WuS - Complete Summary

Ruben Schenk, ruben.schenk@inf.ethz.ch

July 4, 2022

Contents

1 Introduction

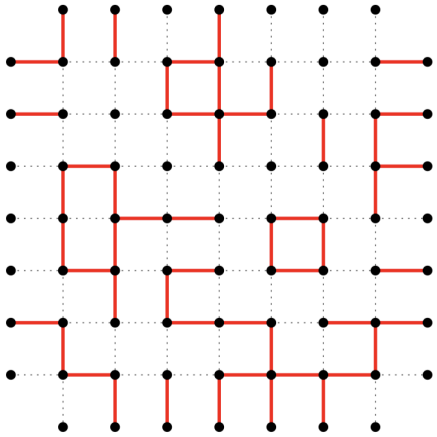
1.1 Percolation Theory

1.1.1 Overview

In physics and mathematics, **percolation theory** describes the behavior of clustered components in random networks. The common intuition is movement and filtering of fluids through porous materials, for example, filtration of water through soil and permeable rocks. In a network, let each node be a cell through which a fluid-like substance may transit to other cells. A network, i.e. a grid, then is a sponge-like substance and percolation is the determination of whether a substance introduced at one cell will reach the other side of the network (or grid).

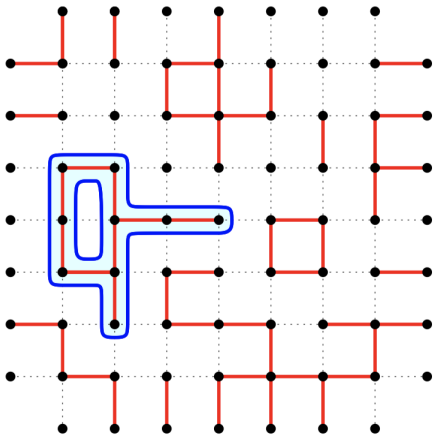
1.1.2 Percolation in a Box

Imagine a box (or grid) with vertices $V = \{-n, ..., n\}^2$ and edges $E = \{e_1, ..., e_N\}$. We introduce parameter p, with $0 \leq p \leq 1$. p denotes the probability that an edge e is *open* ($X_e = 1$). In other words, an edge e is *closed* ($X_e = 0$) with probability $1 - p$. The corresponding model could look something like this:



Note: If an edge is colored red, it means that it's open.

We denote an **open path** as a path consisting of open edges. A **cluster** is the connected component of $(V, \{e : X_e = 1\})$. The following figure shows an example of a cluster (marked in blue):



Theorem [Kesten, 1980]: For the percolation with parameter p we have:

$$\lim_{n\to\infty} \mathbb{P}[\bullet] = \begin{cases} 0, & \text{if } p < \frac{1}{2}, \\ 1, & \text{if } p > \frac{1}{2}. \end{cases}$$

where $\mathbb{P}[\bullet]$ denotes the probability that there exists an open path from the top to the bottom in an $n \times n$ box. Similarly, for the percolation with parameter p we have:

$$\mathbb{P}[\exists \text{ an infinite cluster}] = \begin{cases} 0, & \text{if } p < \frac{1}{2}, \\ 1, & \text{if } p > \frac{1}{2}. \end{cases}$$

1.2 Introduction to Probability

Probability is a mathematical language describing systems involving randomness. Probabilities are used for:

- *Describe random experiments* in the real world, such as coin flips, dice rolling, etc.
- *Express uncertainty.* For example, when a machine performs a measurement, the value is rarely exact. One may use probability theory in this context by saying that the value obtained is equal to the real value plus some small random error.
- *Decision-making.* Probability theory can be used to describe a system when only part of the information is known.
- *Randomized algorithms* in computer science. Sometimes, it is more efficient to add some randomness to perform an algorithm.
- *Simplify complex systems.* Examples include water molecules in water, cars on the highway, etc.

The **goal** of probability theory is to establish general theorems which describe the behavior of multiple random experiments. Example:

Theorem [Law of large numbers]:

$$X_i = \begin{cases} 0, & i^{th} \text{ throw is head}, \\ 1, & i^{th} \text{ throw is number}. \end{cases}$$

It holds, that:

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \frac{1}{2}.$$

2 Mathematical Framework

2.1 Probability Space

2.1.1 Sample Space

Assume we want to model a random experiment. The first mathematical object needed is the set of all possible outcomes of the experiment, denoted by Ω.

The set Ω is called the **sample space.** An element $\omega \in \Omega$ is called an **outcome** (or *elementary experiment*).

Example: If we throw a die, we have the following sample space:

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

2.1.2 Events

Previously, the set of **events** was always $\mathcal{P}(\Omega)$. In this class, we will work with more general sets of events $\mathcal{F} \subset \mathcal{P}(\Omega)$, called sigma algebras.

Definition: A **sigma-algebra** is a subset $\mathcal{F} \subset \mathcal{P}(\Omega)$ satisfying the following properties:

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \implies A^C \in \mathcal{F}$
3. $A_1, A_2, ... \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Example: Following are some (non-) examples of sigma-algebras for $\Omega = \{1, 2, 3, 4, 5, 6\}$:

- $\mathcal{F} = \{\emptyset, \{1, 2, 3, 4, 5, 6\}\}$ is a sigma-algebra.
- $\mathcal{F} = \{\emptyset, \{1, 2\}, \{3, 4, 5, 6\}, \{1, 2, 3, 4, 5, 6\}\}$ is a sigma-algebra.
- $\mathcal{F} = \{\{1, 2, 3, 4, 5, 6\}\}$ is not a sigma-algebra because $P2$ is not satisfied.
- $\mathcal{F} = \{\emptyset, \{1, 2, 3\}, \{4, 5, 6\}, \{1\}, \{2, 3, 4, 5, 6\}, \Omega\}$ is not a sigma-algebra because $P3$ is not satisfied.

2.1.3 Probability Measure

Definition: Let Ω be a sample space, let $\mathcal{F}$ be a sigma-algebra. A **probability measure** on $(\Omega, \mathcal{F})$ is a map

$$\begin{aligned} \mathbb{P} : &\mathcal{F} \to [0, 1] \\ &A \to \mathbb{P}[A] \end{aligned}$$

that satisfies the following two properties:

- **P1.** $\mathbb{P}[\Omega] = 1$.
- **P2. (countable additivity)** $\mathbb{P}[A] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$ if $A = \bigcup_{i=1}^{\infty} A_i$ (*disjoint union*).

2.1.4 Notion of Probability Space

Definition: Let Ω be a sample space, $\mathcal{F}$ a sigma-algebra, and $\mathbb{P}$ a probability measure. The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space.**

2.2 Examples of Probability Space

2.2.1 Example with Ω Finite

We discuss a particular type of probability spaces where the sample space Ω is an arbitrary **finite** set, and all the outcomes have the **same** probability $p_\omega = \frac{1}{|\Omega|}$.

Definition: Let Ω be a finite sample space. The **Laplace model** on Ω is the triple $(\Omega, \mathcal{F}, \mathbb{P})$, where:

- $\mathcal{F} = \mathcal{P}(\Omega)$,
- $\mathbb{P} : \mathcal{F} \to [0, 1]$ is defined by

$$\forall A \in \mathcal{F} \quad \mathbb{P}[A] = \frac{|A|}{|\Omega|}$$

Example: We consider $n \geq 3$ points on a circle, from which we select 2 at random. What is the probability that these two points selected are neighbors? We consider the Laplace model one

$$\Omega = \{E \subset \{1, 2, ..., n\} : |E| = 2\}.$$

The event "the two points of E are neighbors" is given by

$$A = \{\{1, 2\}, \{2, 3\}, ..., \{n-1, n\}, \{n, 1\}\}$$

and we have

$$\mathbb{P}[A] = \frac{|A|}{|\Omega|} = \frac{n}{\binom{n}{2}} = \frac{2}{n-1}.$$

2.2.2 Example with Ω Infinite Countable

Example: We throw a biased coin multiple times, at each throw, the coin falls on head with probability p, and it falls on tail with probability $1-p$ (p is a fixed parameter in $[0, 1]$). We stop at the first time we see a tail. The probability that we stop exactly at time k is given by

$$p_k = p^{k-1}(1-p).$$

For this experiment, one possible probability space is given by:

- $\Omega = \mathbb{N} \setminus \{0\} = \{1, 2, 3, ...\}$
- $\mathcal{F} = \mathcal{P}(\Omega)$
- for $A \in \mathcal{F}$, $\mathbb{P}[A] = \sum_{k \in A} p_k$

2.3 Properties of Events

2.3.1 Operations on Events and Interpretation

The following propositions asserts that the different well-known set operations are allowed.

Proposition (Consequences of the definition): Let $\mathcal{F}$ be a sigma-algebra on Ω. We have:

- **P4.** $\emptyset \in \mathcal{F}$
- **P5.** $A_1, A_2, ... \in \mathcal{F} \implies \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$
- **P6.** $A, B \in \mathcal{F} \implies A \cup B \in \mathcal{F}$
- **P7.** $A, B \in \mathcal{F} \implies A \cap B \in \mathcal{F}$

A short summary of the common set-operations is given below:

- A^C : A does not occur.
- $A \cap B$: A and B occur.
- $A \cup B$: A or B occurs
- $A \Delta B$: one and only one of A or B occurs
- $A \subset B$: If A occurs, then B occurs
- $A \cap B = \emptyset$: A and B cannot occur at the same time
- $\Omega = A_1 \cup A_2 \cup A_3$ with A_1, A_2, A_3 pairwise disjoint: for each outcome ω, one and only one of the events A_1, A_2, A_3 is satisfied.

2.4 Properties of Probability Measures

2.4.1 Direct Consequences of the Definition

Proposition: Let $\mathbb{P}$ be an arbitrary measure on $(\Omega, \mathcal{F})$. We have:

- **P3.** $\mathbb{P}[\emptyset] = 0$.
- **P4. (additivity)** Let $k \geq 1$. let $A_1, ..., A_k$ be k pairwise disjoint events, then $\mathbb{P}[A_1 \cup \cdots \cup A_k] = \mathbb{P}[A_1] + \cdots + \mathbb{P}[A_k]$.
- **P5.** Let A be an event, then $\mathbb{P}[A^C] = 1 - \mathbb{P}[A]$.
- **P6.** If A and B are two events (not necessarily disjoin), then $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$.

2.4.2 Useful Inequalities

Proposition (Monotonicity): Let A, $B \in \mathcal{F}$, then

$$A \subset B \implies \mathbb{P}[A] \leq \mathbb{P}[B].$$

Proposition (Union bound): Let A_1, $A_2, ...$ be a sequence of events (not necessarily disjoint), then we have

$$\mathbb{P}[\bigcup_{i=1}^{\infty} A_i] \leq \sum_{i=1}^{\infty} \mathbb{P}[A_i].$$

Remark: The union bound also applies to a *finite* collection of events.

2.4.3 Continuity Properties of Probability Measures

Proposition: Let (A_n) be an increasing sequence of events (i.e. $A_n \subset A_{n+1}$ for every n). then

$$\lim_{n\to\infty} P[A_n] = \mathbb{P}[\bigcup_{n=1}^{\infty} A_n]. \quad \text{(increasing limit)}$$

Let (B_n) be a decreasing sequence of events (i.e. $B_n \supset B_{n+1}$ for every n). Then

$$\lim_{n\to\infty} P[B_n] = \mathbb{P}[\bigcap_{n=1}^{\infty} B_n]. \quad \text{(decreasing limit)}$$

Remark: By monotonicity, we have $\mathbb{P}[A_n] \leq \mathbb{P}[A_{n+1}]$ and $\mathbb{P}[B_n] \geq \mathbb{P}[B_{n+1}]$ for every n. Hence the limits in the proposition are well defined as monotone limits.

2.5 Conditional Probabilities

Definition (Conditional probability): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space. Let A, B be two events with $\mathbb{P}[B] > 0$. The **conditional probability of A given B** is defined by

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Remark: $\mathbb{P}[B \mid B] = 1$.

Proposition: Let Ω, $\mathcal{F}$, $\mathbb{P}$ be some probability space. Let B be an event with positive probability. Then $\mathbb{P}[\,. \mid B]$ is a probability measure on Ω.

Proposition (Formula of total probability): Let $B_1, ..., B_n$ be a partition of the sample space Ω with $\mathbb{P}[B_i] > 0$ for every $1 \leq i \leq n$. Then, one has

$$\forall A \in \mathcal{F} : \mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A \mid B_i]\mathbb{P}[B_i].$$

Here, a *partition* B_i is such that $\Omega = B_1 \cup \cdots \cup B_n$ and the events are pariwise disjoint.

Proposition (Bayes formula): Let $B_1, ..., B_n \in \mathcal{F}$ be a partition of Ω with $\mathbb{P}[B_i] > 0$ for every i. For every event A with $\mathbb{P}[A] > 0$, we have

$$\forall i = 1, ..., n : \mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[A \mid B_i] \cdot \mathbb{P}[B_i]}{\sum_{j=1}^{n} \mathbb{P}[A \mid B_j] \cdot \mathbb{P}[B_j]}.$$

2.6 Independence

2.6.1 Independence of Events

Definition (Independence of two events): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Two events A and B are said to be **independent** If

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

Remark: If $\mathbb{P}[A] \in \{0, 1\}$, then A is independent of every event, i.e. $\forall B \in \mathcal{F} : \mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B]$. Furthermore we might also state, that A is independent of B if and only if A is independent of B^C.

Proposition: Let $A, B \in \mathcal{F}$ be two events with $\mathbb{P}[A], \mathbb{P}[B] > 0$. Then the following are equivalent:

- $\mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B]$ (A and B are independent)
- $\mathbb{P}[A \mid B] = \mathbb{P}[A]$ (the occurrence of B has no influence on A)
- $\mathbb{P}[B \mid A] = \mathbb{P}[B]$ (the occurrence of A has no influence on B)

Definition: Let I be an arbitrary set of indices. A collection of events $(A_i)_{i \in I}$ is said to be **independent** if

$$\forall J \subset I \text{ infinite} : \mathbb{P}[\bigcap_{j \in J} A_j] = \prod_{j \in J} \mathbb{P}[A_j].$$

3 Random Variables and Distribution Functions

3.1 Abstract Definition

Definition: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A **random variable (r.v.)** is a map $X : \Omega \to \mathbb{R}$ such that for all $a \in \mathbb{R}$,

$$\{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{F}.$$

The condition $\{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{F}$ is needed for $\mathbb{P}[\{w \in \Omega : X(\omega) \leq a\}]$ to be well-defined.

Example (Indicator function of an event): Let $A \in \mathcal{F}$. Consider the **indicator function** $\mathbb{1}_A$ of A, defined by

$$\forall \omega \in \Omega : X(\omega) = \begin{cases} 0 & \text{if } \omega \notin A, \\ 1 & \text{if } \omega \in A. \end{cases}$$

Then $\mathbb{1}_A$ is a random variable. Indeed, we have

$$\{\omega : \mathbb{1}_A(\omega) \leq a\} = \begin{cases} \emptyset & \text{if } a < 0, \\ A^C & \text{if } 0 \leq a \leq 1, \\ \Omega & \text{if } a \geq 1, \end{cases}$$

and $\emptyset$, A^C, and Ω are three elements of $\mathcal{F}$.

Notation: When events are defined in terms of random variables, we will *omit the dependence in* ω. For example, for $a \leq b$ we write:

$$\{X \leq a\} = \{\omega \in \Omega : X(\omega) \leq a\},$$
$$\{a < X \leq b\} = \{\omega \in \Omega : aX(\omega) < b\},$$
$$\{X \in \mathbb{Z}\} = \{\omega \in \Omega : X(\omega) \in \mathbb{Z}\}$$

When considering the probability of the events above, we omit the brackets and, for example, simply write:

$$\mathbb{P}[X \leq a] = \mathbb{P}[\{X \leq a\}] = \mathbb{P}[\{\omega \in \Omega : X(\omega) \leq a\}].$$

3.2 Distribution Function

Definition: Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The **distribution function** of X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$\forall a \in \mathbb{R} : F_X(a) = \mathbb{P}[X \leq a]$$

The idead is that the distribution function F_X encodes the probabilistic properties of the random variable X.

Proposition (Basic identity): Let $a < b$ be two real numbers. Then

$$\mathbb{P}[a < X \leq b] = F(b) - F(a)$$

Theorem (Properties of distribution functions): Let X be a random variable on some probability space $(\Omega, \mathcal{F}), \mathbb{P}$. The distribution function $F = F_X : \mathbb{R} \to [0, 1]$ of X satisfies the following properties:

1. F is nondecreasing.
2. F is right continuous, i.e. $F(a) = \lim_{h \downarrow 0} F(a + h)$ for every $a \in \mathbb{R}$.
3. $\lim_{a \to -\infty} F(a) = 0$ and $\lim_{a \to \infty} F(a) = 1$.

3.3 Independence

3.3.1 Independence of Random Variables

Definition: Let $X_1, ..., X_n$ be n random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that $X_1, ..., X_n$ are **independent** if

$$\forall x_1, ..., x_n \in \mathbb{R} : \mathbb{P}[X_1 \leq x_1, ..., X_n \leq x_n] = \mathbb{P}[X_1 \leq x_1] \cdots \mathbb{P}[X_n \leq x_n].$$

Definition: An infinite sequence $X_1, X_2, ...$ of random variables is said to be:

- **independent** ix $X_1, ..., X_n$ are independent, for every n.
- **independent and identically distributed (iid)** if they are independent and have the same distribution function, i.e. $\forall i,\, y : F_{X_i} = F_{X_j}$.

3.4 Transformation of Random Variables

Once we have some random variables X_1, $X_2, ...$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we can create and consider many new random variables on the same probability space by using operations. For example, one can consider $Z_1 = X_1 + X_2$. However, one should not forget that random variables are maps $\Omega \to \mathbb{R}$. For example, the random variable Z_1 corresponds to the map, defined for every $\omega \in \Omega$, $Z_1(\omega) = X_1(\omega) + X_2(\omega)$.

Formally, we introduce the following notation, which allows us to work with random variables as if they were just real numbers. If X is the random variable, and $\phi : \mathbb{R} \to \mathbb{R}$, then we write

$$\phi(X) := \phi \circ X.$$

This way, $\phi(X)$ is a new mapping $\Omega \to \mathbb{R}$ as show in the following diagram:

$$\begin{aligned} &\Omega \to^X \mathbb{R} \to^\phi \mathbb{R} \\ &\omega \to X(\omega) \to \phi(X(\omega)). \end{aligned}$$

3.5 Construction of Random Variables

The goal of this section is to construct general random variables. Our approach will rely on the abstract theorem of Kolmogorov, that guarantees existences of iid sequences. The construction proceeds in 4 steps:

Step 1: Komogorov theorem and iid sequence of Bernoulli random variables Our construction starts with Bernoulli random variables, that we define now.

Definition: Let $p \in [0, 1]$. A random variable X is said to be a **Bernoulli random variable with parameter** p if

$$\mathbb{P}[X = 0] = 1 - p \text{ and } \mathbb{P}[X = 1] = p.$$

In this case, we write $X \sim \text{Ber}(p)$.

> **Theorem (Existence theorem of Kolmogorov):** There exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an infinite sequence of random variables $X_1, X_2, ...$ (on this probability space) that is an iid sequence of Bernoulli random variables with parameter $\frac{1}{2}$.

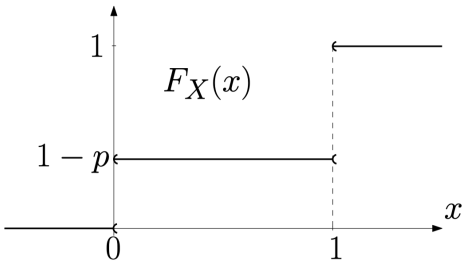
Step 2: Construction of a uniform random variable in $[0, 1]$ Here we use Bernoulli random variables to construct a uniform random variable in $[0, 1]$. Intuitively, one can imagine a droplet of water falling in the interval $[0, 1]$. A uniform random variable in $[0, 1]$ represents the position at which such a droplet falls.

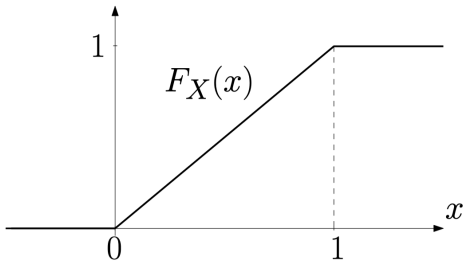
Definition: A random variable U is said to be a **uniform random variable in** $[0, 1]$ if its distribution function is equal to

$$F_U(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \le x \le 1, \\ 1, & x > 1. \end{cases}$$

In this case, we write $U \sim \mathcal{U}([0, 1])$.

The figure below shows the distribution function of a Bernoulli r.v. with parameter p (left) and the distribution function of a uniform random variable in $[0, 1]$ (right).





Let $X_1, X_2, ...$ be a sequence of independent Bernoulli random variables with parameter $\frac{1}{2}$. For every fixed ω, we have $X_1(\omega), X_2(\omega), ... \in \{0, 1\}$. Hence the infinite series

$$Y(\omega) = \sum_{n=1}^{\infty} 2^{-n} X_n(\omega)$$

is absolutely convergent, and we have $Y(\omega) \in [0, 1]$.

Proposition: The mapping $Y : \Omega \to [0, 1]$ defined by the equation above is a uniform random variable in $[0, 1]$.

Step 3: Construction of a random variable with an arbitrary distribution F Let $F : \mathbb{R} \to [0, 1]$ satisfying item $(1) - (3)$ at the beginning of the section. If F is strictly increasing and continuous then F is one-to-one and one can define its inverse F^{-1}. For every $\alpha \in [0, 1]$, $F^{-1}(\alpha)$ is the unique real number x such that $F(x) = \alpha$. In such a case, the defines the inverse distribution function. More generally, we can define a generalized inverse for F.

Definition (Generalized inverse): The generalized inverse of F is the mapping $F^{-1} : (0, 1) \to \mathbb{R}$ defined by

$$\forall \alpha \in (0, 1) : F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

By definition of the infimum and using right continuity of F, we have for every $x \in \mathbb{R}$ and $\alpha \in (0, 1)$

$$(F^{-1}(\alpha) \leq x) \iff (\alpha \leq F(x)).$$

Theorem (inverse transform sampling): Let $F : \mathbb{R} \to [0, 1]$ satisfying items $(1) - (3)$ at the beginning of the section. Let U be a uniform random variable in $[0, 1]$. Then the random variable

$$X = F^{-1}(U)$$

has distribution $F_X = F$.

Step 4: General sequence of independent random variables Finally, we introduce the following theorem:

Let $F_1, F_2, ...$ be a sequence of functions $\mathbb{R} \to [0, 1]$ satisfying items $(1) - (3)$ at the beginning of the section. Then there exists a probability space $(\Omega, \mathcal{F}. \mathbb{P})$ and a sequence of independent random variables $X_1, X_2, ...$ on this probability space such that

- for every i X_i has a distribution function F_i (i.e. $\forall x \mathbb{P}[X_i \leq x] = F_i(x)$), and
- $X_1, X_2, ...$ are independent.

4 Discrete and Continuous Random Variables

4.1 Discontinuity & Continuity Points of F

We have seen that the distribution function $F = F_X$ of a random variable X is always *rigt continuous.* What about left continuous?

Example: For a Bernoulli random variable $X \sim \text{Ber}(p)$ with $p < 1$, we have $F_X(-h) = 0$ for every $h > 0$, but $F_X(0) = 1 - p \neq 0$. Therefore, F_X is not left continuous at 0, i.e.

$$\lim_{h \downarrow 0} F_X(-h) = 0 \neq F_X(0).$$

The following proposition gives an interpretation of the limit

$$F(a-) := \lim_{h \downarrow 0} F(a-h)$$

at a given point a for a general distribution function.

Proposition (probability of a given value): Let $X : \Omega \to \mathbb{R}$ be a random variable with distribution function F. Then for every a in $\mathbb{R}$ we have

$$\mathbb{P}[X = a] = F(a) - F(a-).$$

We give the following interpretation of the above introduce proposition. Fix some $a \in \mathbb{R}$. Then:

- If F is not continuous at a point $a \in \mathbb{R}$, then the "jump size" $F(a) - F(a-)$ is equal to the probability that $X = a$.
- If F is continuous at a point $a \in \mathbb{R}$, then $\mathbb{P}[X = a] = 0$.

4.2 Almost Sure Events

Definition: Let $A \in \mathcal{F}$ be an event. We say that A occurs **almost surely (a.s.)** if

$$\mathbb{P}[A] = 1.$$

Remark: This notion can be extended to any set $A \subset \Omega$: We say that A occurs almost surely if there exists an event $A' \in \mathcal{F}$ such that $A' \subset A$ and $\mathbb{P}[A'] = 1$.

4.3 Discrete Random Variables

Definition (Discrete Random Variables): A random variable $X : \Omega \to \mathbb{R}$ is said to be **discrete** if there exists some set $W \subset \mathbb{R}$ finite or countable such that

$$X \in W \quad a.s.$$

Remark: If the sample space Ω is finite or countable, then every random variable $X : \Omega \to \mathbb{R}$ is discrete.

Definition: Let X be a discrete random variable taking some values in some finite or countable set $W \subset \mathbb{R}$. The **distribution of** X is the sequence of numbers $(p(x))_{x \in W}$ defined by

$$\forall x \in W : p(x) := \mathbb{P}[X = x].$$

Proposition: The distribution $(p(x))_{w \in W}$ of a discrete random variable satisfies

$$\sum_{x \in W} p(x) = 1.$$

Example: Consider the random variable defined by

$$\forall \omega \in \Omega : X(\omega) := \begin{cases} -1, & \text{if } \omega = 1, 2, 3, \\ 0, & \text{if } \omega = 4, \\ 2, & \text{if } \omega = 5, 6. \end{cases}$$

Then X takes values in $W = \{-1, 0, 2\}$ almost surely and its distribution is given by

$$p(-1) = \frac{1}{2}, \quad p(0) = \frac{1}{6}, \quad p(2) = \frac{1}{3}.$$

Remark: Conversely, if we are given a sequence of numbers $(p(x))_{x \in W}$ with values in $[0, 1]$ and such that $\sum_{x \in W} p(x) = 1$, then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X with associated distribution $(p(x))$. This observation is important in practice, it allows us to write: "*Let X be a discrete random variable with distribution $(p(x))_{x \in W}$.*"

4.3.1 From p to F_X

Proposition: Let X be a discrete random variable with values in a finite or countable set W almost surely, and distribution p. Then the distribution function of X is given by

$$\forall x \in \mathbb{R} : F_X(x) = \sum_{y \leq x,\, y \in W} p(y).$$

4.3.2 From F_X to p

Given a discrete random variable X. A random variable with a piecewise cosntant function F is discrete and W and p are given by:

- $W = \{\text{positions of the jumps of } F_X\}$
- $p(x) =$ "height of the jump" at $x \in W$

4.4 Examples of Discrete Random Variables

4.4.1 Bernoulli Distribution

Definition (Bernoulli): Let $0 \leq p \leq 1$. A random variable X is said to be a **Bernoulli random variable with parameter** p if it takes values in $W = \{0, 1\}$ and

$$\mathbb{P}[X = 0] = 1 - p \quad \text{and} \quad \mathbb{P}[X = 1] = p.$$

In that case, we write $X \sim \text{Ber}(p)$.

4.4.2 Binomial Distribution

Definition (Binomial): Let $0 \leq p \leq 1$, let $n \in \mathbb{N}$. A random variable X is said to be a **binomial random variable with parameters** n **and** p if it takes values in $W = \{0, ..., n$ and

$$\forall k \in \{0, ..., n\} : \mathbb{P}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}.$$

In that case we write $X \sim \text{Bin}(n, p)$. This appears in applications when we consider the number of successes in a repetition of Bernoulli experiments.

Proposition (Sum of independent Bernoulli and binomial): Let $0 \leq p \leq 1$, let $n \in \mathbb{N}$. Let $X_1, ..., X_n$ be independent Bernoulli random variables with parameter p. Then

$$S_n := X_1 + \cdots + X_n$$

is a binomial random variable with parameter n and p.

Remark: In particular, the distribution Bin(1, p) is the same as the distribution Ber(p). One can also check that if $X \sim \text{Bin}(m, p)$ and $Y \sim \text{Bin}(n, p)$ and X, Y are independent, then $X + Y = \text{Bin}(m+n, p)$.

4.4.3 Geometric Distribution

Definition (Geometric): Let $0 \leq p \leq 1$. A random variable X is said to be a **geometric random variable with parameter** p if it takes values in $W = \mathbb{N} \setminus \{0\}$ and

$$\forall k \in \mathbb{N} \setminus \{0\} : \mathbb{P}[X = k] = (1-p)^{k-1} \cdot p.$$

In that case, we write $X \sim \text{Geom}(p)$.

The geometric random variable appears naturally as the first success in an infinite sequence of Bernoulli experiments with parameter p. This is formalized by the following proposition.

Proposition: Let $X_1, X_2, ...$ be a sequence of infinitely many independent Bernoulli r.v.'s with parameter p. Then

$$T := \min\{n \geq 1 : X_n = 1\}$$

is a geometric random variable with parameter p.

Proposition: Let $T \sim \text{Geom}(p)$ for some $0 < p < 1$. Then

$$\forall n \geq 0, \forall k \geq 1 : \mathbb{P}[T \geq n + k \,|\, T > n] = \mathbb{P}[T \geq k].$$

4.4.4 Poisson Distribution

Definition: Let $\lambda > 0$ be a positive real number. A random variable X is said to be a **Poisson random variable with parameter** λ if it takes values in $W = \mathbb{N}$ and

$$\forall k \in \mathbb{N} : \mathbb{P}[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}.$$

In this case, we write $X \sim \text{Poisson}(\lambda)$.

The Poisson distribution appears naturally as an approximation of a binomial distribution when the parameter n is large and the parameter p is small, as stated formally in the following proposition.

Proposition (Poisson approximation of the binomial): Let $\lambda > 0$. For every $n \geq 1$, consider a random variable $X_n \sim \text{Bin}(n, \frac{\lambda}{n})$. Then

$$\forall k \in \mathbb{N} : \lim_{n \to \infty} \mathbb{P}[X_n = k] = \mathbb{P}[N = k],$$

where N is a Poisson random variable with parameter λ.

4.5 Continuous Random Variables

Definition (Continuous Random Variables): A random variable $X : \Omega \to \mathbb{R}$ is said to be **continuous** if its distribution function F_X can be written as

$$F_X(a) = \int_{-\infty}^{a} f(x)\, dx \quad \text{for all } a \in \mathbb{R}$$

for some nonnegative function $f : \mathbb{R} \to \mathbb{R}_+$, called the **density** of X.

Intuition: $f(x)\,dx$ represents the probability that X takes a value in the infinitesimal interval $[x,\, x+dx]$.

Proposition: The density f of a random variable satisfies

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

4.5.1 From f to F_X

Let X be a continuous random variable with density f. By definition, the distribution function F_X can be calculated as the integral

$$F_X(x) = \int_{-\infty}^{x} f(y)\,dy.$$

4.5.2 From F_X to f

Since one goes from f to F_X by integrating, it is natural to expect that the reverse operation is to take the derivative. This is in general the case, provided F_X is regular enough. The following theorem will be useful in applications to calculate densities.

Theorem: Let X be a random variable. Assume that the distribution function F_X is continuous and piecewise $\mathcal{C}^1$, i.e. that there exists $x_0 = -\infty < x_1 < \cdots < x_{n-1} < x_n = +\infty$ such that F_X is $\mathcal{C}^1$ on every interval $(x_i,\, x_{i+1})$. Then X is a continuous random variable and a density f can be constructed by defining

$$\forall x \in (x_i,\, x_{i+1}) : f(x) = F_X'(x)$$

and setting arbitrary values at $x_1, ..., x_{n-1}$.

4.6 Examples of Continuous Random Variables

4.6.1 Uniform Distributions

Definition (Uniform distribution in $[a,\, b]$, $a < b$): A coninuous random variable X is said to be **uniform in** $[a,\, b]$ if its density is equal to

$$f_{a,\,b}(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,\, b] \\ 0 & \text{otherwise} \end{cases}$$

In this case, we write $X \sim \mathcal{U}([a,\, b])$.

Intuition: X represents a uniformly chosen point in the interval $[a,\, b]$.

Properties:

- The probability to fall in an interval $[c,\, c+l] \subset [a,\, b]$ depends only on its length l:

$$\mathbb{P}[X \in [c,\, c+l]] = \frac{l}{b-a}.$$

- The distribution function of X is equal to:

$$F_X(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } x \in [a,\, b], \\ 1 & \text{if } x > b. \end{cases}$$

4.6.2 Exponential Distribution

The exponential distribution is the continuous analogue of the geometric distribution.

Definition (Exponential distribution with $\lambda > 0$: A continuous random variable T is said to be **exponential with parameter** $\lambda > 0$ if its density is equal to:

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & x > 0. \end{cases}$$

In that case, we write $T \sim \text{Exp}(\lambda)$.

Intuition: T represents the time of a "clock ring". For example, the time at which the first customer arrives in a shop is well modeled by an exponential random variable.

Properties:

- The waiting probability is exponentially small:
$$\forall t \geq 0 : \mathbb{P}[T > t] = e^{-\lambda t}.$$
- It has the absence of memory property:
$$\forall t, s \geq 0 : \mathbb{P}[T > t + s \,|\, T > t] = \mathbb{P}[T > s].$$

4.6.3 Normal Distribution

Definition: A continuous random variable X is said to be **normal with parameters** m **and** $\sigma^2 > 0$ if its density is equal to:

$$f_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

In that case, we write $X \sim \mathcal{N}(m, \sigma^2)$.

Properties:

- If $X_1, ..., X_n$ are independent random variables with parameters $(m_1, \sigma_1^2), ..., (m_n, \sigma_n^2)$ respectively, then
$$Z = m_0 + \lambda_1 X_i + \cdots \lambda_n X_n$$
is a normal random variable with parameters $m = m_0 + \lambda_1 m_1 + \cdots + \lambda_n m_n$ and $\sigma^2 = \lambda_1^2\sigma_1^2 + \cdots \lambda_n^2\sigma_n^2$.
- In particular, if $X \sin \mathcal{N}(0, 1)$ (in this case we say that X is a **standard normal random variable**), then
$$Z = m + \sigma \cdot X$$
is a normal random variable with parameters m and σ^2.

5 Expectation

5.1 Expectation for General Random Variables

Definition: Let $x : \Omega \to \mathbb{R}_+$ be a random variable with nonnegative values. The **expectation** of X is defined as

$$\mathbb{E}[X] = \sum_0^\infty (1 - F_X(x))\, dx.$$

Proposition: Let X be a nonnegative random variable. Then we have

$$\mathbb{E}[X] \geq 0$$

with equality if and only if $X = 0$ almost surely.

Definition: Let X be a random variable. If $E[|X|] < \infty$, then the expectation of X is defined by

$$\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-].$$

5.2 Expectation of a Discrete Random Variable

Proposition: Let $X : \Omega \to \mathbb{R}$ be a discrete random variable with values in W (finite or countable) almost surely. We have

$$\mathbb{E}[X] = \sum_{x \in W} x \cdot \mathbb{P}[X = x],$$

provided the sum is well defined.

Example 1 (Bernoulli): Let X be a Bernoulli random variable with parameter p. We have

$$\mathbb{E}[X] = p.$$

Example 2 (Poisson): Let X be a Poisson random variable with parameter $\lambda > 0$, then

$$\mathbb{E}[X] = \lambda.$$

Definition: Let $A \in \mathcal{F}$ be an event. Consider the **indicator function** $\mathbb{1}_A$ of A, defined by

$$\forall \omega \in \Omega : \mathbb{1}_A(\omega) = \begin{cases} 0 & \text{if } \omega \notin A, \\ 1 & \text{if } \omega \in A. \end{cases}$$

Then $\mathbb{1}_A$ is a random variable. Ineed, we have:

$$\{\mathbb{1}_A \leq a\} = \begin{cases} \emptyset & \text{if } a > 0, \\ A^C & \text{if } 0 \leq a < 1, \\ \Omega & \text{if } a \geq 1, \end{cases}$$

and $\emptyset$, A^C, Ω are three elements of $\mathcal{F}$. Furthermore, writing $X = \mathbb{1}_A$, we have

$$\mathbb{P}[X = 0] = 1 - \mathbb{P}[A] \quad \text{and} \quad \mathbb{P}[X = 1] = \mathbb{P}[A].$$

Therefore, $\mathbb{1}_A$ is a Bernoulli random variable with parameter $\mathbb{P}[A]$. Hence,

$$\mathbb{E}[\mathbb{1}_A] = \mathbb{P}[A].$$

Proposition: Let $X : \Omega \to \mathbb{R}$ be a discrete random variable with values in W (finite or countable) almost surely. For every $\phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}[\phi(X)] = \sum_{w \in W} \phi(x) \cdot \mathbb{P}[X = x],$$

provided the sum is well defined.

5.3 Expectation of a Continuous Random Variable

Proposition: Let X be a continuous random variable with density f. Then, we have

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

provided the integral is well defined.

Example 1 (Uniform): We have

$$\mathbb{E}[X] = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\cdot(\frac{1}{2}b^2 - \frac{1}{2}a^2).$$

Therefore,

$$\mathbb{E}[X] = \frac{a+b}{2}.$$

Example 2 (Exponential): By integration by parts, we have

$$\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x}\,dx = [-xe^{-\lambda x}]_0^\infty + \int_0^\infty e^{-\lambda x}\,dx.$$

Therefore,

$$\mathbb{E}[X] = \frac{1}{\lambda}.$$

Proposition: Let X be a continuous random variable with density f. Let $\phi : \mathbb{R} \to \mathbb{R}$ be such that $\phi(X)$ is a random variable. Then we have

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

provided the integral is well defined.

5.4 Calculus

Theorem (Linearity of the expectation): Let $X, Y : \Omega \to \mathbb{R}$ be random variables, let $\lambda \in \mathbb{R}$. Provided the expectations are well defined, we have:

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

Application 1 (Binomial): Let S be a binomial random variable with parameters n and p. By definition we have

$$\mathbb{E}[S] = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k(1-p)^{n-k}.$$

By linearity we have $\mathbb{E}[S_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]$, where $X_1, ..., X_n$ are n i.i.d. Bernoulli random variables. Using that $\mathbb{E}[X_i] = p$ for every p, we deduce directly

$$\mathbb{E}[S] = \mathbb{E}[S_n] = np.$$

Application 2 (Normal): By Proposition we have (with $Y \sim \mathcal{N}(0, 1)$)

$$\mathbb{E}[X] = \mathbb{E}[m + \sigma \cdot Y] = m + \sigma \cdot \mathbb{E}[Y],$$

hence it suffices to compute the expectation of Y. Writing $f_{0,1}$ for the density of Y, we have

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} x \cdot f_{0,1}(x)\,dx = 0$$

because $x \cdot f_{0,1}(x)$ is an odd function. Finally, we obtain

$$\mathbb{E}[X] = m.$$

Theorem: Let X, Y be two random variables. If X and Y are *independent,* then

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

5.5 Characterizations via Expectations

5.5.1 Density

Proposition: Let X be a random variable. Let $f : \mathbb{R} \to \mathbb{R}_+$ such that $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Then the following are equivalent:

1. X is continuous with density f.
2. For every function $\phi : \mathbb{R} \to \mathbb{R}$ measurable bounded,

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

5.5.2 Independence

Theorem: Let X, Y be 2 discrete random variables. Then the following two are equivalent:

1. X, Y are independent.
2. For every $\phi : \mathbb{R} \to \mathbb{R}$, $\psi : \mathbb{R} \to \mathbb{R}$ (measurable) bounded

$$\mathbb{E}[\phi(X)\psi(Y)] = \mathbb{E}[\phi(X)]\mathbb{E}[\psi(Y)].$$

5.6 Ungleichungen

5.6.1 Monotonie

Satz: Seien X, Y zwei Z.V., sodass

$$X \leq Y \; f.s.$$

gilt. Falls beide Erwartungswerte wohldefiniert sind foglt dann

$$\mathbb{E}[X] \leq \mathbb{E}[Y]. \; f.s.$$

5.6.2 Markov Ungleichung

Theorem (Markov-Ungleichung): Sei X eine nicht-negative Z.V. Für jedes $a > 0$ gilt dann

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

5.6.3 Jensen Ungleichung

Theorem (Jensen Ungleichung): Sei X eine Z.V. Sei $\phi : \mathbb{R} \to \mathbb{R}$ eine konvexe Funktion. Falls $\mathbb{E}[\phi(x)]$ und $\mathbb{E}[X]$ wohldefiniert sind, gilt

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

Daraus folgt mit $\phi(x) = |x|$, dass $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$ (Dreiecksungleichung) und mit $\phi(x) = x^2$, dass $\mathbb{E}[|X|] \leq \sqrt{\mathbb{E}[X^2]}$.

5.7 Varianz

Def: Sei X eine Zufallsvariable, sodass $\mathbb{E}[X]^2 < \infty$. Wir definieren die **Varianz von** X durch

$$\sigma_X^2 = \mathbb{E}[(X-m)^2], \text{ wobei } m = \mathbb{E}[X].$$

Die Wurzel aus σ_X^2 nennen wir gerade die **Standardabweichung von** X.

Die Standardabweichung ist ein Indikator für die Fluktuation von X um den *Mittelwert* $m = \mathbb{E}[X]$ herum. Allgemein ist eine Zufallsvariable mit geringer Varianz konzentriert um ihren Erwartungswert $m = \mathbb{E}[X]$. Die Tschebyscheffsche Ungleichung formalisiert diese Beobachtung.

Satz: Sei X eine Z.V. mit $\mathbb{E}[X^2] < \infty$. Dann gilt für jedes $a \geq 0$

$$\mathbb{P}[|X-m| \geq a] \leq \frac{\sigma_X^2}{a^2}, \text{ wobei } m = \mathbb{E}[X].$$

Satz (Grundlegende Eigenschaften der Varianz):

1. Sei X eine Z.V. mit $\mathbb{E}[X^2] < \infty$. Dann gilt
$$\sigma_X^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$
2. Sei X eine Z.V. mit $\mathbb{E}[X^2] < \infty$ und sei $\lambda \in \mathbb{R}$. Dann gilt
$$\sigma_{\lambda X}^2 = \lambda^2 \cdot \sigma_X^2.$$
3. Seien $X_1, ..., X_n$ n-viele paarweise unabhängige Z.V. und $S = X_1 + \cdots + X_n$. Dann gilt
$$\sigma_S^2 = \sigma_{X_1}^2 + \cdots + \sigma_{X_n}^2.$$

Anwendung: Sei S eine binomialverteilte Z.V. mit Parametern n und p. Was ist die Varianz von S? Wir benutzen, dass S die selbe Verteilung wie $S_n = X_1 + \cdots X_n$, mit $X_1, ..., X_n$ u.i.v. Bernoulli Z.V. mit Parameter p hat. Dann erhalten wir:

$$\sigma_S^2 = \sigma_{S_n}^2 = \sigma_{X_1}^2 + \cdots + \sigma_{X_n}^2 = n \cdot \sigma_{X_1}^2.$$

Zudem gilt $\sigma_{X_1}^2 = \mathbb{E}[X_i^2] - p^2 = p - 2 = p(1-p)$. Durch Einsetzen erhalten wir:

$$\sigma_S^2 = n \cdot p(1-p).$$

Im Allgemeinen erhalten wir für Summen von u.i.v. Z.V. stets $\mathbb{E}[S] = n \cdot p$ und $\sigma_S = \sqrt{n} \cdot \sqrt{p(1-p)}$.

5.8 Kovarianz

Def: Seien X, Y zwei Z.V. mit endlichen zweiten Momenten $\mathbb{E}[X^2] < \infty$ und $\mathbb{E}[Y^2] < \infty$. Wir definieren die **Kovarianz zwischen** X **und** Y durch

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

Die Kovarianz verschwindet wenn X und Y unabhängig sind, somit gilt:

$$X, Y \text{ unabhängig} \implies \mathrm{Cov}(X, Y) = 0.$$

Achtung: Die umgekehrte Implikation ist falsch!

6 Gemeinsame Verteilungen

6.1 Gemeinsame diskrete Verteilungen

6.1.1 Definition

Def: Seien $X_1, ..., X_n$ n diskrete Zufallsvariablen, sei $W_i \subset \mathbb{R}$ endlich oder abzählbar, wobei $X_i \in W_i$ fast sicher gilt. Die gemeinsame Verteilung von $(X_1, ..., X_n)$ ist eine Familie $p = (p(x_1, ..., x_n))_{x_1 \in W_1, ..., x_n \in W_n}$, wobei jedes Mitglied definiert ist durch:

$$p(x_1, ..., x_n) = \mathbb{P}[X_1 = x_1, ..., X_n = x_n].$$

Beispiel: Seien X, Y zwei unabhängige Bernoulli Z.V. mit Parameter $1/2$. Die gemeinsame Verteilung von (X, Y) ist gegeben durch

$$\forall x, y \in \{0, 1\} \quad p(x, y) = \frac{1}{4}.$$

Die gemeinsame Verteilung von (X, X) ist gegeben durch

$$\forall x, y \in \{0, 1\} \quad p(x, y) = \begin{cases} \frac{1}{2}, & x = y \\ 0, & x \neq y. \end{cases}$$

Satz: Eine gemeinsame Verteilung von Z.V. $X_1, ..., X_n$ erfüllt

$$\sum_{x_1 \in W_1, ..., x_n \in W_n} p(x_1, ..., x_n) = 1.$$

Bemerkung: Man kann auch **Gewichtsfunktion** anstatt Verteilung sagen.

6.1.2 Randverteilung

Unter Kenntnis der Verteilung von $X_1, ..., X_n$ kann man die Verteilung der einzelnen X_i separat ermitteln. In diesem Zusammenhang wird die Verteilung von X_i als i-te **Randverteilung** bezeichnet.

Satz: Seien $X_1, ..., X_n$ diskrete Z.V. mit gemeinsamer Verteilung $p = (p(x_1, ..., x_n))_{x_i \in W_1, ..., x_m \in W_n}$. Für jedes i gilt:

$$\forall z \in W_i \quad \mathbb{P}[X_i = z] = \sum_{x_1, ..., x_{i-1}, x_{i+1}, ..., x_n} p(x_1, ..., x_{i-1}, z, x_{i+1}, ..., x_n).$$

6.1.3 Unabhängigkeit

Satz: Seien $X_1, ..., X_n$ diskrete Zufallsvariablen mit gemeinsamer Verteilung $p = (p(x_1, ..., x_n))_{x_1 \in W_1, ..., x_n \in W_n}$. Die folgenden Aussagen sind äquivalent:

1. $X_1, ..., X_n$ sind unabhängig.
2. $p(x_1, ..., x_n) = \mathbb{P}[X_1 = x_1] \cdots \mathbb{P}[X_n = x_n]$ für jedes $x_1 \in W_1, ..., x_n \in W_n$.

6.2 Stetige Gemeinsame Verteilung

6.2.1 Definition

Def: Sei $n \geq 1$. Wir sagen, dass die Z.V. $X_1, ..., X_n : \Omega \to \mathbb{R}$ eine **stetige gemeinsame Verteilung** besitzen, falls eine Abbildung $f : \mathbb{R}^n \to \mathbb{R}_+$ existiert, sodass

$$\mathbb{P}[X_1 \leq a_1, ..., X_n \leq a_n] = \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_n} f(x_1, ..., x_n)\, dx_n ... dx_1$$

für jedes $a_1, ..., a_n \in \mathbb{R}$ gilt. Obige Abbildung f nennen wir gerade **gemeinsame Dichte von** $(X_1, ..., X_n)$.

Satz: Sei f die gemeinsame Dichte der Zufallsvariablen $(X_1, ..., X_n)$. Dann gilt

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, ..., x_n)\, dx_n ... dx_1 = 1.$$

Intuition: Nehmen wir zum Beispiel zwei Z.V. X, Y. Intuitiv beschreibt $f(x, y)\, dxdy$ die Wahrscheinlichkeit, dass ein Zufallspunkt (X, Y) in einem Rechteck $[x, x + dx] \times [y, y + dy]$ liegt.

6.2.2 Erwartungswert unter Abbildungen

Satz: Sei $\phi : \mathbb{R}^n \to \mathbb{R}$ eine Abbildung. Falls $x_1, ..., X_n$ eine gemeinsame Dichte f besitzen, dann lässt sich der Erwartungswert der Z.V. $Z = \phi(X_1, ..., X_n)$ mittels

$$\mathbb{E}[Z] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi(x_1, ..., x_n) \cdot f(x_1, ..., x_n)\, dx_1 ... dx_n,$$

berechnen (solange das Integral wohldefiniert ist).

Beispiel: Betrachten wir das Paar (X, Y) analog zum obigen Beispiel. Falls wir die Funktion $\phi(x, y) = \mathbb{1}_{(x,y)\in R}$ betrachten, gilt für jedes Rechteck $R = (a, a') \times (b, b') \subseteq [0, 1]^2$:

$$\mathbb{P}[(X, Y) \in R] = \mathbb{E}[\phi(X, Y)] = \int_a^{a'} \int_b^{b'} dxdy = (a' - a)(b' - b) = \text{Flaeche}(R).$$

6.2.3 Randverteilungen

Falls $X.Y$ eine gemeinsame Dichte $f_{X,Y}$ besitzt, dann gilt

$$\begin{aligned}\mathbb{P}[X \leq a] &= \mathbb{P}[X \in [-\infty, a], Y \in [-\infty, \infty]] \\ &= \int_{-\infty}^{a} \Big(\int_{-\infty}^{\infty} f(x, y)\, dy \Big) dx.\end{aligned}$$

Somit is X stetig mit folgender Dichte:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy.$$

Analog ist Y stetig mit folgender Dichte:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx.$$

Bemerkung: Folgende Implikationen gelten:

$$X, Y \text{ diskrete Z.V.} \iff X, Y \text{ gemeinsame diskrete Z.V.}$$

$$X, Y \text{ gemeinsam stetig} \implies X \text{ stetig und } Y \text{ stetig.}$$

Beispiel: Schauen wir uns die Gleichverteilung eines Punktes auf einem Quadrat an. Unter gemeinsamer Dichte $f_{X,Y}(x, y) = \mathbb{1}_{0\leq x,y\leq 1}$ hat X folgende Dichte:

$$f_X(x) = \int_0^1 \mathbb{1}_{0\leq x\leq 1}\mathbb{1}_{0\leq y\leq 1}\, dy = \mathbb{1}_{0\leq x\leq 1}.$$

Analog ist $f_Y(y) = \mathbb{1}_{0\leq y\leq 1}$. Somit sind sowohl X als auch Y gleichverteilte Zufallsvariablen auf $[0, 1]$ ($\mathcal{U} \sim [0, 1]$).

6.2.4 Unabhängigkeit stetiger Zufallsvariablen

Theorem: Seien $X_1, ..., X_n$ Z.V. mit Dichten $f_1, ..., f_n$. Dann sind folgende Aussagen äquivalent:

1. $X_1, ..., X_n$ sind unabhängig,
2. $X_1, ..., X_n$ sind insgesamt stetig mit gemeinsamer Dichte.

$$f(x_1, ..., x_n) = f_1(x_1) \cdots f_n(x_n)$$

Bemerkung: Somit sind zwei unabhängige stetige Z.V. automatisch gemeinsam stetig.

7 Grenzwertsätze

Vorbemerkung: In diesem Kapitel fixieren wir einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ und eine Folge von u.i.v.-Z.V. $X_1, X_2, ...$. Mit anderen Worten, wir erhalten Z.V. $X_i : \Omega \to \mathbb{R}$, so dass

$$\forall i_1 < \cdots i_k, \forall x_1, ..., x_k \in \mathbb{R} \quad \mathbb{P}[X_{i_1} \leq x_1, ..., X_{i_k} \leq x_k] = F(x_1) \cdots F(x_k).$$

wobei F die allgemeine Verteilungsfunktion ist. Für jedes n betrachen wir die Partialsumme

$$S_n = X_1 + \cdots + X_n,$$

und wir interessieren uns für das Verhalten (wenn n gros ist) der folgenden Z.V.

$$\frac{S_n}{n} = \frac{X_1 + \cdots + X_n}{n}.$$

Das wird manchmal auch der **empirische Durchschnitt** genannt.

7.1 Gesetz der grossen Zahlen (GGZ)

Theorem: Sei $\mathbb{E}[|X_1|] < \infty$. Setze $m = \mathbb{E}[X_1]$, dann gilt

$$\lim_{n\to\infty} \frac{X_1 + \cdots X_n}{n} = m \quad f.s.$$

Bemerkung: Da die Z.V. u.i.v. sind, haben wir ebenfalls $\mathbb{E}[|X_i|] < \infty$ und $m = \mathbb{E}[X_i]$ für jedes i.

Beispiele: Sei $X_1, X_2, ...$ eine Folge von u.i.v. Bernoulli Z.V. mit Parameter p. Dann ist

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = p \quad f.s.$$

Sei $T_1, T_2, ...$ eine u.i.v. Folge von exponential verteilten Z.V. mit Parameter λ. Dann gilt

$$\lim_{n\to\infty} \frac{T_1 + \cdots + T_n}{n} = \frac{1}{\lambda} \quad f.s.$$

7.2 Anwendung: Monte-Carlo Integration

Unser Ziel ist es folgendes Integral

$$I = \int_0^1 g(x)\,dx$$

nummerisch zu bestimmen. Die Idee: Wir betrachten I als Erwartungswert und verwenden das GGZ um I zu approximieren. Sei U eine gleichverteilte Z.V. auf [0, 1]. Dann gilt

$$\mathbb{E}[g(U)] = \int_0^1 g(x)\,dx = I.$$

Somit finden wir eine gute Approximation von I, falls wir obigen Erwartungswert $g(U)$ zufriedenstellen bestimmen können. Nun kommt das GGz ins Spiel. Sei $U_1, U_2, ...$ eine u.i.v. Folge von gleichverteilten Z.V. auf [0, 1] und setze $X_{=}g(U_n)$ für jedes n. Somit sind die Folgenglieder $X_1, X_2, ...$ u.i.v. und es gilt

$$\mathbb{E}[|X_1|] = \int_0^1 |g(x)|\,dx > \infty,$$

und $\mathbb{E}[X_1] = I$. Anwendung des GGZ liefert

$$\lim_{n\to\infty} \frac{g(U_1) + \cdots g(U_n)}{n} = I.$$

Somit erhalten wir eine Approximation von I.

7.3 Konvergenz in Verteilung

Def: Seien $(X_n)_{n\in\mathbb{N}}$ und X Z.V. Wir schreiben

$$X_n \overset{Approx}{\approx} X \text{ as } n \to \infty$$

falls für jedes $x \in \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{P}[X_n \le x] = \mathbb{P}[X \le x].$$

7.4 Zentraler Grenzwertsatz

7.4.1 Ein Frage der Fluktuation?

Das GGZ besagt, dass für grosse n der empirische Durchschnitt nahe dem Erwartungswert $m = \mathbb{E}[X_1]$ ist. Eine zweite Frage, die man stellen kann, ist:

$$\text{Wie weit ist } \frac{X_1 + \cdots + X_n}{n} \text{ typischerweise von } m \text{ entfernt?}$$

7.4.2 Fluktuation von Normalverteilten Z.V.

Betrachten wir zuerst den Fall, dass $X_1, X_2, ...$ eine Folge von i.i.d. normalen Z.V. mit den Parametern m und σ^2 ist. Dann sagen uns die Ergebnisse, die wir für normale Z.V. gesehen haben, dass

$$Z = \frac{X_1 + \cdots + X_n}{n} - m$$

wiederum eine normale Z.V. mit Parametern $\bar{m} = 0$ und $\bar{\sigma}^2 = \frac{1}{n}\sigma^2$ ist. Die Standardabweichung $\bar{\sigma} = \frac{1}{\sqrt{n}}\sigma$ stellt die typischen Schwankungen von Z dar. Grobn kann man sage, dass der typische Abstand zwischen $\frac{X_1+\cdots+X_n}{n}$ und m von der Ordnung $\frac{\sigma}{\sqrt{n}}$ ist.

7.4.3 Der zentrale Grenzwertsatz (ZGWS)

Seien X_1, $X2, ...$ nicht normalverteilt. Dann ist die Brechnung der Verteilung

$$\frac{X_1 + \cdots + X_n - n \cdot m}{\sqrt{\sigma^2 n}}$$

nicht immer einfacht. Hier setzt der ZGWS gerade an. Er besagt, dass für immer grösser werdende n die Verteilung der obigen Z.V. sich der Verteilung einer standard normalverteilten Z.V. annähert.

Theorem (ZGWS): Nehme an, dass der Erwartungswert $\mathbb{E}[X_1^2]$ wohldefineirt und endlich ist. Setze $m = \mathbb{E}[X_1]$ und $\sigma^2 = \mathrm{Var}(X_1)$, dann gilt folgender Grenzwert

$$\mathbb{P}\Big[\frac{S_n - n \cdot m}{\sqrt{\sigma^2 n}} \leq a\Big] \to_{n\to\infty} \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2}\, dx$$

für jedes $a \in \mathbb{R}$.

Beachte gerade, dass Φ gerade die Verteilungsfunktion einer Z.V. $Z \sim \mathcal{N}(0,\, 1)$ ist. Der Satz besagt somit, dass für grosse $n \in \mathbb{N}$ die Z.V.

$$Z_n = \frac{S_n - n \cdot m}{\sqrt{\sigma^2 n}}$$

einer Verteilung $Z \sim \mathcal{N}(0,\, 1)$ ähnelt.

8 Statistische Grundideen

Wir befassen uns im Folgenden mit der **induktiven Statistik.** Die Grundidee dabei ist wie folgt: Man fasst die Daten $x_1, ..., x_n$ auf als Realisierung / realisierte Werte $X_1(\omega), ..., X_n(\omega)$ von Z.V. $X_1, ..., X_n$, und sucht dann Aussagen über die Verteilung von $X_1, ..., X_n$.

Wichtig: Man muss immer sauber unterscheiden zwischen den *Daten* $x_1, ..., xn$ und dem generierenden Mechanismus $X_1, ..., X_n$ (bezeichnet mit grossen Buchstaben, sind Z.V., also Funktionen auf einem Ω).

Terminologie: Die Gesamtheit der Beobachtungen $x_1, ..., x_n$ oder Z.V. $X_1, ..., X_n$ nennt man of eine **Stichprobe**, die Anzahl n heisst dann der **Stichprobenumfang.**

9 Schätzer

Setup:

- Parameterraum $\Theta \subset \mathbb{R}$
- Grundraum Ω
- sigma-Algebra $\mathcal{F}$
- $(\mathbb{P}_\theta)_{\theta\in\Theta}$ Familie von Wahrscheinlichkeitsmasse auf $(\Omega,\, \mathcal{F})$
- $X_1, ..., X_n$ Zufallsvariablen auf $(\Omega,\, \mathcal{F})$

9.1 Grundbegriffe

Wir suchen für den Parameter θ einen Schätzer T aufgrund unserer Stichprobe $(X_1, ..., X_n)$.

Def: Ein **Schätzer** ist eine Zufallsvariable $T : \Omega \to \mathbb{R}$ der Form

$$T = t(X_1, ..., X_n),$$

wobei $t : \mathbb{R}^n \to \mathbb{R}$.

Einsetzen von Daten $x_i = X_i(\omega)$, $i = 1, ..., n$ liefert dann **Schätzwerte** $T(\omega) = t(x_1, ..., x_n)$ für θ.

Beispiel: Jemand behauptet zu schmecken, ob in einer Tasse Tee zuerst die Milch oder der Tee eingegossen worden ist. Wie kann man überprüfen, ob das stimmen kann?
Wie geben der Person an n Tagen je zwei Tassen, von welchen Sie sagen soll, in welcher zuerst die Milch und in welcher zuerst der Tee eingegossen wurde. Wir notieren uns dabei die Ergebnisse $x_1, ..., x_n \in \{0, 1\}$ und fassen wie üblich diese Daten als Realisation von Z.V. $X_1, ..., X_n$ auf. Dann ist $S_n = \sum_{i=1}^n X_i$ die zufällige Anzahl der korrekt klassifizierten Tassenpaare, und $s_n = \sum_{i=1}^n x_i$ die beobachtete Anzahl von Erfolgen.
Als Modell nehmen wir nun an, dass die X_i unter $\mathbb{P}_\theta$ i.i.d. $\sim \text{Ber}(\theta)$ mit $\theta \in \Theta = [0, 1]$ sind. Dann ist natürlich $S_n \sim \text{Bin}(n, \theta)$ unter $\mathbb{P}_\theta$, d.h. im Modell $\mathbb{P}_\theta$, das zu θ gehört, ist die Anzahl S_n der Erfolge binomialverteilt mit Parametern n und θ.
Weil wir den Parameter θ nicht kennen, liegt es nahe, zuerst einmal dafür einen Schätzer zu suchen. Eine erste Möglichkeit wäre, einfach das letzte Ergebnis zu nehmen. Unser erster Schätzer $\hat{T}$ für θ wäre also $\hat{T} = X_n$. Ein zweiter naheliegender Schätzer wäre die durchschnittliche Anzahl der Erfolge bei den n Versuchen. Unser zweiter Schätzer wäre also $T = \bar{X_n} = \frac{1}{n} S_n$. Für gegebene Daten $x_1, ..., x_n$ gibt uns das dann zwei Schätzwerte $\hat{t}(x_1, ..., x_n) = x_n$ und $t(x_1, ..., x_n) = \bar{x_n} = \frac{1}{n} \sum_{i=1}^n x_i$, die wir konkret berechnen können.

9.2 Bias

Def: Ein Schätzer T heiss **erwartungstreu (unbiased)** für θ, falls für alle $\theta \in \Theta$ gilt:

$$\mathbb{E}_\theta[T] = \theta$$

Die Interpretation dazu ist wie folgt: Im Mittel (über alle denkbaren Realisationen ω) schätzt T also richtig, und zwar unabhängig davon, welches Modell $\mathbb{P}_\theta$ zu Grunde liegt.

Def: Sei $\theta \in \Theta$ und T ein Schätzer. Der **Bias** (oder erwartete Schätzfehler) von T im Modell $\mathbb{P}_\theta$ ist definiert als

$$\mathbb{E}_\theta[T] - \theta.$$

Der mittlere quadratische Schätzfehler (mean squared error, MSE) von T im Modell $\mathbb{P}_\theta$ ist definiert als

$$\text{MSE}_\theta[T] := \mathbb{E}_\theta[(T - \theta)^2].$$

Bemerkung: Mann kann den MSE zerlegen als

$$\text{MSE}_\theta[T] = \mathbb{E}_\theta[(T - \theta)^2] = \text{Var}_\theta[T] + (\mathbb{E}_\theta[T] - \theta)^2,$$

also in die Summe aus der Varianz des Schätzers T un dem Quadrat des Bias.

9.3 Die Maximum-Likelihood-Methode (ML-Methode)

Ausgangspunkt im folgenden Abschnitt ist immer eine von zwei Situationen, jenachdem ob wir es mit diskreten oder mit stetigen Zufallsvariablen zu tun haben. Wir schreiben oft kurz $\vec{X} = (X_1, ..., X_n)$. In jedem Modell $\mathbb{P}_\theta$ sind $X_1, ..., X_n$ entweder diskret mit gemeinsamer Gewichtsfunktion $p_{\vec{X}}(x_1, ..., x_n; \theta)$ oder stetig mit gemeinsamer Dichtefunktion $f_{\vec{X}}(x_1, ..., x_n; \theta)$. Meistens sind sogar die X_i unter $\mathbb{P}_\theta$ i.i.d. mit individueller Gewichtsfunktion $p_X(x; \theta)$ bzw. Dichtefunktion $f_X(x; \theta)$. Dann ist also die gemeinsame Gewichtsfunktion

$$p_{\vec{X}}(x_1, ..., x_n; \theta) = \prod_{i=1}^n p_X(x_i; \theta)$$

bzw. die gemeinsame Dichtefunktion

$$f_{\vec{X}}(x_1, ..., x_n; \theta) = \prod_{i=1}^n f_X(x_i; \theta).$$

Anschaulich ist

$$p_{\vec{X}}(x_1, ..., x_n; \theta) = \mathbb{P}_\theta[X_1 = x_1, ..., X_n = x_n]$$

gerade die Wahrscheinlichkeit im Modell $\mathbb{P}_\theta$, dass unsere Strichprobe $X_1, ..., X_n$ die Werte $x_1, ..., x_n$ liefert, und $f_X(x_1, ..., x_n; \theta)$ ist das übliche stetige Analog.

Def: Die **Likelihood-Funktion** ist:

$$L(x_1, ..., x_n; \theta) := \begin{cases} p_{\vec{X}}(x_1, ..., x_n; \theta) & \text{im diskreten Fall,} \\ f_{\vec{X}}(x_1, ..., x_n; \theta) & \text{im stetigen Fall.} \end{cases}$$

Die Funktion $\log L(x_1, ..., x_n; \theta)$ heisst die **log-Likelihood-Funktion.** Sie hat gegenüber der Likelihood-Funktion den Vorteil, dass sie im i.i.d.-Fall durch eine Summe (statt ein Produkt) gegeben und damit zum Rechnen oft wesentlich einfacher ist.

Def: Für jedes $x_1, ..., x_n$ sei $t_{ML}(x_1, ..., x_n) \in \mathbb{R}$ der Wert, der $\theta \to L(x_1, ..., x_n; \theta)$ als Funktion von θ maximiert. D.h.,

$$L(x_1, ..., x_n; t_{ML}(x_1, ..., x_n)) = \max_{\theta \in \Theta} L(x_1, ..., x_n; \theta).$$

Ein **Maximum-Likelihood-Schätzer (ML-Schätzer)** T_{ML} für θ wird definiert durch

$$T_{ML} = t_{ml}(X_1, ..., X_n).$$

Meistens sind $X_1, ..., X_n$ i.i.d. unter $\mathbb{P}_\theta$. Die Likelihood-Funktion L ist dann ein Produkt, und es ist bequemer, statt L die log-Likelihood-Funktion $\log L$ zu maximieren, weil diese eine Summe ist. Statt zu maximieren sucht man ferner meistens nur *Nullstellen der Ableitung (nach θ).*

10 Konfidenzintervalle

Die Grundidee ist wie folgt: Wie im vorigen Abschnitt suchen wir aus einer Familie $(\mathbb{P}_\theta)_{\theta \in \Theta}$ von Modellen eines, das zu unseren Daten $x_1, ..., x_n$ passt. Ein Schätzer für θ gibt uns dabei einen einzelnen zufälligen möglichen Parameterwert. Weil es schwirig ist, mit diesem einen Wert den richtigen Parameter zu treffen, suchen wir nun stattdessen eine **zufällige Teilmenge des Parameterbereichs,** die hoffentlich den wahren Parameter enthält.

10.1 Definitionen

Eir reichhaltig sind diese Schätzer? Werfen wir zum Beispiel eine Münze 100 mal, ohne die Wahrscheinlichkeit p von Kopf zu kennen. Falls wir 70 mal Kopf erhalten, ist der Maximum-Likelihood-Schätzer für p $T_{ML} = 0.7$. Wie weit ligt T_{ML} von dem wahren Wert p entfernt? Um diese Art von Fragen zu beantworten, führen wir den Begriff der Konfidenzintervalle ein.

Def: Sei $\alpha \in [0, 1]$. Ein **Konfidenzintervall für θ mit Niveau** $1-\alpha$ ist ein Zufallsintervall $I = [A, B]$, sodass gilt

$$\forall \theta \in \Theta \quad \mathbb{P}_\theta[A \leq \theta \leq B] \geq 1 - \alpha,$$

wobei A, B Zufallsvariablen der Form $A = a(X_1, ..., X_n)$, $B = b(X_1, ..., X_n)$ mit $a, b : \mathbb{R}^n \to \mathbb{R}$ sind.

Zu bemerken ist hier, dass θ in dieser Gleichung deterministisch und nicht zufällig ist. Nur die Schranken $A = a(X_1, ..., X_n)$ und $B = b(X_1, ..., X_n)$ sind Zufallsvariablen.

Beispiel (Konfidenzintervall für normales Modell mit bekannter Varianz): Seien $X_1, ..., X_n$ u.i.v. normalverteilte Zufallsvariablen mit Parametern m und $\sigma^2 = 1$. Wir betrachten somit ein stochastisches Modell mit bekannter Varianz ($\sigma^2 = 1$) aber unbekanntem Mittelwert μ ($X_1 \sim \mathcal{N}(\mu, 1)$). Mann kann zeigen, dass der Maximum-Likelihood Schätzer gegeben ist durch

$$T = T_{ML} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

mit $T_{ML} \sim \mathcal{N}(\mu, \frac{1}{n})$, also $Z = \sqrt{n}(T_{ML} - \mu) \sim \mathcal{N}(0, 1)$ (wir normalisieren T_{ML}). Wir suchen also für μ Konfidenzintervalle der Form

$$I = [T_{ML} - \frac{c}{\sqrt{n}}, T_{ML} + \frac{c}{\sqrt{n}}].$$

Zuerst betrachten wir

$$\mathbb{P}_\theta[T_{ML} - \frac{c}{\sqrt{n}} \leq mu \leq T_{ML} + \frac{c}{\sqrt{n}}] = \mathbb{P}_\theta[-c \leq Z \leq c].$$

Somit können wir die obige Wahrscheinlichkeit explizit bestimmen:

$$\mathbb{P}_\theta[-c \leq Z \leq c] = \mathbb{P}_\theta[Z \leq c] - \mathbb{P}_\theta[Z < -c] \,(= 1 - \mathbb{P}_\theta[Z \leq c]) = 2\Phi(c) - 1,$$

wobei $\Phi(c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} \exp(\frac{-x^2}{2})\, dx$ ist und mittels einer Tabelle der Standardnormalverteilung nachgelesen werden kann. Somit ergibt sich $2\Phi(1.96) - 1 \geq 0.95$ und schliesslich:

$$I = [T_{ML} - \frac{1.96}{\sqrt{n}}, T + \frac{1.96}{\sqrt{n}}].$$

10.2 Verteilungsaussagen

Def: Eine stetige Z.V. X heisst χ^2**- Verteilung mit** m **Freiheitsgraden** falls ihre Dichte gegeben ist durch

$$f_X(y) = \frac{1}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})} y^{\frac{m}{2}-1} e^{-\frac{1}{2}y} \quad \text{fuer } y \geq 0.$$

Wir schreiben dann $X \sim \chi^2_m$. Dabei ist die sogenannte Gamma-Funktion für $v \geq 0$ gegeben durch

$$\Gamma(v) := \int_0^\infty t^{v-1} e^{-t}\, dt.$$

Es gilt $\Gamma(n) = (n-1)!$ für $v = n \in \mathbb{N}$.

Bemerkung: Die χ^2 Verteilung mit m Freiheitsgraden ist der Spezialfall einer $Ga(\alpha, \lambda)$-Verteilung mit $\alpha = \frac{m}{2}$ und $\lambda = \frac{1}{2}$. Für $m = 2$ ergibt sich eine Exponentialverteilung mit $X \sim Exp(\frac{1}{2})$.

Satz: Sind die Z.V. $X_1, ..., X_m$ u.i.v. $\sim \mathcal{N}(0, 1)$, so ist die Summe $Y := \sum_{i=1}^{m} X_i^2 \sim \chi^2_m$.

Def: Eine stetige Z.V. X heisst t**verteilt mit** m **Freiheitsgraden** falls ihre Dichte gegeben ist durch

$$f_X(x) = \frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\Gamma(\frac{m}{2})} \Big(1 + \frac{x^2}{m}\Big)^{-\frac{m+1}{2}} \quad \text{fuer } x \in \mathbb{R}.$$

Wir schreiben dann $X \sim t_m$.

Bemerkung: Für $m = 1$ ist das eine Cauchy-Verteilung, und für $m \to \infty$ erhält man asymptotisch eine $\mathcal{N}(0, 1)$-Verteilung. Wie die $\mathcal{N}(0, 1)$-Verteilung ist die t-Verteilung symmetrisch um 0, sie ist aber langschwänziger (d.h. ihre Dichte geht langsamer gegen 0, wenn das Argument gegen $\pm\infty$ geht), und zwar umso mehr, je kleiner m ist.

Satz: Sind X und Y unabhängig mit $X \sim \mathcal{N}(0,1)$ und $Y \sim \chi^2_m$, so ist der Quotient

$$Z := \frac{X}{\sqrt{\frac{1}{m}Y}}.$$

10.3 Normalverteilung mit σ und m unbekannt

Wir erinnern an die Notationen

$$\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X_n})^2$$

für das Stichprobenmittel und die Stichprobenvarianz.

Satz: Seien $X_1, ..., X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. Dann sind $\bar{X_n}$ und S^2 unabhängig.

In diesem Fall sind $\bar{X_n}$ und S^2 unsere Schätzer für μ und σ^2. Es gilt zudem $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

Beispiel: Betrachten wir nun $A = \frac{\bar{X_n}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ und $B = \sqrt{\frac{\frac{n-1}{\sigma^2}S^2}{n-1}}$, wobei $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$, dann gilt

$$Z := \frac{A}{B} = \frac{\bar{X_n} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}.$$

Damit können wir ein Konfidenzintervall für μ der Form

$$[A, B] = \left[\bar{X_n} - c \cdot \sqrt{\frac{S^2}{n}}, \bar{X_n} + c \cdot \sqrt{\frac{S^2}{n}}\right],$$

wobei wir dazu wieder die Tabelle für die t_{n-1}-Verteilung benutzen, um einen geigneten Wert für c zu finden.

10.4 Approximative Konfidenzintervalle

Einen allgemeinen **approximativen Zugang** liefert der zentralge Grenzwertsatz. Oft ist ein Schätzer T eine Funktion einer Summe $\sum_{i=1}^n Y_i$, wobei die Y_i im Modell $\mathbb{P}_\theta$ i.i.d. sind. Das einfachste Beispiel ist $T = \bar{Y_n} = \frac{1}{n}\sum_{i=1}^n Y_i$. Nach dem zentralen Grenzwertsatz ist dann für grosse n

$$\sum_{i=1}^{n} Y_i \quad \text{approximativ normalverteilt unter } \mathbb{P}_\theta$$

mit Parametern $\mu = n \cdot \mathbb{E}_\theta[Y_i]$ und $\sigma^2 = n \cdot \mathrm{Var}_\theta[Y_i]$. Das kann man benutzen, um für die Verteilung von T approximative Aussagen zu bekommen und damit gewisse Fragen zumindest approximativ zu beantworten.

11 Tests

Tagesproblem: Sophie ist Statistikstudentin und Velokurier. Sie hat eine Lieferung mit einem Beutel fairer Münzen ($p = 0.5$) und einem Beutel gezinkter Münzen ($p = 0.7$). Die fairen Münzen sollen ins Casion, die gezinkten sollen entsorgt werden. Bei einem Velounfall werden alle Münzen vermischt. Wie kann Sopie entscheiden, welche Münzen ins Casion sollen, und welche entsorgt werden (ohne jede Münzen 10'000 Mal zu werfen)?

11.1 Null- und Alternativhypothese

Ausgangspunkt ist wieder eine Stichprobe $X_1, ..., X_n$. Wir betrachten wieder eine Familie von Wahrscheinlichkeiten $\mathbb{P}_\theta$ mit $\theta \in \Theta$, die unsere möglichen Modelle beschreiben. Wie bisher kann θ eine ein- oder mehrdimensionaler Parameter sein.

Das Grundproblem ist, eine Entscheidung zwischen zwei konkurrierenden Modellklassen zu treffen – der **Nullhypothese** $\Theta_0 \subseteq \Theta$ und der **Alternativhypothese** $\Theta_A \subseteq \Theta$, wobei $\Theta_0 \cap \Theta_A = \emptyset$. Meist schreibt man das als

$$\text{Nullhypothese } H_0 : \theta \in \Theta_0,$$
$$\text{Alternativhypothese } H_A : \theta \in \Theta_A.$$

Ist keine explizite Alternative spezifiziert, so hat man $\Theta_A = \Theta_0^C = \Theta \setminus \Theta_0$. Null- und/oder Alternativhypothese heissen **einfach,** falls Θ_0 bzw. Θ_A aus einem einzelnen Wert, θ_0 bzw. θ_A, bestehen. Sonst heissen sie **zusammengesetzt.**

> **Beispiel:** In unserem Tagesproblem sind also:
>
> $$H_0 : \theta = 0.7 \quad (\theta \in \Theta_0 = \{0.7\})$$
> $$H_A : \theta = 0.5 \quad (\theta \in \Theta_A = \{0.5\})$$

11.2 Test und Entscheidung

Def: Ein **Test** ist ein Paar (T, K), wobei

- T eine Z.V. der Form $T = t(X_1, ..., X_n)$ ist, und
- $K \subseteq \mathbb{R}$ eine (deterministische) Teilmenge von $\mathbb{R}$ ist.

Die Z.V. $T = t(X_1, ..., X_n)$ heisst dann **Teststatistik,** und K heisst **kritischer Bereich** oder **Verwerfungsbereich.**

Def: Die **Entscheidungsregel** ist wie folgt:

- die Hypothese H_0 wird verworfen, falls $T(\omega) \in K$,
- die Hypothese H_A wird nicht verworfen, bzw. angenommen, falls $T(\omega) \notin K$.

Die Entscheidung bei einem Test kann auf zwei verschiedene Arten falsch herauskommen:

1. Bei einem **Fehlert 1. Art** wir die Nullhypothese zu Unrecht verworfen, d.h. obwohl sie richtig ist. Das passiert für $\theta \in \Theta_0$ und $T \in K$, deshalb heisst $\mathbb{P}_\theta[T \in K]$ für $\theta \in \Theta_0$ die Wahrscheinlichkeit für einen Fehler 1. Art.
2. Bei einem **Fehler 2. Art** wird die Nullhypothese zu Unrecht nicht verworfen, d.h. man akzeptiert die Nullhypothese, obwohl sie falsch ist. Das passiert für $\theta \in \Theta_A$ und $T \notin K$, und deshalb heisst $\mathbb{P}_\theta[T \notin K] = 1 - \mathbb{P}_\theta[T \in K]$ für $\theta \in \Theta_A$ die Wahrscheinlichkeit für einen Fehler 2. Art.

> **Beispiel:** In unserem Tagesproblem ist:
>
> - Fehler 1. Art: Sophie bringt eine gezinkte Münze ins Casino.
> - Fehler 2. Art: Sophie entsorgt eine faire Münze.

11.3 Signifikanzniveau und Macht

Def: (Fehler 1. Art vermeiden) Sei $\alpha \in (0, 1)$. Ein Test (T, K) besitzt **Signifikanzniveau** α, falls

$$\forall \theta \in \Theta_0 \quad \mathbb{P}_\theta[T \in K] \leq \alpha.$$

Def: (Fehler 2. Art vermeiden) Die **Macht** eines Tests (T, K) wird definiert als folgende Funktion

$$\beta : \Theta_A \to [0, 1], \quad \theta \to \beta(\theta) := \mathbb{P}_\theta[T \in K].$$

Bemerkung:

- α *klein* bedeuted, dass die Wahrscheinlichkeit eines Fehlers 1. Art klein ist
- β *gross* bedeuted, dass die Wahrscheinlichkeit eines Fehlers 2. Art klein ist

Wir können das **asymmetrische Verhalten** zwischen Null- und Alternativhypothese anhand unseres Tagesproblems beobachten:

Beispiel: Betrachten wir nochmals unser Tagesproblem. Wir nehmen an, dass Sophie jede Münze $n = 10$ Mal wirft. $T = \sum_{i=1}^{10} X_i$ ist die Anzahl der Kopf-Würfe und $K = (-\infty, j)$, wobei $j \in \{0, 1, ..., 10\}$ fixiert wird.

j	Signifikanzniveau $\alpha = \mathbb{P}_{0.7}[T \leq j]$	Macht $\beta(0.5) = \mathbb{P}_{0.5}[T \leq j]$
4	$\approx 4.7\%$	$\approx 62.3\%$
5	$\approx 15\%$	$\approx 37.7\%$
6	$\approx 35\%$	$\approx 17.2\%$
7	$\approx 62\%$	$\approx 5.5\%$

11.4 Konstruktion von Tests

Seien $\theta_0 \neq \theta_A$ zwei fixierte Zahlen. In diesem Abschnitt nehmen wir stets an, dass sowohl die Nullhypothese als auch die Alternativhypothese von der einfachen From

$$H_0 : \theta = \theta_0,$$
$$H_A : \theta = \theta_A,$$

ist. Ferner nehmen wir an, dass die Z.V. $X_1, ..., X_n$ entweder diskret, oder gemeinsam stetig unter $\mathbb{P}_{\theta_0}$ und unter $\mathbb{P}_{\theta_A}$ sind.

Die Grundidee ist wie folgt: Wir fixieren α (klein), da wir unbedingt den Fehler 1. Art klein halten wollen. Danach suchen wir einen Test, sodass β möglichst gross ist.

Def: Für jedes $x_1, ..., x_n$ definieren wir den **Likelihood-Quotienten** durch

$$R(x_1, ..., x_n) := \frac{L(x_1, ..., x_n; \theta_A)}{L(x_1, ..., x_n; \theta_0)}.$$

Dabei ist $L(x_1, ..., x_n; \theta)$ die Likelihood-Funktion und wir setzen als Konvention $R(x_1, ..., x_n) = +\infty$, falls $L(x_1, ..., x_n; \theta_0) = 0$.

Es ligt somit nahe, als Teststatistik $T := R(X_1, ..., X_n)$ und als kritischen Bereich $K := (c, \infty)$ zu wählen, wenn man θ_0 gegen θ_A testen will. Schliesslich wird gerade dann die Hypothese H_0 verworfen, falls der Quotient R gross ist.

Def: Sei $c \geq 0$. Der **Likelihood-Quotienten-Test (LQ-Test) mit Parameter** c ist ein Test (T, K), wobei Teststatistik und Verwerfungsbereich gegeben sind durch

$$T = R(X_1, ..., X_n) \quad \text{und} \quad K = (c, \infty).$$

Beispiel: Betrachten wir das Tagesbeispiel von letzter Woche. Seien $X_1, ..., X_n \sim Ber(\theta)$. Zudem haben wir:

- $H_0 : \theta = 0.7$
- $H_A : \theta = 0.5$

Für $X_i \sim Ber(\theta)$ gilt:

$$L(x_1, ..., x_n; \theta) = \mathbb{P}_\theta[X_1 = x_1, ..., X_n = x_n] = \mathbb{P}[X_1 = x_1] \cdots \mathbb{P}[X_n = x_n] = \theta^{|X|}(1-\theta)^{n-|X|},$$

wobei $|X| = \sum_{i=1}^n X_i$. In unserem Fall ist nun:

$$R(x_1, ..., x_n) = \left(\frac{3}{7}\right)^{|X|} \frac{1}{0.6^n}.$$

Somit erhalten wir für $n = 10$ und $c = 1$

- $\alpha = 0.35$
- $\beta = 0.82$

sowie für $c = 10$

- $\alpha = 0.0101$
- $\beta = 0.1719$

Theorem (Neyman-Pearson-Lemma): Sei $c \geq 0$. Sei (T, K) ein Likelihood-Quotienten-Test mit Parameter c und Signifikanzniveau $\alpha^* := \mathbb{P}_{\theta_0}[T > c]$. ist (T', K') ein anderer Test mit Signifikanzniveau $\alpha \leq \alpha^*$, dann gilt

$$P_{\theta_A}[T' \in K'] \leq \mathbb{P}_{\theta_A}[T \in K].$$

11.5 Beispiele

Betrachten wir folgendes Modell: $(\mathbb{P}_\theta)_{\theta \in \mathbb{R}}$ und $X_1, ..., X_n$ u.i.v. $\sim \mathcal{N}(\theta, 1)$ unter $\mathbb{P}_\theta$. Zudem gilt:

- $H_0 : \theta = 0$
- $H_A : \theta \neq 0$

Betrachten wir nun zwei verschiedene Tests:

- Test 1: $T = \frac{X_1 + \cdots + X_n}{\sqrt{n}}$ und $K = (1.65, \infty)$
- Test 2: $T = \frac{X_1 + \cdots + X_n}{\sqrt{n}}$ und $K = (2.33, \infty)$

Bemerkung: Unter $\mathbb{P}_0$ ist $T \sim \mathcal{N}(0, 1)$.
Für die Signifikanz gilt:

- Test 1: $\alpha = \mathbb{P}_{H_0}[H_0 \text{ verworfen}] = \mathbb{P}_{\theta=0}[T > 1.65] = 5\%$
- Test 2: $\alpha = \mathbb{P}_{H_0}[H_0 \text{ verworfen}] = \mathbb{P}_{\theta=0}[T > 2.33] = 1\%$

Beobachten wir nun verschiedene Daten. Wir haben $T(\omega) = 2.967$. Dann wird H_0 sowohl für Test 1, sowie auch für Test 2 verworfen. Somit ist unsere Aussage viel kräftiger als wenn wir z.B. nur Test 1 betrachtet hätten.

Für dieses Beispiel ist der **p-Wert** definiert als:

$$p\text{-Wert} := \mathbb{P}_0[T > 2.976] = 0.2\%.$$

In anderen Worten gibt uns der p-Wert in diesem Fall das Signifikanzniveau vom "besten" Test, der H_0 verwirft.

11.6 p-Wert

Def: Sei $X_0 : \theta = \theta_0$ eine einfache Nullhypothese. Sei $(T, K_t)_{t \geq 0}$ eine geordnete Familie von Tests. Der **p-Wert** ist definiert alls die Z.V.

$$p\text{-Wert} = G(T),$$

wobei $G : \mathbb{R}_+ \to [0, 1]$ mittels $G(t) = \mathbb{P}_{\theta_0}[T \in K_t]$ definiert ist.