution theorem holds:

$$\begin{aligned} \chi^2_{n_1} * \chi^2_{n_2} &= \mathsf{Ga}\left(\frac{n_1}{2}, \frac{1}{2}\right) * \mathsf{Ga}\left(\frac{n_2}{2}, \frac{1}{2}\right) \\ &= \mathsf{Ga}\left(\frac{n_1 + n_2}{2}, \frac{1}{2}\right) \\ &= \chi^2_{n_1+n_2} \end{aligned} \tag{527}$$

Convolution of gamma rv's

Note: The summation theorem does not hold for exponential distributions with different intensities:

$$\mathrm{Ex}(\lambda_1) * \mathrm{Ex}(\lambda_2) = \mathrm{Ga}(1, \lambda_1) * \mathrm{Ga}(1, \lambda_2) \neq \mathrm{Ex}(\lambda_1 + \lambda_2) \tag{528}$$

If $\lambda_1 = \lambda_2 = \lambda$ then $\mathrm{Ex}(\lambda_1) * \mathrm{Ex}(\lambda_2) = \mathrm{Ga}(2, \lambda)$.

Theorem: Summation theorem for the Poisson Distribution

For independent Poisson rv's $X_1 \sim \mathrm{Po}(\lambda_1)$ and $X_2 \sim \mathrm{Po}(\lambda_2)$ it holds

$$X_1 + X_2 \sim \mathrm{Po}(\lambda_1 + \lambda_2) \tag{529}$$

Finally, a summation theorem also holds for independent binomially distributed random variables provided they have the same success parameter $p \in (0, 1)$:

$$\mathrm{Bi}(n_1, p) * \mathrm{Bi}(n_2, p) = \mathrm{Bi}(n_1 + n_2, p).$$

7.3 Ratios of random variables

7.3.1 Jacobi Matrix

In section 2.5 we learned about the change-of-variable theorem when considering transformations of one-dimensional rv's. We now study a multivariate version of this theorem.

Consider a random vector $\underline{X} = (X_1, \ldots, X_n)^T$ and a transformed random vector $\underline{Z} = (Z_1, \ldots, Z_n)^T$ resulting from a multivariate mapping

$$\underline{Z} = \underline{g}(\underline{X}) \text{ with } \underline{g} : \mathbb{R}^n \to \mathbb{R}^n.$$

We assume that the mapping $\underline{g}$ is one-to-one and has differentiable components $g_i, i = 1, \ldots, n$. Then,

$$\underline{X} = \underline{g}^{-1}(\underline{Z}),$$

where $\underline{g}^{-1}$ denotes the inverse mapping. Further, we need the partial

Jacobi Matrix

derivatives of the components of $\underline{X}$ w.r.t. the components of $\underline{Z}$.

Definition: The matrix

$$J(\underline{z}) = \left(\frac{\partial x_i}{\partial z_j} \right)_{i,j=1,..,n} \tag{530}$$

is called **Jacobi matrix** associated with the transformation $\underline{g}$. The determinant $\det(J(\underline{z}))$ of this matrix is called the **Jacobian**.

Theorem: Multivariate change-of-variable theorem

Let be $\underline{X}$ an $n-$ dimensional continuous random vector and $\underline{g}$ a one-to-one mapping with continuously differentiable component functions. Then the transformed random vector $\underline{Z} = \underline{g}(\underline{X})$ has pdf

$$f_{\underline{Z}}(\underline{z}) = f_{\underline{X}}\left(\underline{g}^{-1}(\underline{z})\right) |\det J(\underline{z})| \tag{531}$$

7.3.2 Application: probability densities of ratios

Let $\underline{X} = (X_1, X_2)^T$ be a two-dimensional random vector with pdf $f_{(X_1,X_2)}(x_1, x_2)$. We want to determine the distribution of $\frac{X_1}{X_2}$. To this, define

$$\underline{Z} = \underline{g}(\underline{X}) = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} \frac{X_1}{X_2} \\ X_2 \end{pmatrix} \tag{532}$$

We then have to determine the marginal distribution of $Z_1 = \frac{X_1}{X_2}$. From the above transformation we get

$$x_1 = z_1 z_2, \; x_2 = z_2. \tag{533}$$

The Jacobi matrix then becomes

$$J(\underline{z}) = \begin{pmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_1}{\partial z_2} \\ \frac{\partial x_2}{\partial z_1} & \frac{\partial x_2}{\partial z_2} \end{pmatrix} = \begin{pmatrix} z_2 & z_1 \\ 0 & 1 \end{pmatrix} \tag{534}$$

The determinant of the Jacobi matrix is $det(J(\underline{z})) = z_2$. Using the above change-of-variable theorem, we get

$$f_{\underline{Z}}(z_1, z_2) = f_{(X_1,X_2)}(z_1 z_2, z_2) \ |z_2| \tag{535}$$

Corollary: If X_1 and X_2 are independent rv's, then it holds

$$f_{Z_1}(z_1) = \int_{-\infty}^{\infty} f_{X_1}(z_1 z_2) \ f_{X_2}(z_2) \ |z_2| \ dz_2 \tag{536}$$

7.3.3 Student-t-Distribution

Definition: The random variable T is said to be **Student-t-distributed** with n degrees of freedom, briefly: $T \sim t_n$, if it can be represented as a ratio

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} \quad where \ X \sim N(0,1), \ Y \sim \chi^2_n \tag{537}$$

and, in addition, X and Y are independent.

Student-t- distribution

Problem: Find the pdf f_T of $T = \frac{X}{\sqrt{Y/n}}$

$$\text{Transformation } \underline{g} : \underline{X} = \begin{pmatrix} X \\ Y \end{pmatrix} \to \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} X/\sqrt{Y/n} \\ Y \end{pmatrix}$$

$$Z_1 = \frac{X}{\sqrt{Y/n}} \Rightarrow X = Z_1\sqrt{\frac{Z_2}{n}}$$

$$Z_2 = Y \Rightarrow Y = Z_2$$

$$\Rightarrow \text{Jacobi matrix } J(\underline{z}) = \begin{pmatrix} \frac{\partial x}{\partial z_1} & \frac{\partial x}{\partial z_2} \\ \frac{\partial y}{\partial z_1} & \frac{\partial y}{\partial z_2} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{z_2}{n}} & \frac{z_1}{\sqrt{n}}\frac{1}{2\sqrt{z_2}} \\ 0 & 1 \end{pmatrix}$$

$$\Rightarrow \det(J(\underline{z})) = \sqrt{\frac{z_2}{n}}$$

$$\Rightarrow f_{z_1}(z_1) = \int_0^\infty \underbrace{f_X(z_1\sqrt{\frac{z_2}{n}})}_{N(0,1)} \underbrace{f_Y(z_2)}_{Ga(\frac{n}{2},\frac{1}{2})} \sqrt{\frac{z_2}{n}}\, dz_2$$

Student-t- distribution

$$= \int_0^\infty \frac{1}{\sqrt{2\pi}} \, exp\left(-\frac{1}{2}\frac{z_1^2 z_2}{n}\right) \, \frac{(\frac{1}{2})^{(\frac{n}{2})}}{\Gamma(\frac{n}{2})} \, z_2^{\frac{n}{2}-1} exp(-\frac{z_2}{2}) \sqrt{\frac{z_2}{n}} \, dz_2$$

$$= \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})\sqrt{2\pi n}} \int_0^\infty \underbrace{z_2^{(n-1)/2} \, exp\left(-\frac{z_2}{2}(1+\frac{z_1^2}{n})\right)}_{\text{gamma kernel: } Ga\left(\frac{n+1}{2}, \frac{1}{2}(1+\frac{z_1^2}{n})\right)} \, dz_2$$

Observing that for $W \sim \text{Ga}(\alpha, \lambda)$ it holds

$$\int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} w^{\alpha-1} e^{-\lambda w} \, dw = 1 \Rightarrow \int_0^\infty w^{\alpha-1} e^{-\lambda w} \, dw = \frac{\Gamma(\alpha)}{\lambda^\alpha},$$

we finally obtain for $f_{z_1}(z_1)$:

Student-t- distribution

$$= \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})\sqrt{2\pi n}} \frac{\Gamma(\frac{n+1}{2})}{(\frac{1}{2})^{(n+1)/2}\left(1+\frac{z_1^2}{n}\right)^{(n+1)/2}}$$

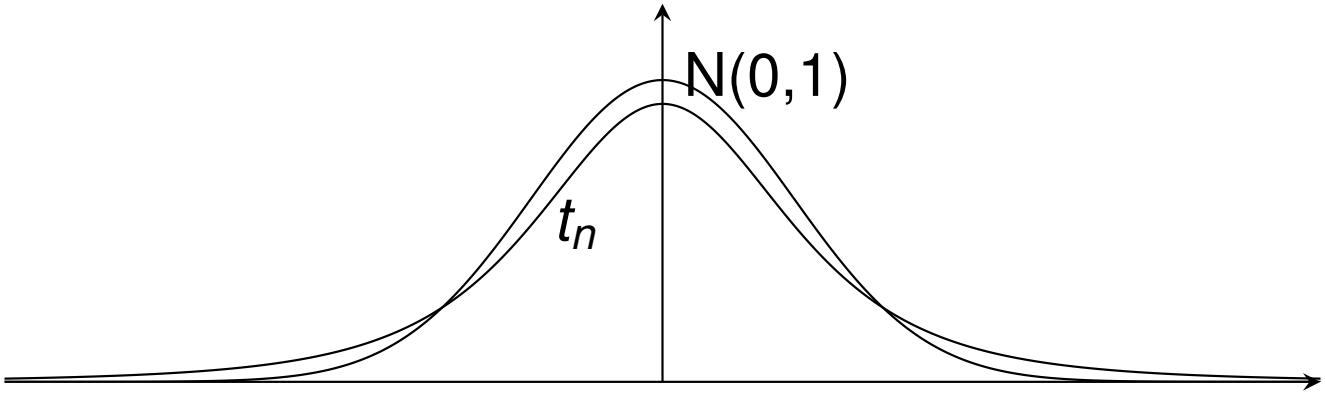
$$= \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi n}}\left(1+\frac{z_1^2}{n}\right)^{-(n+1)/2}$$

Corollary: The pdf of the Student-t_n-distribution reads as follows:

$$f_T(t) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \quad -\infty < t < \infty$$

Student-t- distribution

Graphical illustration:



For $n \to \infty$ we have: $f_T(t) \to \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}$,

observing that $e^a = \lim_{n\to\infty} (1 + \frac{a}{n})^n$

Special case: $n = 1 \Rightarrow$ t_1 = Cauchy distribution.

In general: For $T \sim t_n$ it holds

$$\mathbb{E}(T) = 0 \text{ if } n > 1, \text{Var}(T) = \frac{n}{n-2} \text{ if } n > 2.$$

Statistical Applications of Student-t-Distribution

t-Test for means (one- and two-sample problems), see e.g.
https://www.wiley.com/en-sg/Applied+Statistics%3A+Theory+and+Problem+Solutions+with+R-p-9781119551522

tSNE for clustering high-dimensional data

7.3.4 Fisher-F- distribution

Definition: Let $X \sim \chi^2_n$ and $Y \sim \chi^2_m$ be independent rv's, then the ratio

$$Z = \frac{X/n}{Y/m} \tag{538}$$

is said to be **Fisher-F- distributed** with (n, m) degrees of freedom, briefly: $Z \sim \mathrm{F}_{n,m}$.

Corollary: If $Z \sim \mathrm{F}_{n,m}$ then Z has pdf

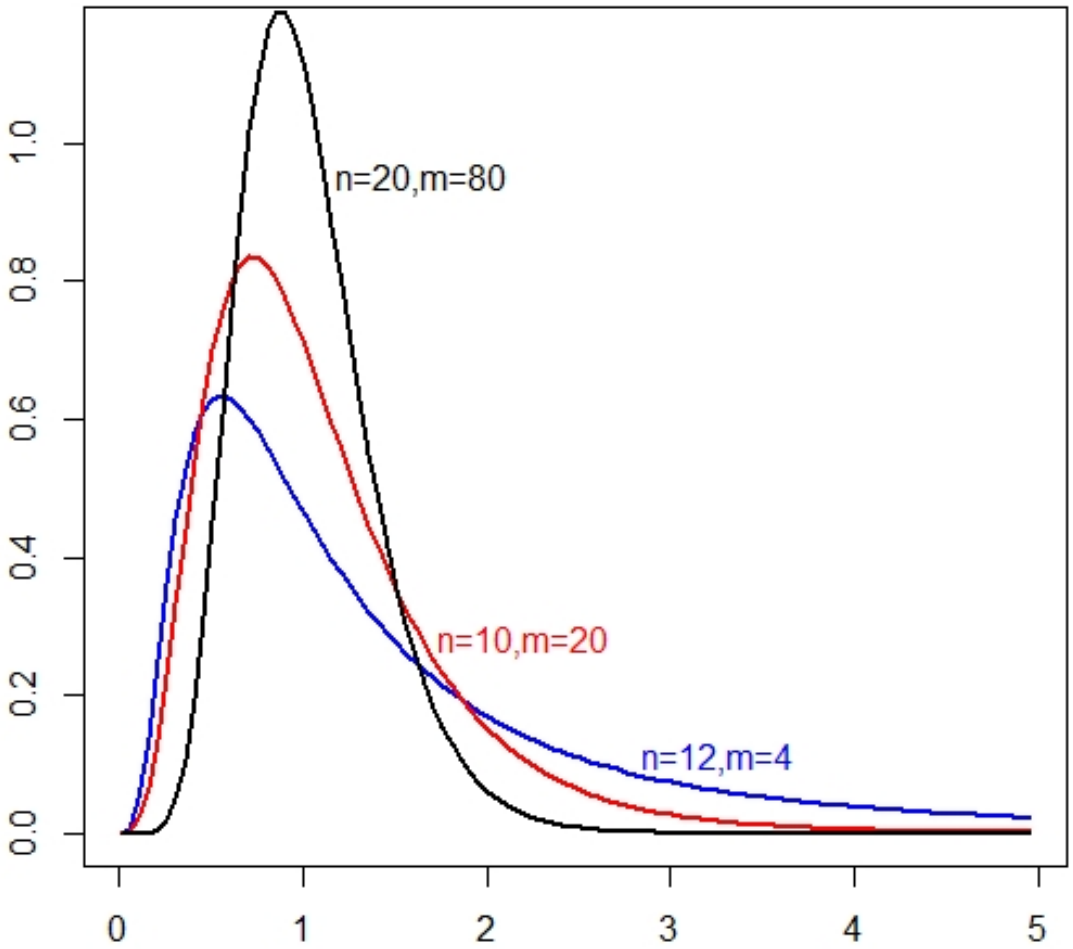
$$f_Z(z) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \cdot \left(\frac{n}{m}\right)^{\frac{n}{2}} \cdot \frac{z^{n/2-1}}{\left(1+\frac{n}{m}z\right)^{(m+n)/2}} \tag{539}$$

The proof follows again from the corresponding Jacobi transformation as illustrated in the previous section for the Student-t- distribution.

Statistical Applications: F-Test for variance components

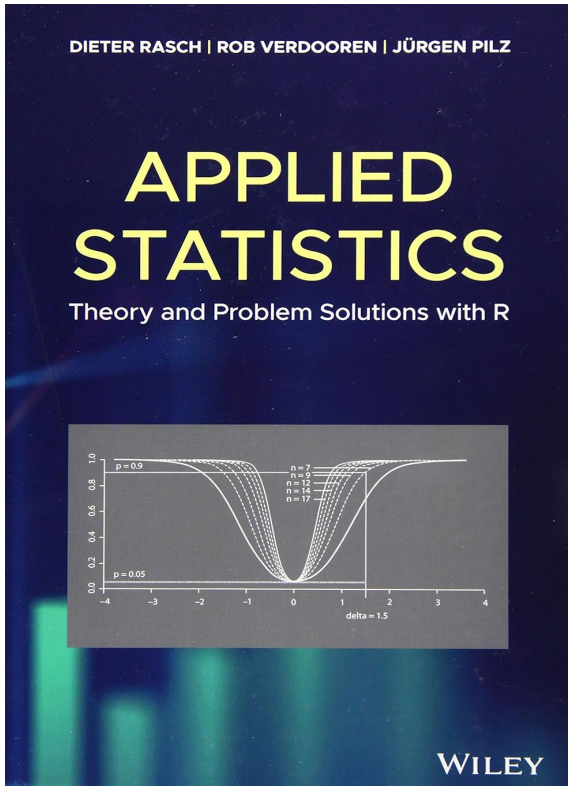
Fisher pdf

The figure below shows the pdf's for Fisher distributions with different combinations of degrees of freedom (n, m).



Statistics & Applications

For classical statistical modeling (estimation, testing, regression, anova, DoE,...) including sample size determination see e.g. our book

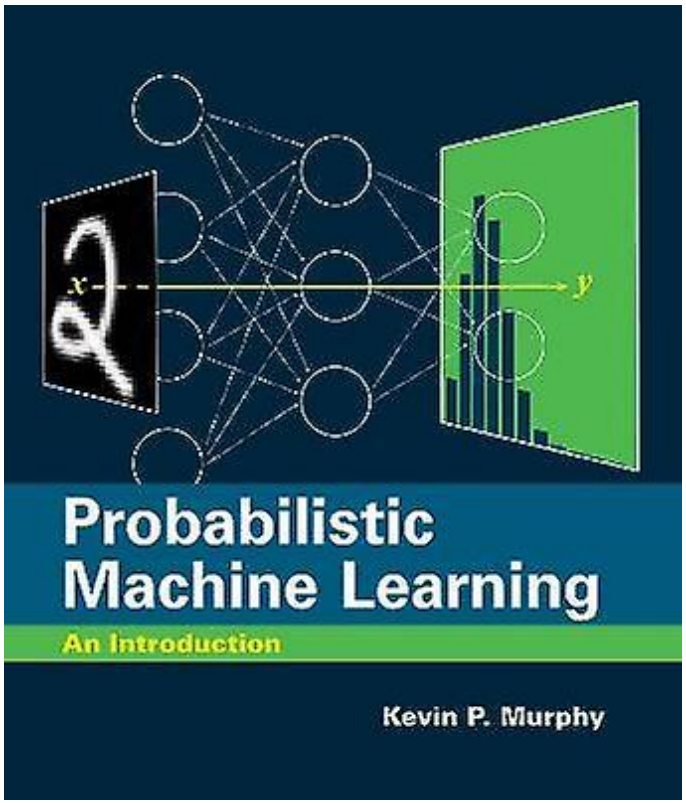


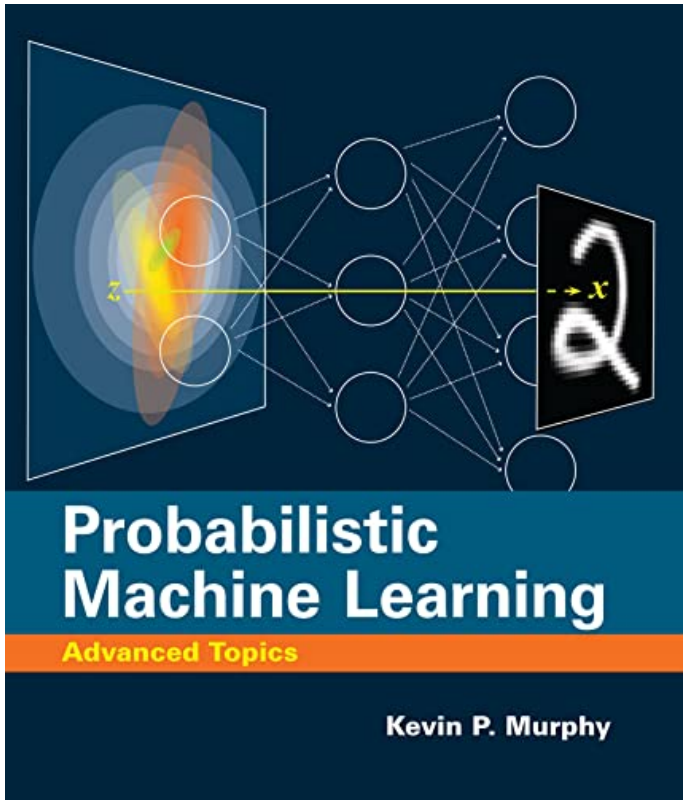
D. Rasch, R. Verdooren and J. Pilz: Applied Statistics - Theory and Problem Solutions with R. Wiley, Oxford 2019.

Perfect textbook interfacing PIT and ML

The perfect textbook for learning about the interface of Probability Theory, Information Theory and Statistical Learning Theory is

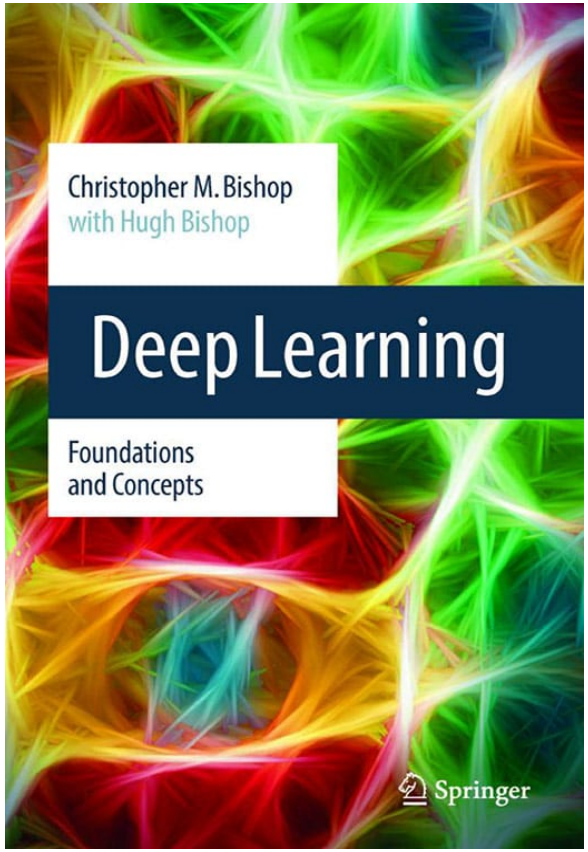
Kevin Murphy's compendium **"Probabilistic Machine Learning"**





K.P. Murphy: Probabilistic Machine Learning. MIT Press 2022, 2023

Deep Learning with solid anchoring in PIT



Ch.M. Bishop, H. Bishop: Deep Learning - Foundations and Concepts. Springer 2024. https://www.bishopbook.com