Probability & Statistics

Fadoua Balabdaoui

Typed and edited by Eric Cegli

Drawings illustrated by Heidi Hertach

Contents

Probability Theory

Statistics

Appendix

Probability Theory

The Concept of Probability

The set of all possible outcomes is called the *sample space* and is usually denoted by Ω. An element $\omega \in \Omega$ is called an *elementary outcome*.

Examples 0.1. Different experiments require different choices of Ω, where Ω can be any set. Here we list a few first examples.

- The experiment of tossing a coin can be modeled by $\Omega := \{\text{Head}, \text{Tails}\}$. One could also take $\tilde{\Omega} := \{H, T\}$, so the choice of a sample space is all but unique.
- Drawing one of six balls can be modeled by $\Omega := \{1, 2, \ldots, 6\}$.
- Assume we want to model the motion of a small particle in a fluid. Such a motion can be interpreted as a continuous function $f : \mathbb{R}_+ \to \mathbb{R}^3$, hence we choose $\Omega := C(\mathbb{R}_+, \mathbb{R}^3)$.
- To model the random number of emails received during a weekday we may choose $\Omega := \mathbb{N}_0$.

Let us now consider $\mathcal{A} =$ "the set of all observable events". For now, we take $\mathcal{A} = 2^\Omega$ (which denotes the *powerset* of Ω). For an $A \in \mathcal{A}$ we say that A *occurs* if the element ω belongs to A, so if we have $\omega \in A$.

Example 0.2. Consider the experiment of throwing a die, so $\Omega := \{1, 2, \ldots, 6\}$ is a suitable choice. Let us consider the event $A := \{\text{the number is} < 5\} = \{1, 2, 3, 4\} \in \mathcal{A}$. In this case, if the die falls on $1, 2, 3$ or 4, we say that the event A occurs.

After choosing $(\Omega, \mathcal{A})$, we will define a map $\mathbb{P} : \mathcal{A} \to [0, 1]$ which, if it fulfills certain properties, is be called a *probability measure* and $\mathbb{P}(A)$ is called the *probability* with which the event A occurs. The triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is then called a *probability space*.

1 Discrete Probability Spaces

1.1 Introduction

In this section, we will put ourselves in the case where Ω is either *finite* or *infinitely countable* and always consider $\mathcal{A} := 2^\Omega$. We will assume that to each $\omega \in \Omega$ we can assign a *weight* $p(\omega) \in [0,1]$ such that

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

For an event $A \in \mathcal{A} = 2^\Omega$ we then set

$$\mathbb{P}(A) := \sum_{\omega \in A} p(\omega).$$

Note that $\mathbb{P} : \mathcal{A} \to [0,1]$ is now completely determined and we have the following properties:

- $\forall \Omega \in \Omega : \ \mathbb{P}(\{\omega\}) = p(\omega)$.
- $\mathbb{P}(\Omega) = 1$.

- For any sequence $(A_k)_{k\in\mathbb{N}} \subseteq \mathcal{A}$ of pairwise disjoint events we have

$$\mathbb{P}\Big(\bigsqcup_k A_k\Big) = \sum_k \mathbb{P}(A_k).$$

  Indeed, observe that

$$\mathbb{P}\Big(\bigsqcup_k A_k\Big) = \sum_{\omega\in\bigsqcup_k A_k} p(\omega) = \sum_k \sum_{\omega\in A_k} p(\omega) = \sum_k \mathbb{P}(A_i)$$

  holds.

This construction motivates the following definition.

Definition 1.1. A set function $\mathbb{P} : \mathcal{A} \to [0, 1]$ is called a *probability measure* if

- $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$,
- for any sequence $(A_k)_{k\in\mathbb{N}} \subseteq \mathcal{A}$ of pairwise disjoint events we have

$$\mathbb{P}\Big(\bigsqcup_k A_k\Big) = \sum_k \mathbb{P}(A_k), \tag{1.1}$$

  called *sigma additivity* of $\mathbb{P}$.

In this case, the triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is called a *probability space.*

Proposition 1.2. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $A, B \in \mathcal{A}$ arbitrary. Then*

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$,
- *if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$, called* monotonicity *of $\mathbb{P}$.*

Proof.

- By sigma additivity (1.1) we immediately get

$$\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(A \sqcup A^c) = \mathbb{P}(\Omega) = 1$$

  which proves $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

- Observe that we have $A \cup B = A \sqcup (B \setminus A)$. Hence again by (1.1) it we get

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A).$$

  On the other hand, we have $B = (A \cap B) \sqcup (B \setminus A)$, so again by (1.1)

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

so we arrive at

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

- Assume $A \subseteq B$ and thus $A \cap B = A$. Then again by (1.1) we have

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}((A \cap B) \sqcup (B \setminus A)) \\ &= \mathbb{P}(A) + \underbrace{\mathbb{P}(B \setminus A)}_{\geq 0} \geq \mathbb{P}(A),\end{aligned}$$

which concludes the proof. □

Remark 1.3. Note that the second identity of Proposition 1.2 can be generalized to any subsets $A_1, \ldots, A_k \in \mathcal{A}$ by

$$\mathbb{P}\Big(\bigcup_{i=1}^{k} A_i\Big) = \sum_{j=1}^{k} (-1)^{j+1} \sum_{1 \leq i_1 < \ldots < i_j \leq k} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_j}).$$

Examples 1.4.

- A coin is tossed. Let $p \in (0,1)$ be the probability that the coin falls on heads. Set $\Omega = \{0,1\}$ and $p(1) = p$, so $p(0) = 1 - p(1) = 1 - p$. If $\mathbb{P}$ is the associated probability measure, then we have

$$\mathbb{P}(A) = \begin{cases} p & \text{if } A = \{1\} \\ 1-p & \text{if } A = \{0\} \\ 1 & \text{if } A = \Omega \\ 0 & \text{if } A = \emptyset. \end{cases}$$

for any $A \in \mathcal{A} = 2^{\Omega}$. The triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is then called a *Bernoulli-model*.

- Consider the same coin but now toss it n times. We denote by ω the number of Heads obtained by drawing this n tosses, so $\Omega = \{1, 2, \ldots, n\}$. Under an additional assumption, we can then show that

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

are the "right" weights for $\omega \in \Omega$, where $\binom{n}{\omega} = \frac{n!}{\omega!(n-\omega)!}$. The associated s probability measure is given by

$$\mathbb{P}(A) = \sum_{\omega \in A} \binom{n}{\omega} p^{\omega} (1-p)^{n-\omega}$$

for any $A \in 2^{\Omega}$. The triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is then called *Binomial-model* with *success probability* p and *number of trials* n.

- Let ω be the number of calls an SBB employee receives between 8:00 and 10:00, so $\Omega = \mathbb{N}_0$. For $\omega \in \Omega$ consider

$$p(\omega) := \frac{e^{-\lambda}\lambda^{\omega}}{\omega!}$$

for a fixed $\lambda > 0$ called *intensity/rate*. Then

$$\mathbb{P}(A) = \sum_{\omega \in A} \frac{e^{-\lambda}\lambda^{\omega}}{\omega!}$$

and $(\Omega, \mathcal{A}, \mathbb{P})$ is called a *Poisson-model* with intensity λ. In this case, for

$$A = \{\text{“At least one call received”}\}$$

we have

$$\mathbb{P}(A) = 1 - \mathbb{P}(\{0\}) = 1 - e^{-\lambda}.$$

1.2 Random Variables

Definition 1.5. Any map $X : \Omega \to \mathbb{R}$ is called a *random variable.*

Remark 1.6. If Ω is finite, then $X(\Omega)$ is also finite and if Ω is infinitely countable then $X(\Omega)$ is at most also infinitely countable.

Examples 1.7.

- Consider the experiment of tossing a coin twice. Put $\omega = (\omega_1, \omega_2)$ with $\omega_i =$ “the face of the i-th toss”. Then

$$\Omega = \{1, 2, \ldots, 6\}^2.$$

Set $X(\omega) := \omega_1 + \omega_2$ and $Y(\omega) := \omega_1\omega_2$ for $\omega \in \Omega$. Then X and Y are both random variables.

- Let ω be the random number of emails received on a day. Set $\Omega = \mathbb{N}_0$ and $X(\omega) = \mathbb{1}_{\{\omega=0\}}$. Then X is a *Bernoulli random variable.*

Now let X be any random variable on a sample space Ω and for $x \in X(\Omega)$ consider the event

$$\{X = x\} := \{\omega \in \Omega \mid X(\omega) = x\}.$$

We will also write

$$\mathbb{P}(X = x) := \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\})$$

The values $\mathbb{P}(X = x)$, as values in $X(\Omega)$, induce a new probability measure on $2^{X(\Omega)}$. We denote this new probability measure by $\mathbb{P}^X$. Then for any $B \in 2^{X(\Omega)}$ we have

$$\mathbb{P}^X(B) = \sum_{x \in B} \mathbb{P}(X = x).$$

The probability measure $\mathbb{P}^X$ is called the *distribution* of X.

Example 1.8. Let $\Omega = \mathbb{N}_0$ and $p(\omega) = \frac{e^{-1}}{\omega!}$ for $\omega \in \Omega$ and put $X(\omega) := \omega^2$. Then we have

$$\begin{aligned}
\mathbb{P}(X \geq 3) &= \mathbb{P}^X([3, \infty)) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \geq 3\} \\
&= \mathbb{P}(\{\omega \in \Omega \mid \omega \geq \sqrt{3}\}) = \mathbb{P}(\{\omega \in \Omega \mid \omega \geq 2\}) \\
&= 1 - \mathbb{P}(\{0\}) - \mathbb{P}(\{1\}) = 1 - 2e^{-1} \approx 0.26.
\end{aligned}$$

1.3 Expectations

We start with the case where X is a non-negative random variable.

Definition 1.9. Let $X \geq 0$ be a random variable defined on Ω with given probability weights $p(\omega)$ for $\omega \in \Omega$. Then the *expectation* of X is defined by

$$\mathbb{E}[X] := \sum_{\omega \in \Omega} X(\omega) p(\omega) \in [0, \infty].$$

Now for any function $f : \Omega \to \mathbb{R}$ set $f_+ := \max(f, 0)$ and $f_- = \max(-f, 0)$. Then we have $f = f_+ - f_-$ and $|f| = f_+ + f_-$ with $f_+, f_- \geq 0$.

Definition 1.10. Let X be any random variable on Ω. If $\min(\mathbb{E}[X_+], \mathbb{E}[X_-]) < \infty$ then we define the *expectation* of X by

$$\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-] \in [-\infty, \infty]$$

say that X is *integrable*.

Proposition 1.11. *If X is a non-negative or integrable random variable then we have*

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x \cdot \mathbb{P}(X = x).$$

Proof. In both cases, we have

$$\begin{aligned}
\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) p(\omega) &= \sum_{x \in X(\Omega)} \sum_{\omega \colon X(\omega) = x} \underbrace{X(\omega)}_{=x} p(\omega) \\
&= \sum_{x \in X(\Omega)} x \cdot \Big(\sum_{\omega \colon X(\omega) = x} p(\omega) \Big) \\
&= \sum_{x \in X(\Omega)} x \cdot \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\} \\
&= \sum_{x \in X(\Omega)} x \cdot \mathbb{P}(X = x)
\end{aligned}$$

which proves the claim. □

Examples 1.12.

- For any $A \in 2^{\Omega}$ we have

$$\begin{aligned}\mathbb{E}[\mathbb{1}_A] &= 0 \cdot \mathbb{P}(\mathbb{1}_A = 0) + 1 \cdot \mathbb{P}(\mathbb{1}_A = 1)\\ &= \mathbb{P}(\mathbb{1}_A = 1) = \mathbb{P}(A).\end{aligned}$$

- Let $\Omega = \mathbb{N}_0$ and $X(\omega) = \omega$ with the weights $p(\omega) = \frac{e^{-\lambda}\lambda^{\omega}}{\omega!}$ for $\omega \in \Omega$ and a fixed $\lambda > 0$. Then we have

$$\begin{aligned}\mathbb{E}[X] &= \sum_{\omega \in \mathbb{N}} \omega \cdot p(\omega) = \sum_{\omega=1}^{\infty} \omega \frac{e^{-\lambda}\lambda^{\omega}}{\omega!}\\ &= \sum_{\omega=1}^{\infty} \frac{e^{-\lambda}\lambda^{\omega}}{(\omega-1)!} = \lambda e^{-\lambda} \underbrace{\sum_{\omega=0}^{\infty} \frac{\lambda^{\omega}}{\omega!}}_{=e^{\lambda}} = \lambda.\end{aligned}$$

- Let $\Omega = \{1, 2, \dots, n\}$ with weights $p(\omega) = \binom{n}{\omega} p^{\omega}(1-p)^{n-\omega}$ and set $X(\omega) := \omega$. Then we have

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

Proposition 1.13. *Let X, Y be two integrable random variables.*

(i) If $X \leq Y$ holds then we have $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

(ii) We have $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$.

(iii) For any $\alpha, \beta \in \mathbb{R}$ we have $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$.

Example 1.14. Consider the experiment of tossing a coin n times with $p =$ "probability of obtaining heads". We are interested in the expectation of the number of times the coin falls on heads. The outcome of each experiment can be written as

$$\omega = (\omega_1, \dots, \omega_n)$$

where

$$\omega_i = \begin{cases} 1 & \text{if we obtain heads in toss } i \\ 0 & \text{if we obtain heads in tails } i. \end{cases}$$

with $\Omega = \{0,1\}^n$. Now set

$$X(\omega) := \sum_{i=1}^{n} \omega_i.$$

Then $X(\Omega) = \{1, 2, \dots, n\}$ and X represents the number of times we obtained heads. Furthermore, set

$$X_i(\omega) := \omega_i$$

for each $i \in \{1, 2, \dots, n\}$ which represents the outcome of the i-th toss. Then we have

$$\mathbb{E}[X_i] = \mathbb{P}(X_i = 1) = p$$

and $\sum_{i=1}^{n} X_i = X$. Hence by using linearity of the expectation we get

$$\mathbb{E}[X] = \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \underbrace{\mathbb{E}[X_i]}_{=p} = np.$$

Lemma 1.15. *Let X be a $\mathbb{N}$-valued random variable. Then we have*

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} \mathbb{P}(X > n).$$

Proof. Observe that we have

$$\{X > n\} = \bigsqcup_{k=n+1}^{\infty} \{X = k\}$$

for all $n \in \mathbb{N}$. Thus using sigma additivity of $\mathbb{P}$ we get

$$\mathbb{P}(X > n) = \sum_{k=n+1}^{\infty} \mathbb{P}(X = k) = \sum_{k=0}^{\infty} \mathbb{1}_{\{n \le k-1\}} \mathbb{P}(X = k).$$

Now by using Fubini's theorem we have

$$\begin{aligned} \sum_{n=0}^{\infty} \mathbb{P}(X > n) &= \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \mathbb{1}_{\{n \le k-1\}} \mathbb{P}(X = k) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(X = k) \underbrace{\sum_{n=0}^{\infty} \mathbb{1}_{\{n \le k-1\}}}_{=k} \\ &= \sum_{k=0}^{\infty} k \cdot \mathbb{P}(X = k) = \mathbb{E}[X] \end{aligned}$$

which concludes the proof. □

Example 1.16. Let X be a *geometric random variable* with success probability $p \in (0,1)$, that is

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

for $k \in \mathbb{N}_0$. Intuitively, X represents the "random waiting time" before a success which may happen with probability p. Using Lemma 1.15 we get

$$\begin{aligned}\mathbb{E}[X] &= \sum_{n=0}^{\infty} \mathbb{P}(X > n) = \sum_{n=0}^{\infty} \sum_{k=n+1}^{\infty} p(1-p)^k \\ &= \sum_{n=0}^{\infty} p(1-p)^{n+1} \underbrace{\sum_{k=0}^{\infty} (1-p)^k}_{=\frac{1}{p}} \\ &= (1-p) \sum_{n=0}^{\infty} (1-p)^n = \frac{1-p}{p} = \frac{1}{p} - 1.\end{aligned}$$

Note that for a coin with success probability $p = \frac{1}{2}$ we get $\mathbb{E}[X] = 1$.

1.4 Laplace Models

A *Laplace model* assumes that all elementary events $\omega \in \Omega$ have the same probability to occur. This model only "makes sense" if Ω is finite. In this case, we have

$$p(\omega) = \frac{1}{|\Omega|}$$

for all $\omega \in \Omega$ and

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

for all events $A \in \mathcal{A}$. $\mathbb{P}$ is also called the *(discrete) uniform probability measure* on $(\Omega, \mathcal{A})$.

Examples 1.17.

(1) Throw a fair die. All faces have probability $p = \frac{1}{6}$ to occur. Let

$$A := \{\text{the received number is odd}\} = \{1, 3, 5\}.$$

Then we have $\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{1}{2}$.

(2) Consider an urn with N balls numbered from 1 to N and K of them are red, $N - K$ are white.

Experiment. We draw $n \leq N$ balls from the urn with replacement.

Let $k \in \{1, \ldots, n\}$ and consider the event

$$R_k := \{\text{exactly } k \text{ red balls were drawn}\}.$$

Question. What is $\mathbb{P}(R_k)$ under the assumption of a Laplace model?

$\longrightarrow$ Here we have

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n) \mid 1 \le \omega_i \le N\} = \{1, \ldots, N\}^n$$

and thus $|\Omega| = N^n$. Hence under a Laplace model we have

$$\mathbb{P}(R_k) = \frac{|R_k|}{|\Omega|} = \frac{|R_k|}{N^n}.$$

Now WLOG we may assume that all red balls are numbered from 1 to K. Then we have

$$\omega \in R_k \iff \exists 1 \le i_1 < \ldots < i_k \le n \text{ such that } \omega_i \in \begin{cases} \{1, \ldots, K\} & \text{if } i \in \{i_1, \ldots, i_k\} \\ \{K+1, \ldots, N\} & \text{else,} \end{cases}$$

which shows that

$$|R_k| = \binom{n}{k} K^k (N-K)^{n-k}$$

and thus

$$\begin{aligned} \mathbb{P}(R_k) &= \frac{\binom{n}{k} K^k (N-K)^{n-k}}{N^n} = \frac{\binom{n}{k} K^k (N-K)^{n-k}}{N^k N^{n-k}} \\ &= \binom{n}{k} \left(\frac{K}{N}\right)^k \left(1 - \frac{K}{N}\right)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

for $p := \frac{K}{N}$. Hence we obtain a Binomial model with success probability p and number of trials n.

1.5 Conditional Probabilities

Let Ω be a finite or infinitely countable sample space and $\mathcal{A} = 2^\Omega$. Consider an event $B \in \mathcal{A}$ such that $\mathbb{P}(B) > 0$. We are now interested in the case where B occured.

Question. What is the probability of $A \in \mathcal{A}$ given that B already occured?

Definition 1.18. The *conditional probability* of A given B is defined by

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

Intuition. Consider $\mathbb{P}$ to be a finite distribution on a finite Ω. We know that then

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

holds for all events A. Now if it is known that B occurred, it is as if the whole Ω is replaces by B. Thus we arrive at

$$\mathbb{P}(A \mid B) = \frac{|A \cap B|}{|\Omega \cap B|} = \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Remark 1.19. If Ω is finite and $\mathbb{P}$ is the uniform distribution on $(\Omega, 2^\Omega)$ then the probability measure

$$2^\Omega \to [0,1],\ A \mapsto \mathbb{P}(A \mid B)$$

is again uniform on $(B, 2^B)$. In fact, for all $\omega \in B$ we have

$$\mathbb{P}(\{\omega\} \mid B) = \frac{\mathbb{P}(\{\omega\} \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{\omega\})}{\mathbb{P}(B)} = \frac{1/|\Omega|}{|B|/|\Omega|} = \frac{1}{|B|}.$$

Examples 1.20.

(1) Consider a fair die, so $\Omega = \{1, \ldots, 6\}$ and $\mathbb{P}(\{\omega\}) = \frac{1}{6}$. Let $A := \{1, 2, \ldots, 5\}$ and $B := \{2, 4, 6\}$. Then

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{|A \cap B|}{|B|} = \frac{2}{3}.$$

(2) Let X be a geometric random variable with success probability $p \in (0,1)$. For $r \in \mathbb{N}_0$ consider the event

$$W_r := \{X \geq r\} = \{\omega \in \Omega \mid X(\omega) \geq r\}.$$

For any $s > r$ let us compute

$$\mathbb{P}(W_s \mid W_r) = \frac{\mathbb{P}(W_s \cap W_r)}{\mathbb{P}(W_r)} = \frac{\mathbb{P}(W_s)}{\mathbb{P}(W_r)},$$

where

$$\begin{aligned} \mathbb{P}(W_r) = \mathbb{P}(X \geq r) &= \sum_{k=r}^{\infty} p(1-p)^k \\ &= p(1-p)^r \underbrace{\sum_{k=0}^{\infty} (1-p)^k}_{=\frac{1}{p}} = (1-p)^r. \end{aligned}$$

Hence we get

$$\mathbb{P}(W_s \mid W_r) = \frac{(1-p)^s}{(1-p)^r} = (1-p)^{s-r}.$$

Observe that this conditional probability depends only on the elapsed time $s - r$. This property is called the *memoryless property*.

Theorem 1.21 (Law of total probability)**.** *Let $(B_i)_{i \in I}$ be a partition of Ω, that is $\Omega = \bigsqcup_{i \in I} B_i$. Then for any $A \in \mathcal{A}$ we have*

$$\mathbb{P}(A) = \sum_{i:\, \mathbb{P}(B_i) > 0} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

Proof. Since $(B_i)_i$ is a partition of Ω, we have

$$A = \bigsqcup_{i \in I} (A \cap B_i)$$

and since all $A \cap B_i$ are pairwise disjoint, using sigma additivity we get

$$\begin{aligned}
\mathbb{P}(A) = \sum_{i \in I} \mathbb{P}(A \cap B_i) &= \sum_{i:\, \mathbb{P}(B_i)>0} \mathbb{P}(A \cap B_i) + \underbrace{\sum_{i:\, \mathbb{P}(B_i)=0} \overbrace{\mathbb{P}(A \cap B_i)}^{\leq \mathbb{P}(B_i)=0}}_{=0} \\
&= \sum_{i:\, \mathbb{P}(B_i)>0} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i),
\end{aligned}$$

which concludes the proof. □

Examples 1.22.

(1) Assume that the participation rate in the vote for a new mayor depends conditionally on the age of the voters as follows:

- $\frac{1}{4}$ if age $\in [18, 30]$,
- $\frac{1}{2}$ if age $\in (30, 50]$,
- $\frac{2}{3}$ if age $\in (50, \infty)$.

Furthermore, we know that the proportion of the voters

- whose age $\in [18, 30]$ is 20%,
- whose age $\in (30, 50]$ is 35%,
- whose age $\in (50, \infty)$ is 45%.

QUESTION. What is the global participation rate?

$\longrightarrow$ We have Ω = "Population of voters" $= \{\omega_1, \dots, \omega_N\}$ and the experiment is selecting an individual of Ω at random (meaning that $\mathbb{P}$ is the discrete uniform probability measure). Consider the events

$$\begin{aligned}
A_1 &:= \{\text{the person selected has age} \in [18, 30]\}, \\
A_2 &:= \{\text{the person selected has age} \in (30, 50]\}, \\
A_3 &:= \{\text{the person selected has age} \in (50, \infty)\}, \\
V &:= \{\text{the selected person participates in the vote}\}.
\end{aligned}$$

Then we have

$$\mathbb{P}(A_1) = 0.2, \quad \mathbb{P}(V \mid A_1) = 0.25,$$
$$\mathbb{P}(A_2) = 0.35, \quad \mathbb{P}(V \mid A_2) = 0.5,$$
$$\mathbb{P}(A_3) = 0.45, \quad \mathbb{P}(V \mid A_3) = \frac{2}{3}$$

and we want to compute $\mathbb{P}(V)$. Now by applying the Law of total probability 1.21 we get

$$\mathbb{P}(V) = \sum_{i=1}^{3} \mathbb{P}(V \mid A_i)\mathbb{P}(A_i) = 0.525.$$

(2) We have two urns:

- In urn 1 there are 2 white balls and 1 black ball.
- In urn 2 there are 3 white balls and 3 black balls.

Experiment. First, we select one urn at random where urn 1 is selected with probability p. Secondly, we select one ball from the selected urn uniformly at random.

Question. What is the probability of selecting a black ball?

$\longrightarrow$ Put

$$B := \{\text{the selected ball is black}\}, \quad A_i := \{\text{urn } i \text{ is selected}\}$$

for $i \in \{1, 2\}$. Using the Law of total probability 1.21 we have that

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B \mid A_1)\mathbb{P}(A_1) + \mathbb{P}(B \mid A_2)\mathbb{P}(A_2) \\ &= \frac{1}{3}p + \frac{1}{2}(1-p) = \frac{1}{2} - \frac{p}{6}.\end{aligned}$$

In the following, we are going to also provide a more detailed solution to the problem. An elementary outcome of this experiment is given by a pair (`urn`, `ball`) with `urn` $\in \{U_1, U_2\}$ and `ball` $\in \{W_1, W_2, B_1, W_1', W_2', W_3', B_1', B_2', B_3'\}$ such that

$$\Omega = \{(U_1, W_1), (U_2, W_2), (U_1, B_1), (U_2, W_1'), \ldots (U_2, B_3')\} =: \{\omega_1, \ldots, \omega_9\}.$$

Furthermore, for the weights $p(\omega) := \mathbb{P}(\{\omega\})$ we have

- $p(\omega_1) = p(\omega_2) = p(\omega_3)$,
- $p(\omega_4) = \ldots = p(\omega_9)$

and

$$\mathbb{P}(A_1) = p(\omega_1) + p(\omega_2) + p(\omega_3) = 3p(\omega_1) \overset{!}{=} p$$

which now implies

$$p(\omega_1) = p(\omega_2) = p(\omega_3) = \frac{p}{3}.$$

Similarly, we also get

$$p(\omega_4) = \ldots = p(\omega_9) = \frac{1-p}{6}$$

and thus we can now compute

$$\mathbb{P}(B) = p(\omega_3) + p(\omega_7) + p(\omega_8) + p(\omega_9) = \frac{1}{2} - \frac{p}{6}.$$

Theorem 1.23. *For any events $A_1, \ldots, A_n \in \mathcal{A}$ such that $\mathbb{P}(A_1 \cap \ldots \cap A_n) > 0$ we have*

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_n \mid A_1 \cap \ldots \cap A_{n-1}).$$

Proof. First, note that since $A_1 \cap \ldots \cap A_j \subseteq \bigcap_{i=1}^n A_i$ holds, we have

$$\mathbb{P}(A_1 \cap \ldots \cap A_j) \geq \mathbb{P}(A_1 \cap \ldots \cap A_n) > 0$$

by assumption for every $j \in \{1, \ldots, n\}$. Furthermore, by definition we have

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)$$

and thus by induction we get

$$\begin{aligned}\mathbb{P}(A_1 \cap \ldots \cap A_n) &= \mathbb{P}(A_1 \cap \ldots \cap A_{n-1})\mathbb{P}(A_n \mid A_1 \cap \ldots \cap A_{n-1}) \\ &= \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_n \mid A_1 \cap \ldots \cap A_{n-1})\end{aligned}$$

which concludes the proof. □

Example 1.24 (Birthday paradox)**.** Consider a group of n people.
QUESTION. What is the provability that 2 people share their birthday?
$\longrightarrow$ For simplicity, we assume that

- every year has 365 days,
- all days have the same probability to be a birthday.

Put

$$E := \{\text{at least 2 people have the same birthday}\}.$$

Note that if $n > 365$ then $\mathbb{P}(E) = 1$. Hence we assume that $n \leq 365$ holds. We have

$$E^c = \{\text{all n people have different birthdays}\}$$

and

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n) \mid 1 \leq \omega_i \leq 365\} = \{1, 2, \ldots, 365\}^n,$$

so

$$E^c = \{\omega \in \Omega \mid \forall i \neq j : \omega_i \neq \omega_j\} \subseteq \Omega. \tag{1}$$

Now consider the following events:

- $A_1 := \{\text{person 1 has birthday on some date} \in \{1, \ldots, 365\}\}$,
- $A_2 := \{\text{person 2 has birthday that is different from person 1}\}$,
- $A_2 := \{\text{person 3 has birthday that is different from person 1 and 2}\}$,
- $\ldots$
- $A_n := \{\text{person } n \text{ has birthday that is different from person } 1, 2, \ldots, n-1\}$.

Note that then $E^c = A_1 \cap \ldots \cap A_n$ and thus by Theorem 1.23 we get

$$\begin{aligned}\mathbb{P}(E^c) &= \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1) \cdots \mathbb{P}(A_n \mid A_1 \cap \ldots \cap A_{n-1}) \\ &= 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{366-n}{365}\end{aligned}$$

which implies

$$\mathbb{P}(E) = 1 - \prod_{i=1}^{n-1} \frac{365-i}{365}.$$

Now we are going to also present a solution with Ω and the assumption of a Laplace model. Under this assumption, by using (1) we get

$$\begin{aligned}\mathbb{P}(E) = 1 - \mathbb{P}(E^c) = 1 - \frac{|E^c|}{|\Omega|} &= 1 - \frac{\prod_{i=0}^{n-1}(365-i)}{365^n} \\ = 1 - \frac{\prod_{i=1}^{n-1}(365-i)}{365^{n-1}} &= 1 - \prod_{i=1}^{n-1} \frac{365-i}{365}.\end{aligned}$$

1.6 Bayes Rule

Let A and B be two events with $\mathbb{P}(A), \mathbb{P}(B) > 0$. Then we have

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

which is called *Bayes rule* and directly follows from the definition of conditional probability (Definition 1.18). Now we can combine Bayes rule with the Law of total probability 1.21 as follows: If we have $0 < \mathbb{P}(B) < 1$ then

$$\mathbb{P}(A) = \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c)$$

and thus

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c)}.$$

Example 1.25 (false positive / false negative)**.** In a certain population, the probability that and individual is affected by an illness K is $p = \frac{1}{100}$. It is possible to get tested but the test is not perfect. This means that if we set

$$B := \{\text{person has illness } K\}, T := \{\text{test is positive}\},$$

then we have $\mathbb{P}(T \mid B) = 0.96$ and $\mathbb{P}(T^c \mid B^c) = 0.94$, so with 4% we get a false negative and with 6% a false positive.
Question. What is the probability that a person is ill, given that he tested positive?
$\longrightarrow$ Using Bayes rule, we get

$$\mathbb{P}(B \mid T) = \frac{\mathbb{P}(T \mid B)\mathbb{P}(B)}{\mathbb{P}(T \mid B)\mathbb{P}(B) + \mathbb{P}(T \mid B^c)\mathbb{P}(B^c)} \approx 0.14.$$

Theorem 1.26. *Let $(B_i)_{i\in I}$ be a countable partition of Ω such that $\mathbb{P}(B_i) > 0$ for all $i \in I$. Then for every event $A \in \mathcal{A}$ with $\mathbb{P}(A) > 0$ we have*

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\sum_{j\in I} \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}.$$

1.7 Independence

Definition 1.27. A collection of events $(A_i)_{i\in I}$ is said to be *independent* if

$$\forall J \subseteq I \text{ with } |J| < \infty \text{ we have } \mathbb{P}\Big(\bigcap_{j\in J} A_j\Big) = \prod_{j\in J} \mathbb{P}(A_j).$$

In this case, we also say that the events $(A_i)_{i\in I}$ are *mutually independent*.

Remarks 1.28.

- Note that in Definition 1.27 the indexing set I can be arbitrary.
- For any event $A \in \mathcal{A}$ the collections $\{A, \Omega\}$ and $\{A, \emptyset\}$ are independent.
- For any events $A, B \in \mathcal{A}$ with $\mathbb{P}(A), \mathbb{P}(B) > 0$ we have that A and B are independent if and only if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ or $\mathbb{P}(B \mid A) = \mathbb{P}(B)$ holds.
- In Definition 1.27 the requirement "$\forall J \subseteq I$ with $|J| < \infty$" cannot be relaxed in general as shown in the following example.

Example 1.29. Consider the experiment of tossing a fair coin twice and the events

$$A := \{\text{the first toss results in heads}\},$$
$$B := \{\text{the second toss results in heads}\},$$
$$C := \{\text{the tosses result in different outcomes}\}.$$

Then $\Omega = \{0,1\}^2$ and under the assumption of a Laplace model, we have $\mathbb{P}(\{\omega\}) = \frac{1}{4}$ for all $\omega \in \Omega$. Furthermore, we have

$$\mathbb{P}(A) = \frac{1}{2}, \mathbb{P}(B) = \frac{1}{2}, \mathbb{P}(C) = \frac{1}{2}$$

and

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(B), \\
\mathbb{P}(A \cap C) &= \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(C), \\
\mathbb{P}(B \cap B) &= \frac{1}{4} = \mathbb{P}(B)\mathbb{P}(B),
\end{aligned}$$

which means that the events A, B, C are *pairwise independent*. But we have

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\emptyset) = 0 \neq \frac{1}{8} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$$

and thus the evens A, B, C are not mutually independent.

Lemma 1.30. *Let $(A_i)_{i \in I}$ be an independent collection of events. If we set $B_i := A_i$ or $B_i := A_i^c$ for all $i \in I$ then the new collection $(B_i)_{i \in I}$ is again independent.*

Definition 1.31. A collection of random variables $(X_i)_{i \in I}$ defined on the same probability space Ω is said to be *independent* if the events $\{X_i = x_i\}_{i \in I}$ are independent for every choice of $x_i \in X_i(\Omega)$.

Notation. In probability theory, when we have two functions $X : \Omega \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ we often write $g(X) := g \circ X$ to denote their composition.

Proposition 1.32. *Let $X_1, \ldots, X_n$ be independent random variables. Then for any functions $g_i : X_i(\Omega) \to \mathbb{R}$ for $1 \leq i \leq n$ such that $\mathbb{E}[|g_i(X_i)|] < \infty$ we have*

$$\mathbb{E}\left[\prod_{i=1}^{n} g_i(X_i)\right] = \prod_{i=1}^{n} \mathbb{E}[g_i(X_i)].$$

To prove this result, we need the following lemma.

Lemma 1.33. *Let $X : \Omega \to \mathbb{R}^d$ be a* random vector of dimension d, *which means that all its components $X_i : \Omega \to \mathbb{R}$ are random variables. If $g : X(\Omega) \to \mathbb{R}$ is any function with $g \geq 0$ or $\mathbb{E}[|g(X)|] < \infty$ then*

$$\mathbb{E}[g(X)] = \sum_{x \in X(\Omega)} g(x)\mathbb{P}(X = x).$$

Proof. In both cases, we have

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_{\omega\in\Omega} g(X(\omega))p(\omega) \\ &= \sum_{x\in X(\Omega)} \sum_{\omega:X(\omega)=x} g(X(\omega))p(\omega) \\ &= \sum_{x\in X(\Omega)} g(x) \sum_{\omega:X(\omega)=x} p(\omega) \\ &= \sum_{x\in X(\Omega)} g(x)\mathbb{P}(X=x), \qquad (1)\end{aligned}$$

where $p(\omega) := \mathbb{P}(\{\omega\})$ and at (1) we used σ-additivity of $\mathbb{P}$. $\square$

Proof of Proposition 1.32. Put $X := (X_1, \dots, X_n)$ and consider the function

$$h : X(\Omega) \to \mathbb{R}, (x_1, \dots, x_n) \mapsto \prod_{i=1}^{n} |g(x_i)|.$$

Then $h \geq 0$ and thus by Lemma 1.33 we can write

$$\mathbb{E}\left[\prod_{i=1}^{n} |g_i(X_i)|\right] = \mathbb{E}[h(X)] = \sum_{x\in X(\Omega)} h(x)\mathbb{P}(X=x).$$

Now note that by independence of $X_1, \dots, X_n$ we have

$$\begin{aligned}\mathbb{P}(X=x) &= \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i).\end{aligned}$$

Hence we get

$$\begin{aligned}\mathbb{E}[h(X)] &= \sum_{(x_1,\dots,x_n)\in X(\Omega)} \Big(\prod_{i=1}^{n} |g_i(x_i)|\Big) \cdot \Big(\prod_{i=1}^{n} \mathbb{P}(X_i = x_i)\Big) \\ &= \sum_{x_n}\sum_{x_{n-1}}\cdots\sum_{x_1}\prod_{i=1}^{n} |g_i(x_i)|\mathbb{P}(X_i = x_i) \\ &= \prod_{i=1}^{n}\sum_{x_i} |g_i(x_i)|\mathbb{P}(X_i = x_i) \\ &= \prod_{i=1}^{n} \mathbb{E}[|g_i(X_i)|] < \infty\end{aligned}$$

because we have $\mathbb{E}[|g_i(X_i)|] < \infty$ for all $1 \leq i \leq n$ by assumption. Hence $\mathbb{E}[\prod_{i=1}^{n} g_i(X_i)]$ is finite as well and by replacing h by $h(x) = \prod_{i=1}^{n} g_i(x_i)$ we can conclude. $\square$

Example 1.34 (false positive / false negative)**.** Recall the setting of Example 1.25. Let

$$T_1 := \{1^{\text{st}} \text{ test is positive}\}, \quad T_2 := \{2^{\text{nd}} \text{ test is positive}\}$$

and assume that

- $\mathbb{P}(T_1 \cap T_2 \mid B) = \mathbb{P}(T_1 \mid B)\mathbb{P}(T_2 \mid B)$,
- $\mathbb{P}(T_1 \cap T_2 \mid B^c) = \mathbb{P}(T_1 \mid B^c)\mathbb{P}(T_2 \mid B^c)$,
- $\mathbb{P}(T_1 \mid B)\mathbb{P}(T_2 \mid B) = 0.96$,
- $\mathbb{P}(T_1 \mid B^c)\mathbb{P}(T_2 \mid B^c) = 0.06$.

Then we have

$$\mathbb{P}(B \mid T_1 \cap T_2) = \frac{\mathbb{P}(T_1 \cap T_2 \mid B)\mathbb{P}(B)}{\mathbb{P}(T_1 \cap T_2 \mid B)\mathbb{P}(B) + \mathbb{P}(T_1 \cap T_2 \mid B^c)\mathbb{P}(B^c)} = 0.72.$$

by Bayes rule.

Example 1.35. Consider n independent tosses of a p-coin, so $\Omega = \{0,1\}^n$. For $1 \leq i \leq n$ define

$$X_i : \Omega \to \{0,1\}, \omega \mapsto \omega_i,$$

so X_i represents the (random) outcome of the i-th toss. Let $\mathbb{P}$ be the probability measure on $\mathcal{A} := 2^\Omega$ such that

(1) $\mathbb{P}(X_i = 1) = p$,

(2) $X_1, \ldots, X_n$ are independent.

Question. What is $\mathbb{P}$?
$\longrightarrow$ Note that (1) is equivalent to

$$\forall x \in \{0,1\} : \quad \mathbb{P}(X_i = x) = p^x(1-p)^{1-x}.$$

Now let $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$. Then

$$\begin{aligned} \mathbb{P}(\{\omega\}) = \mathbb{P}(X_1 = \omega_1, \ldots, X_n = \omega_n) &= \prod_{i=1}^n \mathbb{P}(X_i = \omega_i), \qquad (3) \\ &= \prod_{i=1}^n p^{\omega_i}(1-p)^{1-\omega_i} = p^{\sum_{i=1}^n \omega_i}(1-p)^{n-\sum_{i=1}^n \omega_i} \\ &= p^k(1-p)^k, \end{aligned}$$

for $k := \sum_{i=1}^n \omega_i = |\{i \mid \omega_i = 1\}|$ where at (3) we used independence. Hence $\mathbb{P}$ is uniquely given by

$$\mathbb{P}(\{\omega\}) = p^k(1-p)^{1-k}.$$

Now put

$$S_n := \sum_{i=1}^{n} X_i = \text{“the number of Heads obtained in } n\text{-tosses”},$$

so $S_n(\Omega) = \{0, 1, \dots, n\}$. For $k \in S_n(\Omega)$ we have

$$\begin{aligned}
\mathbb{P}(S_n = k) &= \mathbb{P}\Big(\sum_{i=1}^{n} X_i = k\Big) \\
&= \sum_{1 \le i_1 < \dots < i_k \le n} \mathbb{P}(X_{i_1} = 1, \dots, X_{i_k} = 1, X_i = 0, \forall i \notin \{i_1, \dots, i_k\}) \\
&= \binom{n}{k} p^k (1-p)^{n-k}.
\end{aligned}$$

Furthermore, we have

$$\mathbb{E}[S_n] = \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np$$

and

$$\begin{aligned}
\mathbb{E}[S_n^2] &= \sum_{k=0}^{n} k^2 \mathbb{P}(S_n = k) \\
&= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} = \dots
\end{aligned}$$

by Lemma 1.33. Note that this will result in an involved computation. A simpler way to compute $\mathbb{E}[S_n^2]$ is the following :

$$\begin{aligned}
\mathbb{E}[S_n^2] = \mathbb{E}\Big[\Big(\sum_{i=1}^{n} X_i\Big)^2\Big] &= \mathbb{E}\Big[\sum_{i=1}^{n} \underbrace{X_i^2}_{=X_i} + \sum_{1 \le i \ne j \le n} X_i X_j\Big] \\
&= \mathbb{E}[S_n] + \mathbb{E}\Big[\sum_{1 \le i \ne j \le n} X_i X_j\Big] \\
&= np + \sum_{1 \le i \ne j \le n} \underbrace{\mathbb{E}[X_i X_j]}_{=\mathbb{E}[X_i]\mathbb{E}[X_j] = p^2} \\
&= np + \sum_{1 \le i \ne j \le n} p^2 \\
&= np + (n^2 - n)p^2 = np + n^2 p^2 - np^2 \\
&= np(1-p) + \mathbb{E}[S_n]^2
\end{aligned}$$

and thus

$$\mathbb{E}[S_n^2] - \mathbb{E}[S_n]^2 = np(1-p)$$

which is the *variance* of S_n.

1.8 Conditional Expectation

Let Ω be countable, $\mathcal{A} = 2^\Omega$ and $\mathbb{P}$ the probability measure associated with given weights $\{p(\omega) \mid \omega \in \Omega\}$.

Definition 1.36. Let $B \in \mathcal{A}$ with $\mathbb{P}(B) > 0$. For a random variable X with $\mathbb{E}[|X|] < \infty$ we define the *conditional expectation* of X given B by

$$\mathbb{E}[X \mid B] := \frac{\mathbb{E}[X \mathbb{1}_B]}{\mathbb{P}(B)}.$$

Remarks 1.37.

- We have $\mathbb{E}[X \mid \Omega] = \mathbb{E}[X]$.
- We have
$$\mathbb{E}[X \mid B] = \frac{\sum_{\omega \in \Omega} X(\omega)\mathbb{1}_B(\omega)p(\omega)}{\mathbb{P}(B)} = \sum_{\omega \in \Omega} X(\omega)\frac{p(\omega)\mathbb{1}_B(\omega)}{\mathbb{P}(B)}.$$
- The map $\mathcal{A} \to [0,1], A \mapsto \mathbb{P}(A \mid B)$ is σ-additive In fact, if $(A_i)_{i \in \mathbb{N}} \subseteq \mathcal{A}$ is a collection of pairwise disjoint sets, then
$$\begin{aligned} \mathbb{P}\Big(\bigsqcup_i A_i \mid B\Big) &= \frac{\mathbb{P}(\bigsqcup_i A_i \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(\bigsqcup_i (A_i \cap B))}{\mathbb{P}(B)} \\ &= \frac{\sum_i \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} \\ &= \sum_i \mathbb{P}(A_i \mid B). \end{aligned}$$
Using this, we get
$$\mathbb{E}[X \mid B] = \sum_{x \in X(\Omega)} \mathbb{P}(X = x \mid B).$$

Definition 1.38. Let $\mathcal{B} = (B_i)_i$ be a partition of Ω with $\mathbb{P}(B_i) > 0$ and let X be a random variable on Ω with $\mathbb{E}[|X|] < \infty$. Then we define

$$\mathbb{E}[X \mid \mathcal{B}] : \Omega \to \mathbb{R},\ \omega \mapsto \sum_i \mathbb{E}[X \mid B_i]\mathbb{1}_{B_i}(\omega),$$

called the *conditional expectation map* of X given $\mathcal{B}$.

Convention. Note that in Definition 1.36 we assumed $\mathbb{P}(B) > 0$. For the following definition, we use the convention $\mathbb{E}[X \mid B] = 0$ if $\mathbb{P}(B) = 0$.

Definition 1.39. Let X and Y be two random variables on Ω such that $\mathbb{E}[|X|] < \infty$ and let $\{y_i\}_i = Y(\Omega)$. Then the *conditional expectation* of X given Y is defined to be the random variable

$$\mathbb{E}[X \mid Y] := \mathbb{E}[X \mid \{Y = y_i\}_i] = \sum_i \mathbb{E}[X \mid Y = y_i] \mathbb{1}_{\{Y=y_i\}}.$$

Proposition 1.40. *Let X, X' and Y be random variables on Ω such that $\mathbb{E}[|X|], \mathbb{E}[|X'|] < \infty$.*

(i) The conditional expectation map is linear, so for all $\alpha \in \mathbb{R}$ we have

$$\mathbb{E}[\alpha X + X' \mid Y] = \alpha \mathbb{E}[X \mid Y] + \mathbb{E}[X' \mid Y].$$

(ii) Let $g : X(\Omega) \to \mathbb{R}$ be a map with $\mathbb{E}[|g(X)|] < \infty$. Then we have

$$\mathbb{E}[g(X) \mid X] = g(X).$$

(iii) If X and Y are independent then $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$.

(iv) We always have $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$, called the tower rule.

Proof.

(i) Follows from linearity of the expectation.

(ii) Set $X(\Omega) = (x_i)_i$. Then by definition we have

$$\mathbb{E}[g(X) \mid X] = \sum_i \mathbb{E}[g(X) \mid X = x_i] \mathbb{1}_{\{X=x_i\}}.$$

Now, for $\mathbb{P}(X = x_i) > 0$, compute

$$\begin{aligned}
\mathbb{E}[g(X) \mid X = x_i] &= \frac{\mathbb{E}[g(X)\mathbb{1}_{\{X=x_i\}}]}{\mathbb{P}(X = x_i)} \\
&= \frac{\mathbb{E}[g(x_i)\mathbb{1}_{\{X=x_i\}}]}{\mathbb{P}(X = x_i)} \\
&= g(x_i)\frac{\mathbb{E}[\mathbb{1}_{\{X=x_i\}}]}{\mathbb{P}(X = x_i)} \\
&= g(x_i)\frac{\mathbb{P}(X = x_i)}{\mathbb{P}(X = x_i)} = g(x_i).
\end{aligned}$$

Hence we get

$$\mathbb{E}[g(X) \mid X] = \sum_i g(x_i) \mathbb{1}_{\{X=x_i\}} = g(X).$$

(iii) Write $Y(\Omega) = \{y_i\}_i$ and observe that

$$\begin{aligned}
\mathbb{E}[X \mid Y] &= \sum_i \mathbb{E}[X \mid Y = y_i]\mathbb{1}_{\{Y=y_i\}} \\
&= \sum_i \sum_{x \in X(\Omega)} x\mathbb{P}(X = x \mid Y = y_i)\mathbb{1}_{\{Y=y_i\}} \\
&= \sum_i \sum_{x \in X(\Omega)} x\mathbb{P}(X = x)\mathbb{1}_{\{Y=y_i\}} \qquad (1) \\
&= \mathbb{E}[X],
\end{aligned}$$

where at (1) we used independence of X, Y.

(iv) By definition, we have

$$\mathbb{E}[X \mid Y] = \sum_i \mathbb{E}[X \mid Y = y_i]\mathbb{1}_{\{Y=y_i\}}$$

and thus

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}\Big[\sum_i \mathbb{E}[X \mid Y = y_i]\mathbb{1}_{\{Y=y_i\}}\Big].$$

Now using the triangular inequality, we get

$$\begin{aligned}
\mathbb{E}[|\mathbb{E}[X \mid Y]|] &\leq \mathbb{E}\Big[\sum_i |\mathbb{E}[X \mid Y = y_i]|\mathbb{1}_{\{Y=y_i\}}\Big] \\
&= \sum_i |\mathbb{E}[X \mid Y = y_i]|\mathbb{E}[\mathbb{1}_{Y=y_i}] \qquad (2) \\
&= \sum_i |\mathbb{E}[X \mid Y = y_i]|\mathbb{P}(Y = y_i) \qquad (2) \\
&= \sum_i \Big| \sum_{x \in X(\Omega)} x\mathbb{P}(X = x \mid Y = y_i)\Big|\mathbb{P}(Y = y_i) \qquad (2) \\
&\leq \sum_i \sum_{x \in X(\Omega)} |x|\mathbb{P}(X = x \mid Y = y_i)\mathbb{P}(Y = y_i) \\
&= \sum_{x \in X(\Omega)} |x| \sum_i \mathbb{P}(X = x, Y = y_i) \\
&= \sum_{x \in X(\Omega)} |x|\mathbb{P}(X = x) = \mathbb{E}[|X|] < \infty.
\end{aligned}$$

where at (2) we used linearity. A similar computation now shows that

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$$

holds. □

Proposition 1.41 (Jensen's inequality)**.** *Let X be a random variable with $\mathbb{E}[X^2] < \infty$. Then we have*

$$\mathbb{E}[|X|] \leq \sqrt{\mathbb{E}[X^2]}$$

and thus $\mathbb{E}[|X|] < \infty$.

Proof. Observe that we have

$$\begin{aligned}
\mathbb{E}[|X|] &= \sum_{x \in X(\Omega)} |x| \mathbb{P}(X = x) \\
&= \sum_{x \in X(\Omega)} |x| \sqrt{\mathbb{P}(X = x)} \sqrt{\mathbb{P}(X = x)} \\
&\leq \sqrt{\sum_{x \in X(\Omega)} x^2 \mathbb{P}(X = x)} \underbrace{\sqrt{\sum_{x \in X(\Omega)} \mathbb{P}(X = x)}}_{=1} \qquad (1) \\
&= \sqrt{\mathbb{E}[X^2]} < \infty,
\end{aligned}$$

where at (1) we used the Cauchy-Schwarz inequality. □

Now consider the function

$$\psi(c) := \mathbb{E}[(X - c)^2].$$

QUESTION. What is the value of $\arg\min_{c \in \mathbb{R}} \psi(c)$?

Observe that we have

$$\psi(c) = \mathbb{E}[X^2 - 2Xc + c^2] = \mathbb{E}[X^2] - 2c\mathbb{E}[X] + c^2$$

and thus

$$\psi'(c) = -2\mathbb{E}[X] + 2c = 2(c - \mathbb{E}[X]) \overset{!}{=} 0. \iff c = \mathbb{E}[X]$$

Hence $c^* := \mathbb{E}[X]$ is the unique stationary point of ψ, which is a strictly convex function, so

$$\arg\min_{c \in \mathbb{R}} \mathbb{E}[(X - c)^2] = \mathbb{E}[X].$$

Theorem 1.42. *Let X be a random variable on Ω such that $\mathbb{E}[X^2] < \infty$. If $\mathcal{B} = (B_i)_i$ ias a partition of Ω such that $\mathbb{P}(B_i) > 0$ then*

$$\mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{B}])^2] = \underset{c_i : \sum_i c_i^2 \mathbb{P}(B_i) < \infty}{\arg\min} \mathbb{E}\Big[\Big(X - \sum_i c_i \mathbb{1}_{B_i}\Big)^2\Big].$$

Corollary 1.43. *Let X and Y be two random variables on Ω such that $\mathbb{E}[|X|] < \infty$. Then*

$$\mathbb{E}[X \mid Y] = \underset{\substack{g: Y(\Omega) \to \mathbb{R} \\ \mathbb{E}[g(Y)^2] < \infty}}{\arg\min} \mathbb{E}[(X - g(Y))^2].$$

We omit the proof of Theorem 1.42.

Proof of Corollary 1.43. By definition we have

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid \mathcal{B}]$$

for $\mathcal{B} := (B_i)_i$ with $B_i = \{Y = y_i\}$ if we write $Y(\Omega) = \{y_i\}_i$. By Theorem 1.42 we thus know that

$$\mathbb{E}[X \mid Y] = \underset{c_i : \sum_i c_i^2 \mathbb{P}(Y = y_i) < \infty}{\arg\min} \mathbb{E}\Big[\Big(X - \sum_i c_i \mathbb{1}_{\{Y = y_i\}}\Big)^2\Big].$$

Now for any given $(c_i)_i$ there exists a function

$$g : Y(\Omega) \to \mathbb{R}$$

such that $g(Y) = \sum_i c_i \mathbb{1}_{\{Y=y_i\}}$. Conversely, let $g : Y(\Omega) \to \mathbb{R}$ be any function and define $c_i := g(y_i)$. Then again $g(Y) = \sum_i c_i \mathbb{1}_{\{Y=y_i\}}$ holds. Now observe that then

$$\sum_i c_i^2 \mathbb{P}(Y = y_i) = \sum_i g(y_i)^2 \mathbb{P}(Y = y_i) = \mathbb{E}[g(Y)^2]$$

which concludes the proof. $\square$

Examples 1.44.

(1) Let X, Y be two random variables defined on Ω with

$$X(\Omega) = Y(\Omega) = \{0, 1\},$$

$p \in (0, 1)$ and

$$\begin{aligned} \mathbb{P}(X = Y = 0) &= \frac{p}{2}, \\ \mathbb{P}(X = 0, Y = 1) &= \frac{1-p}{2}, \\ \mathbb{P}(X = 1, Y = 0) &= \frac{1-p}{2}, \\ \mathbb{P}(X = Y = 1) &= \frac{p}{2}. \end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}[X \mid Y] &= \sum_{y \in Y(\Omega)} \mathbb{E}[X \mid Y = y] \mathbb{1}_{\{Y=y\}} \\
&= \mathbb{E}[X \mid Y = 0] \mathbb{1}_{\{Y=0\}} + \mathbb{E}[X \mid Y = 1] \mathbb{1}_{\{Y=1\}} \\
&= \mathbb{P}(X = 1 \mid Y = 0) \mathbb{1}_{\{Y=0\}} + \mathbb{P}(X = 1 \mid Y = 1) \mathbb{1}_{\{Y=1\}} \\
&= (1-p) \mathbb{1}_{\{Y=0\}} + p \mathbb{1}_{\{Y=1\}}.
\end{aligned}$$

Similarly, we get

$$\mathbb{E}[Y \mid X] = (1-p) \mathbb{1}_{X=0} + p \mathbb{1}_{X=1}.$$

(2) Consider random variables X, Y on Ω such that $Y \sim \text{Pois}(\lambda)$ for $\lambda > 0$ and conditionally on $Y = k$, X has a binomial distribution with success probability $p \in (0,1)$ and k number of trials. Then

$$\mathbb{P}(Y = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

and

$$\mathbb{P}(X = j \mid Y = k) = \begin{cases} 0 & \text{if } j > k \\ \binom{k}{j} p^j (1-p)^{k-j} & \text{if } j \leq k. \end{cases}$$

Question. What are $\mathbb{E}[X \mid Y]$ and $\mathbb{E}[X]$?

$\longrightarrow$ Here we can use that fact that if $S \sim \text{Binomial}(n, p)$ then $\mathbb{E}[S] = np$. Hence we have

$$\mathbb{E}[X \mid Y = k] = kp$$

and thus

$$\begin{aligned}
\mathbb{E}[X \mid Y] &= \sum_{k \geq 0} kp \mathbb{1}_{\{Y=k\}} \\
&= p \sum_{k \geq 0} k \mathbb{1}_{\{Y=k\}} = pY.
\end{aligned}$$

Now using the tower rule, we get

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[pY] = p\mathbb{E}[Y] = p\lambda.$$

(3) Let X and Y be two independent random variables such that

$$X \sim \text{Pois}(\lambda_1), \quad Y \sim \text{Pois}(\lambda_2)$$

for $\lambda_1, \lambda_2 > 0$ and put $S := Y_1 + Y_2 \sim \text{Pois}(\lambda_1 + \lambda_2)$.

QUESTION. What is $\mathbb{E}[X \mid S]$?

$\longrightarrow$ To determine this, we need to compute $\mathbb{P}(X = x \mid S = s)$ for all $x, s \in \mathbb{N}_0$. Observe that

$$\begin{aligned}\mathbb{P}(X = x \mid S = s) &= \frac{\mathbb{P}(X = x, S = s)}{\mathbb{P}(S = s)} \\ &= \frac{\mathbb{P}(X = x, S = s)}{\sum_{x'} \mathbb{P}(X = x', S = s)}.\end{aligned}$$

We have

$$\{X = x, S = s\} = \{X = x, X + Y = s\} = \{X = x, Y = s - x\}$$

and thus

$$\mathbb{P}(X = x, S = s) = \begin{cases} 0 & \text{if } x > s \\ \mathbb{P}(X = x)\mathbb{P}(Y = s - x) = \frac{e^{-\lambda_1}\lambda_1^x}{x!}\frac{e^{-\lambda_2}\lambda_2^{s-x}}{(s-x)!} & \text{if } x \leq s. \end{cases}$$

Hence

$$\begin{aligned}\sum_{x'} \mathbb{P}(X = x', S = s) &= \sum_{x'=0}^{s} \mathbb{P}(X = x', Y = s - x') \\ &= \sum_{x'=0}^{s} \frac{e^{-\lambda_1}\lambda_1^{x'}}{x'!}\frac{e^{-\lambda_2}\lambda_2^{s-x'}}{(s-x')!} \\ &= \frac{e^{-\lambda_1-\lambda_2}}{s!}\sum_{x'=0}^{s}\binom{s}{x'}\lambda_1^{x'}\lambda_2^{s-x'} \\ &= \frac{e^{-\lambda_1-\lambda_2}}{s!}(\lambda_1 + \lambda_2)^s \\ &= \mathbb{P}(S = s).\end{aligned}$$

Now we get

$$\begin{aligned}\mathbb{P}(X = x \mid S = s) &= \frac{s!}{x!(s-x)!}\frac{\lambda_1^x\lambda_2^{s-x}}{(\lambda_1 + \lambda_2)^s} \\ &= \binom{s}{x}\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^x\left(1 - \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^{s-x}\end{aligned}$$

and thus

$$X \mid S = s \sim \text{Binomial}\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}, s\right),$$

where $X \mid S = s$ denotes the random variable X given the event $S = s$. Hence

$$\mathbb{E}[X \mid S = s] = \frac{\lambda_1}{\lambda_1 + \lambda_2}s$$

and thus

$$\begin{aligned}\mathbb{E}[X \mid S] &= \sum_{s\geq 0} \mathbb{E}[X \mid S = s]\mathbb{1}_{\{S=s\}} \\ &= \sum_{s\geq 0} \frac{\lambda_1}{\lambda_1+\lambda_2} s\mathbb{1}_{\{S=s\}} = \frac{\lambda_1}{\lambda_1+\lambda_2} S.\end{aligned}$$

In particular, if $\lambda_1 = \lambda_2$ then $\mathbb{E}[X \mid S] = \frac{S}{2}$.

Note that in the case of $\lambda_1 = \lambda_2$ there is a quicker method to compute $\mathbb{E}[X \mid S]$. Observe that in this case

$$\begin{aligned}\mathbb{E}[X \mid S] &= \mathbb{E}[X + Y - Y \mid S] \\ &= \mathbb{E}[S \mid S] - \mathbb{E}[Y \mid S] \\ &= S - \mathbb{E}[X \mid S]\end{aligned}$$

since X and Y “play the same role” and thus $\mathbb{E}[X \mid S] = \frac{S}{2}$.

2 Random Walks

2.1 Introduction

Let $N \in \mathbb{N}$ and consider the sample space $\Omega := \{-1, 1\}^N$ with the uniform probability measure on $\mathcal{A} = 2^\Omega$, that is

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{2^N}$$

for all $A \in \mathcal{A}$.

Definition 2.1. Set $S_0(\omega) := 0$ and for $1 \leq n \leq N$

$$S_n(\omega) := \sum_{k=1}^{n} \omega_i$$

for $\omega = (\omega_1, \ldots, \omega_N) \in \Omega$. Then the sequence $(S_n)_{n \geq 0}$ is called a *random walk* with N steps starting at 0.

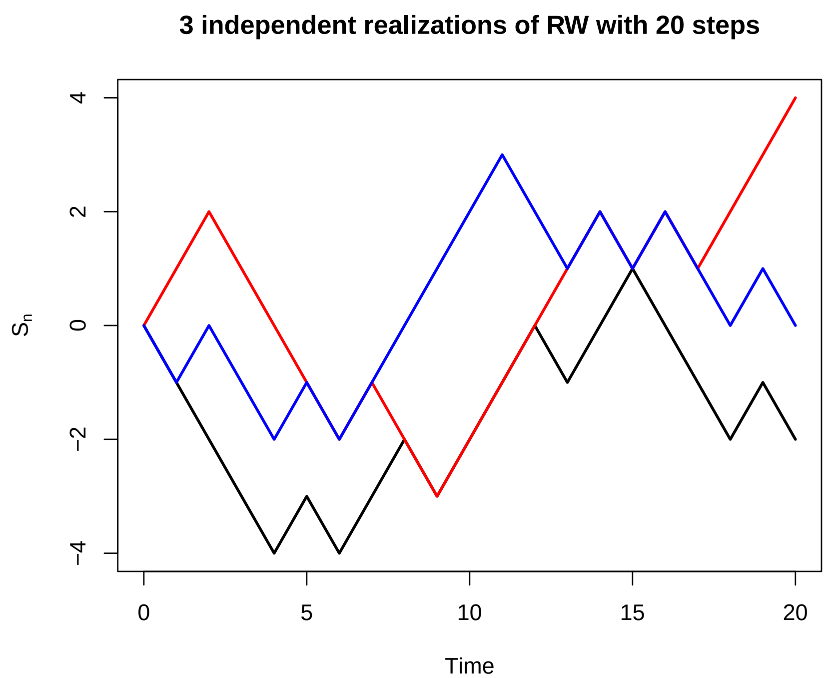


Figure 1: Three random walks starting at 0 with $N = 20$.

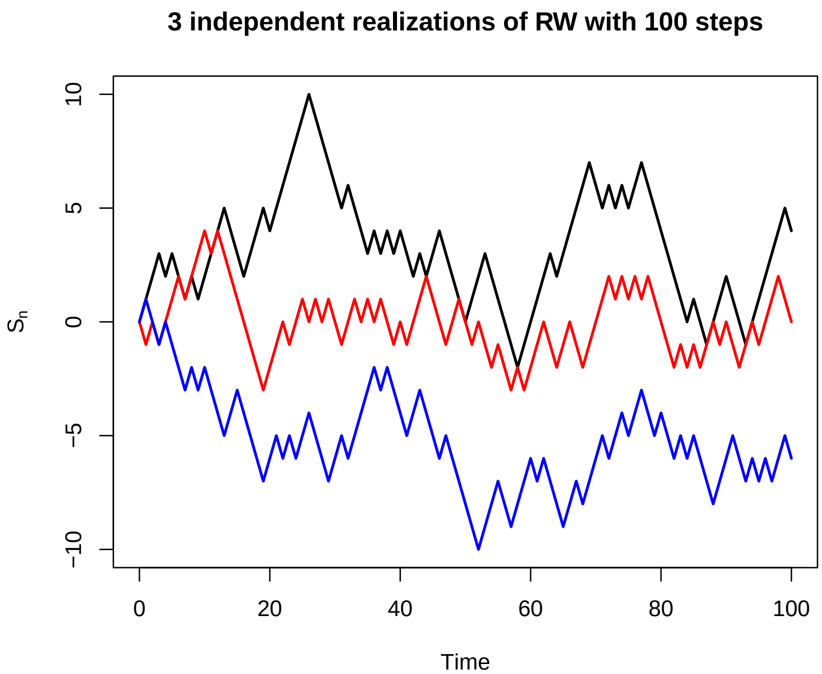


Figure 2: Three random walks starting at 0 with $N = 100$.

For $k \in \{1, \ldots, N\}$ set $X_k(\omega) := \omega_i$, so we have

$$S_n = \sum_{k=1}^{n} X_k.$$

Now observe that

$$|\{X_k = 1\}| = |\{\omega \in \Omega \mid X_k(\omega) = 1\}| = 2^{N-1}$$

and thus $\mathbb{P}(X_k = 1) = \frac{1}{2}$. Now fix integers $1 \leq k_1 < \ldots < k_l \leq N$ and $x_{k_1}, \ldots, x_{k_l} \in \{-1, 1\}$. Then we have

$$|\{X_{k_1} = x_{k_1}, \ldots, X_{k_l} = x_{k_l}\}| = 2^{N-l}$$

and thus

$$\mathbb{P}(X_{k_1} = x_{k_1}, \ldots, X_{k_l} = x_{k_l}) = \frac{1}{2^l}.$$

But this means that

$$\mathbb{P}\Big(\bigcap_{j=1}^{l} \{X_{k_j} = x_{k_j}\}\Big) = \frac{1}{2^l} = \prod_{j=1}^{l} \underbrace{\mathbb{P}(X_{k_j} = x_{k_j})}_{=\frac{1}{2}}$$

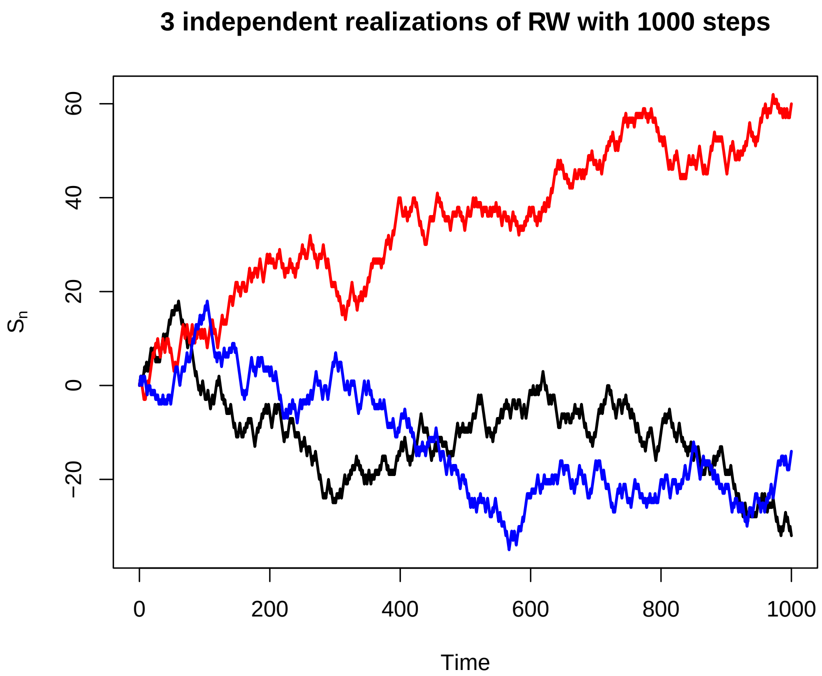


Figure 3: Three random walks starting at 0 with $N = 1000$.

which proves that $X_1, \dots, X_N$ are independent.

Remarks 2.2.

- For a given Ω, the graph of the points $(n, S_n(\omega))_{0\leq n\leq N}$ is called the *trajectory* of the random walk.

- For $k \in \{1, \dots, N\}$ we have

$$\mathbb{E}[X_k] = 0.$$

  We can even say that $X_k = 2U_k - 1$ holds for $U_k := \mathbb{1}_{\{X_k=1\}}$ and $U_1, \dots, U_N \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\frac{1}{2})$, so $U_1, \dots, U_N$ can be viewed as the outcome of tossing a fair coin N times.

- We have

$$\mathbb{E}[S_n] = \mathbb{E}\Big[\sum_{k=1}^{N} X_k\Big] = \sum_{k=1}^{N} \mathbb{E}[X_k] = 0.$$

Theorem 2.3. *Let $n \in \{1, \dots, N\}$. Then we have*

$$S_n(\Omega) = \{2k - n \mid 0 \leq k \leq n\}.$$

Moreover, for $k \in \{0, \dots, n\}$ we have

$$\mathbb{P}(S_n = 2k - n) = \binom{n}{k} 2^{-n}.$$

Proof. By Remarks 2.2 we know that

$$X_k = 2U_k - 1$$

for $1 \leq k \leq N$ where $U_1, \dots, U_N \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\frac{1}{2})$. Then by definition of S_n we get

$$S_n = 2\sum_{k=1}^{n} U_k - n$$

where $\sum_{k=1}^{n} U_k \sim \text{Binomail}(n, \frac{1}{2})$. Hence we have

$$S_n(\Omega) = \{2k - n \mid 0 \leq k \leq n\}$$

and

$$\begin{aligned}
\mathbb{P}(S_n = 2k - 1) &= \mathbb{P}\Big(\sum_{j=1}^{n} U_j = k\Big) \\
&= \binom{n}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{n-k} \\
&= \binom{n}{k} 2^{-n}
\end{aligned}$$

which concludes the proof. $\square$

Lemma 2.4. *Let $(S_n)_{1 \leq n \leq N}$ be a random walk with N steps starting at 0. Then for all $n \in \{1, \dots, N\}$ we have*

$$\max_{x \in S_n(\Omega)} \mathbb{P}(S_n = x) = \begin{cases} \mathbb{P}(S_n = 0) & \text{if } n \text{ is even} \\ \mathbb{P}(S_n = 1) = \mathbb{P}(S_n - 1) & \text{if } n \text{ is odd.} \end{cases}$$

Proof. For $k \in \{0, \dots, n\}$ we have

$$\begin{aligned}
C &:= \frac{\mathbb{P}(S_n = 2k - n)}{\mathbb{P}(S_n = 2(k-1) - n)} = \frac{\binom{n}{k} 2^{-n}}{\binom{n}{k-1} 2^{-n}} \\
&= \frac{(k-1)!(n-k+1)!}{k!(n-k)!} = \frac{n-k+1}{k} \geq 1 \\
&\iff n - k + 1 \geq k \iff n + 1 \geq 2k \\
&\iff k \leq \frac{n+1}{2}.
\end{aligned}$$

This means that if n is even, then $C \geq 1$ if and only if $k \in \{0, \dots, \frac{n}{2}\}$. In this case, the function $k \mapsto \mathbb{P}(S_n = 2k - n)$ is increasing on $\{0, \dots, \frac{n}{2}\}$ and decreasing on $\{\frac{n}{2}, \dots, n\}$. Hence $\mathbb{P}(S_n = 2k - n)$ is maximal for $k = \frac{n}{2}$, so

$$\max_{0 \leq k \leq n} \mathbb{P}(S_n = 2k - n) = \mathbb{P}(S_n = 0).$$

A similar analysis proves the other case where n is odd. $\square$

Remark 2.5. With the help of *Stirling's formula*

$$n! \underset{n \to \infty}{\sim} \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

we can show that

$$\mathbb{P}(S_n = 0) \underset{n \to \infty}{\sim} \frac{1}{\sqrt{\pi \frac{n}{2}}}$$

holds if n is even. The same holds for $\mathbb{P}(S_n = \pm 1)$ if n is odd.

2.2 The Reflection Principle

Let $(S_n)_{1 \leq n \leq N}$ be a random walk starting at 0.
For a given $c \in \bigcup_{n=0}^{N} S_n(\Omega) = \{-N, \dots, N\}$ define

$$T_c(\omega) := \min(\{1 \leq n \leq N \mid S_n(\omega) = c\} \cup \{N+1\}).$$

Lemma 2.6 (Reflection principle)**.** *For $a > 0$ and $b \geq -a$ we have*

$$\mathbb{P}(T_{-a} \leq n, S_n = b) = \mathbb{P}(S_n = -2a - b).$$

Theorem 2.7. *For $a > 0$ we have*

$$\begin{aligned} \mathbb{P}(T_{-a} \leq n) &= 2\mathbb{P}(S_n < -a) + \mathbb{P}(S_n = -a) \\ &= \mathbb{P}(S_n \notin (-a, a]). \end{aligned}$$

Proof. Observe that we have

$$\{T_{-a} \leq n\} = \bigsqcup_{b=-\infty}^{\infty} \{T_{-a} \leq n, S_n = b\}.$$

Hence by σ-additivity we get

$$\begin{aligned} \mathbb{P}(T_{-a} \leq n) &= \sum_{b=-\infty}^{\infty} \mathbb{P}(T_{-a} \leq n, S_n = b) \\ &= \underbrace{\sum_{b=-\infty}^{-a-1} \mathbb{P}(T_{-a} \leq n, S_n = b)}_{=:C_1} + \underbrace{\sum_{b=-a}^{\infty} \mathbb{P}(T_{-a} \leq n, S_n = b)}_{=:C_2}. \end{aligned}$$

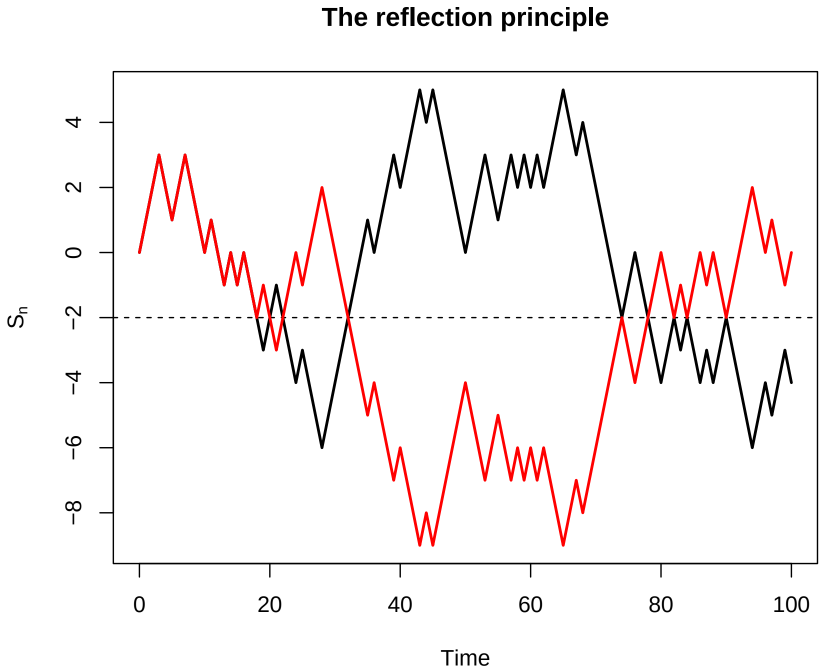


Figure 4: A random walk with $N = 100$ reflected at $-a = -2$.

Now using the Reflection principle 2.6, we can write

$$\begin{aligned} C_2 &= \sum_{b=-a}^{\infty} \mathbb{P}(S_n = -2a - b) \\ &= \sum_{t=-\infty}^{-a} \mathbb{P}(S_n = t) = \mathbb{P}(S_n \leq -a). \end{aligned}$$

Furthermore, we have

$$\begin{aligned} C_1 &= \sum_{b=-\infty}^{-a-1} \mathbb{P}(T_{-a} \leq n, S_n = b) \\ &= \sum_{b=-\infty}^{-a-1} \mathbb{P}(S_n = b) \\ &= \mathbb{P}(S_n \leq -a - 1) = \mathbb{P}(S_n < -a). \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{P}(T_{-a} \leq n) &= \mathbb{P}(S_n < -a) + \mathbb{P}(S_n \leq -a) \\ &= 2\mathbb{P}(S_n < -a) + \mathbb{P}(S_n = -a). \end{aligned}$$

Now we show that $\mathbb{P}(T_{-a} \leq n) = \mathbb{P}(S_n \notin (-a, a])$. Observe that we have

$$\begin{aligned}
\mathbb{P}(S_n < -a) &= \sum_{t<-a} \mathbb{P}(S_n = t) \\
&= \sum_{t<-a} \mathbb{P}(S_n = -t) \\
&= \sum_{z>a} \mathbb{P}(S_n = z) = \mathbb{P}(S_n > a).
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{P}(T_{-a} \leq n) &= 2\mathbb{P}(S_n < -a) + \mathbb{P}(S_n = -a) \\
&= \mathbb{P}(S_n < -a) + \mathbb{P}(S_n < -a) + \mathbb{P}(S_n = -a) \\
&= \mathbb{P}(S_n > a) + \mathbb{P}(S_n \leq -a) \\
&= \mathbb{P}(\{S_n > a\} \sqcup \{S_n \leq -a\}) \\
&= \mathbb{P}(S_n \notin (-a, a])
\end{aligned}$$

which concludes the proof. □

Corollary 2.8. *For $a \neq 0$ we have*

- $\mathbb{P}(T_a > N) \searrow 0$ *as* $N \to \infty$,
- $\mathbb{E}[T_a] \nearrow \infty$ *as* $N \to \infty$.

Proof. Observe that we have

$$\begin{aligned}
\mathbb{P}(T_a > N) &= \mathbb{P}(S_n \in (-a, a]) \\
&= \sum_{k=-a+1}^{a} \mathbb{P}(S_N = k) \\
&\leq 2a \cdot \begin{cases} \mathbb{P}(S_N = 0) & \text{if } N \text{ is even} \\ \mathbb{P}(S_N = 1) & \text{if } N \text{ is odd} \end{cases} \\
&\underset{n\to\infty}{\sim} 2a \frac{1}{\sqrt{\pi \frac{n}{2}}} \xrightarrow{n\to\infty} 0.
\end{aligned}$$

Now by Lemma 1.15 we have

$$\mathbb{E}[T_a] = \sum_{n=0}^{\infty} \mathbb{P}(T_a > n).$$

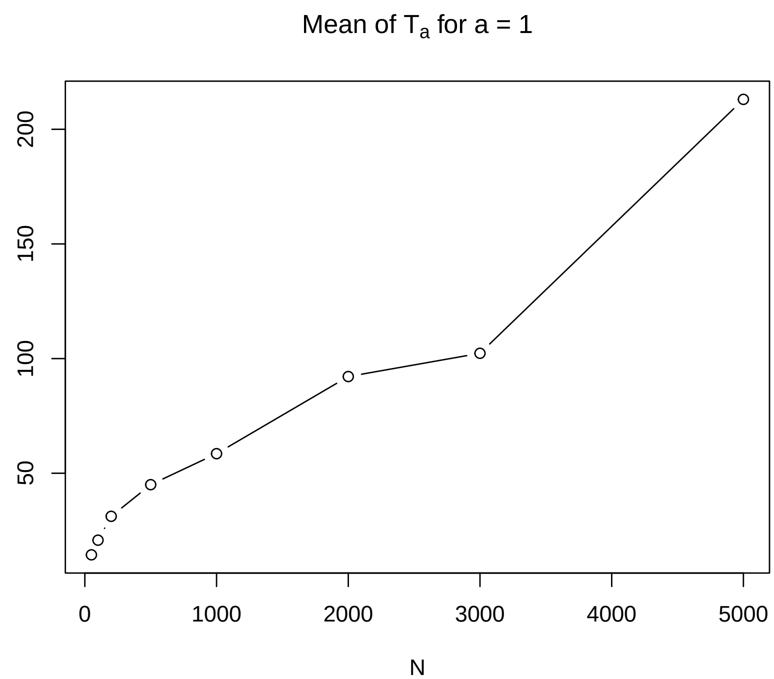


Figure 5: $\mathbb{E}[T_a]$ for $a = 1$ as N grows.

Note that $\{T_a > n\} = \emptyset$ for $n \geq N + 1$ and thus

$$\begin{aligned}
\mathbb{E}[T_a] &= \sum_{n=0}^{N} \mathbb{P}(T_a > n) \\
&= \sum_{n=0}^{N} \mathbb{P}(S_n \in (-a, a]) \\
&\geq \sum_{n=0}^{N} \mathbb{P}(S_n = 0) \\
&\geq \begin{cases} \sum_{k=0}^{\frac{N}{2}} \mathbb{P}(S_{2k} = 0) & \text{if } N \text{ is even} \\ \sum_{k=0}^{\frac{N-1}{2}} \mathbb{P}(S_{2k} = 0) & \text{if } N \text{ is odd.} \end{cases}
\end{aligned}$$

But we also have $\mathbb{P}(S_{2k} = 0) \underset{n\to\infty}{\sim} \frac{1}{\sqrt{\pi k}}$ and since

$$\sum_{k\geq 1} \frac{1}{\pi\sqrt{k}} = \infty$$

we can conclude that

$$\mathbb{E}[T_a] \xrightarrow{N\to\infty} \infty$$

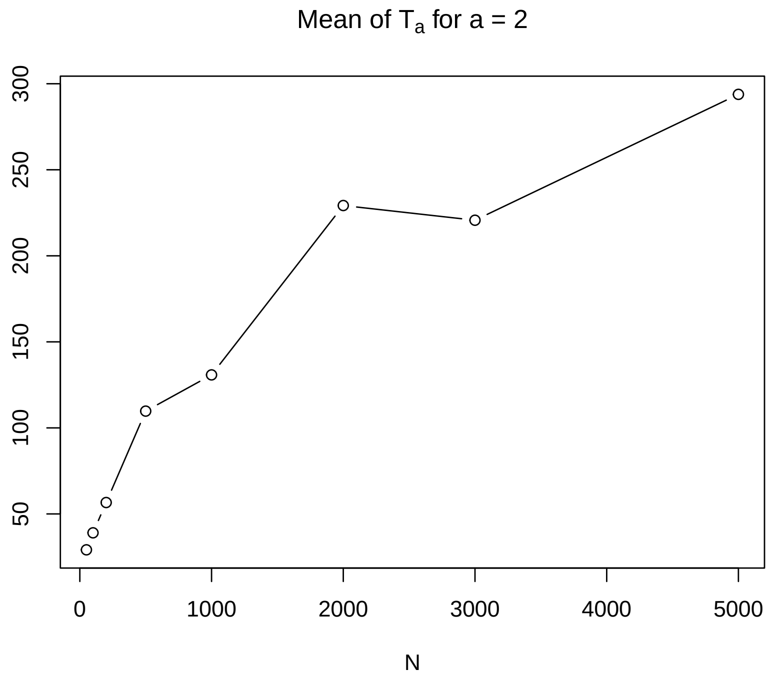


Figure 6: $\mathbb{E}[T_a]$ for $a = 2$ as N grows.

holds. $\square$

Theorem 2.9. *For $N \in \mathbb{N}$ and $2n \leq N$ it holds that*

$$\mathbb{P}(T_0 > 2n) = \mathbb{P}(S_{2n} = 0).$$

Example 2.10. Take $N \geq 3$ and $n = 1$. We want to check that Theorem 2.9 holds, i.e. that we have

$$\mathbb{P}(T_0 > 2) = \mathbb{P}(S_2 = 0),$$

without using it. Observe that we have

$$\begin{aligned}
\mathbb{P}(T_0 > 2) &= 1 - \mathbb{P}(T_0 \leq 2) \\
&= 1 - \mathbb{P}(T_0 = 2) \qquad (1) \\
&= 1 - \mathbb{P}(S_2 = 0) \\
&= 1 - \frac{1}{2} = \frac{1}{2} = \mathbb{P}(S_2 = 0)
\end{aligned}$$

where at (1) we used the fact that $T_0 > 0$ holds by definition and T_0 is always even.

Example 2.11. Take $N \geq 5$ and $n = 2$. We again want to check that

$$\mathbb{P}(T_0 > 4) = \mathbb{P}(S_4 = 0)$$

holds. Compute

$$\mathbb{P}(S_4 = 0) = 6 \cdot \left(\frac{1}{4}\right)^4 = \frac{3}{8}$$

and observe that

$$\begin{aligned}\mathbb{P}(T_0 > 4) &= 1 - \mathbb{P}(T_0 \leq 4) \\ &= 1 - \mathbb{P}(T_0 = 2) - \mathbb{P}(T_0 = 4) \\ &= 1 - \frac{1}{2} - \frac{1}{8} = \frac{3}{8} \\ &= \mathbb{P}(S_4 = 0).\end{aligned}$$

Remark 2.12. Recall that by Remark 2.5 we have

$$\mathbb{P}(S_{2n} = 0) \underset{n\to\infty}{\sim} \frac{1}{\sqrt{\pi n}}$$

and thus by Theorem 2.9 we get

$$\lim_{n\to\infty} \mathbb{P}(T_0 > 2n) = \lim_{n\to\infty} \mathbb{P}(S_{2n} = 0) = 0.$$

Hence we have

$$\lim_{n\to\infty} \mathbb{P}(T_0 \leq 2n) = 1,$$

which means that the random walk is *recurrent*.

2.3 The arcsin Law

Let $N \in \mathbb{N}$ and consider a random walk $(S_n)_{0\leq n\leq 2N}$ starting at 0. Set

$$L(\omega) := \max\{0 \leq n \leq 2N \mid S_n(\omega) = 0\}$$

to be the *last visit at 0* of the random walk.

Theorem 2.13. *For $n \leq N$ we have*

$$\begin{aligned}\mathbb{P}(L = 2n) &= \mathbb{P}(S_{2n} = 0)\mathbb{P}(S_{2(N-n)} = 0) \\ &= \frac{1}{2^{2N}}\binom{2n}{n}\binom{2(N-n)}{N-n}\end{aligned}$$

Proof. Define

$$\begin{aligned}A &:= \{L = 2n\} = \{\omega \in \Omega \mid L(\omega) = 2n\}\\ &= \{\omega \in \Omega \mid S_{2n}(\omega) = 0 \text{ and } S_k(\omega) \neq 0 \text{ for } k > 2n\}.\end{aligned}$$

Now observe that we have

$$\begin{aligned}A &= \Big\{(x_1, \dots, x_{2N}) \in \{-1,1\}^{2N} \mid \sum_{i=1}^{2n} x_i = 0,\ \forall k > 2n : \sum_{i=2n+1}^{k} x_i \neq 0\Big\}\\ &= \Big\{(x_1, \dots, x_{2n}, y_1, \dots, y_{2N-2n}) \in \{1,-1\}^{2N} \mid \sum_{i=1}^{2n} x_i = 0,\ \forall 1 \leq j \leq 2(N-n) : \sum_{i=1}^{j} y_i \neq 0\Big\}.\end{aligned}$$

Now set

$$\begin{aligned}B_1 &:= \Big\{(x_1, \dots, x_{2n}) \in \{-1,1\}^{2n} \mid \sum_{i=1}^{2n} x_i = 0\Big\},\\ B_2 &:= \Big\{(y_1, \dots, y_{2(N-n)}) \in \{-1,1\}^{2(N-n)} \mid \forall 1 \leq j \leq 2(N-n) : \sum_{i=1}^{j} y_i \neq 0\Big\}\end{aligned}$$

and observe that then we have

$$A = B_1 \times B_2.$$

This implies

$$|A| = |B_1| \cdot |B_2|.$$

Note that we have

$$\{S_{2n} = 0\} = \Big\{(x_1, \dots, x_{2N}) \in \{-1,1\}^{2N} \mid \sum_{i=1}^{2n} x_i = 0\Big\}$$

and thus

$$|\{S_{2n} = 0\}| = |B_1| \cdot 2^{2(N-n)}$$

which implies

$$\mathbb{P}(S_{2n} = 0) = \frac{|\{S_{2n} = 0\}|}{|\Omega|} = \frac{|B_1| \cdot 2^{2(N-n)}}{2^{2N}} = \frac{|B_1|}{2^{2n}}.$$

Furthermore, we have

$$\begin{aligned}\{T_0 > 2(N-n)\} &= \{\omega \in \Omega \mid S_1(\omega) \neq 0, \dots, S_{2(N-n)} \neq 0\}\\ &= \Big\{(x_1, \dots, x_{2N}) \in \{-1,1\}^{2N} \mid \forall 1 \leq j \leq 2(N-n) : \sum_{i=1}^{j} x_i \neq 0\Big\}\end{aligned}$$

which implies

$$\begin{aligned}\mathbb{P}(T_0 > 2(N-n)) &= \frac{|\{T_0 > 2(N-n)\}|}{2^{2N}} \\ &= \frac{|B_2| \cdot 2^{2n}}{2^{2N}} = \frac{|B_2|}{2^{2(N-n)}}.\end{aligned}$$

Now by applying Theorem 2.9 we get

$$\mathbb{P}(T_0 > 2(N-n)) = \mathbb{P}(S_{2(N-n)} = 0)$$

and thus, putting everything together, we have

$$\begin{aligned}\mathbb{P}(A) = \frac{|A|}{2^{2N}} &= \frac{|B_1| \cdot |B_2|}{2^{2n} \cdot 2^{2(N-n)}} \\ &= \mathbb{P}(S_{2n} = 0)\mathbb{P}(S_{2(N-n)} = 0) \\ &= \frac{1}{2^{2N}} \binom{2n}{n} \binom{2(N-n)}{N-n}\end{aligned}$$

which concludes the proof. □

Now Theorem 2.13 implies the following interesting result.

The arcsin law 2.14. *We have* $\mathbb{P}(S_{2n} = 0) \approx \frac{1}{\sqrt{\pi n}}$ *and thus*

$$\mathbb{P}(L = 2n) \approx \frac{1}{\pi\sqrt{n(N-n)}} = \frac{1}{N} f\left(\frac{n}{N}\right)$$

for $f(x) := \frac{1}{\pi\sqrt{x(1-x)}}$. *Hence we have*

$$\mathbb{P}\left(\frac{L}{2N} \le z\right) \approx \sum_{n:\frac{n}{N} \le z} \frac{1}{N} f\left(\frac{n}{N}\right) \approx \int_0^z f(x)\, dx = \frac{2}{\pi} \arcsin\sqrt{z}.$$

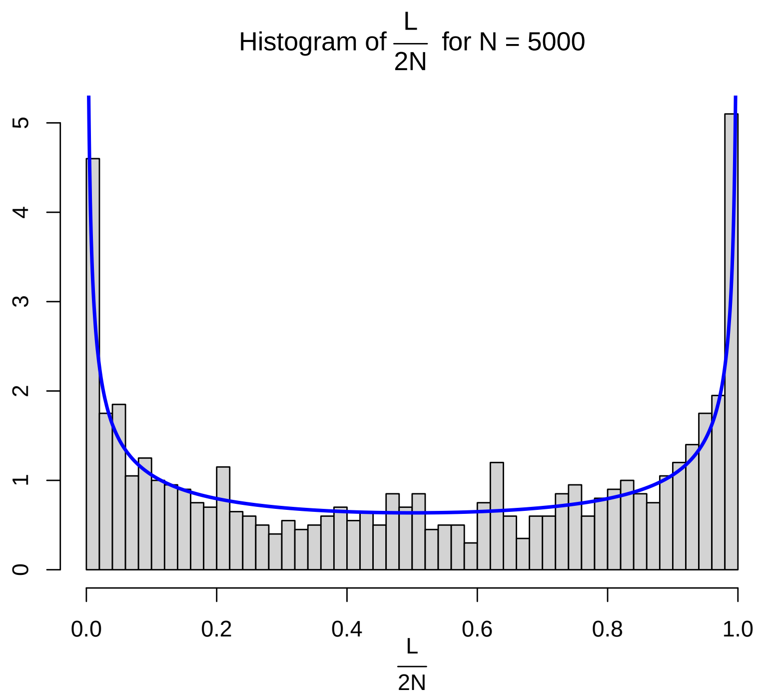


Figure 7: Histogram of $\frac{L}{2N}$ for $N = 5000$ and the function $\frac{2}{\pi}\arcsin\sqrt{z}$ in blue.

3 General Models

3.1 Introduction

Definition 3.1. A triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is said to be a *probability space* if the following hold.

(a) $\mathcal{A}$ is a *σ-algebra*, that is

- $\Omega \in \mathcal{A}$,
- $A \in \mathcal{A} \implies A^c \in \mathcal{A}$,
- $(A_i)_i \subseteq \mathcal{A} \implies \bigcup_i A_i \in \mathcal{A}$.

(b) $\mathbb{P}$ is a *probability measure*, that is

- $\mathbb{P}(\Omega) = 1$,
- $\mathbb{P}$ is *σ-additive*, so if $(A_i)_i \subseteq \mathcal{A}$ are pairwise disjoint then $\mathbb{P}(\bigsqcup_i A_i) = \sum_i \mathbb{P}(A_i)$ holds.

Examples 3.2.

- Let Ω be countable, $\mathcal{A} = 2^\Omega$ and let $\{p(\omega) \in [0,1] \mid \omega \in \Omega\}$ be given weights with $\sum_{\omega \in \Omega} p(\omega) = 1$. Then
$$\mathbb{P} : \mathcal{A} \to [0,1], A \mapsto \sum_{\omega \in A} p(\omega)$$
is a probability measure.

- Let $\Omega = [0,1]$ and let $\mathcal{A} = \mathcal{B}([0,1])$ be the *Borel σ-algebra*, i.e. the smallest σ-algebra containing all closed intervals $[a,b] \subseteq [0,1]$. Then it can be shown that there exists a unique probability measure $\mathbb{P}$ on $\mathcal{A}$ such that
$$\mathbb{P}([a,b]) = b - a$$
holds for all such intervals. This measure $\mathbb{P}$ is also called the *uniform distribution* on $[a,b]$.

Remarks 3.3.

- For $(A_i)_i \subseteq \mathcal{A}$ we have $\bigcap_i A_i \in \mathcal{A}$.
- For $(A_i)_i \subseteq \mathcal{A}$ we define
$$A_\infty := \text{“infinitely many } A_i \text{ occur”}.$$
Then
$$A_\infty = \{\forall n \in \mathbb{N}\ \exists k \geq n : A_k \text{ occurs}\} = \bigcap_{n \geq 1} \bigcup_{k \geq n} A_k.$$

- Instead of σ-additivity, we may also define a weaker version called *additivity* by the following property: If $A_1, \dots, A_n \in \mathcal{A}$ are pairwise disjoint, then

$$\mathbb{P}\Big(\bigsqcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

Theorem 3.4. *Let $\mathbb{P} : \mathcal{A} \to [0,1]$ be an additive measure. Then the following are equivalent.*

(a) $\mathbb{P}$ *is σ-additive.*

(b) If $A_1 \subseteq A_2 \subseteq \dots$ are all in $\mathcal{A}$ then

$$\mathbb{P}\Big(\bigcup_{n} A_n\Big) = \lim_{n\to\infty} \mathbb{P}(A_n),$$

which is called continuity from below.

(c) If $A_1 \supseteq A_2 \supseteq \dots$ are all in $\mathcal{A}$ then

$$\mathbb{P}\Big(\bigcap_{n} A_n\Big) = \lim_{n\to\infty} \mathbb{P}(A_n),$$

which is called continuity from above.

Proof. (a) $\Longrightarrow$ (b). Take $A_1 \subseteq A_2 \subseteq \dots$ all in $\mathcal{A}$ and define

$$B_1 := A_1, \quad B_n := A_n \setminus A_{n-1}$$

for $n \geq 2$. Then $(B_n)_{n\geq 1}$ are pairwise disjoint with

$$\bigcup_{k} A_k = \bigsqcup_{k} B_k$$

and

$$A_n = \bigsqcup_{k=1}^{n} B_k.$$

Now using σ-additivity, we get

$$\begin{aligned}
\mathbb{P}\Big(\bigcup_{k} A_k\Big) &= \mathbb{P}\Big(\bigsqcup_{k} B_k\Big) = \sum_{k=1}^{\infty} \mathbb{P}(B_k) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(B_k) \overset{(1)}{=} \lim_{n\to\infty} \mathbb{P}\Big(\bigsqcup_{k=1}^{n} B_n\Big) \\
&= \lim_{n\to\infty} \mathbb{P}(A_n),
\end{aligned}$$

where at (1) we used additivity of $\mathbb{P}$.

(b) $\implies$ (a). Let $(A_k)_k \subseteq \mathcal{A}$ be pairwise disjoint. Define

$$B_1 := A_1, \quad B_n := A_n \cup B_{n-1}.$$

Then we have $B_1 \subseteq B_2 \subseteq \dots$ and all are in $\mathcal{A}$. Hence by (b) we get

$$\begin{aligned}\mathbb{P}\Big(\bigsqcup_k A_k\Big) = \mathbb{P}\Big(\bigcup_k B_k\Big) &= \lim_{n\to\infty} \mathbb{P}(B_n)\\ &\overset{(1)}{=} \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(A_k) = \sum_{k=1}^{\infty} \mathbb{P}(A_k),\end{aligned}$$

where at (1) we again used additivity of $\mathbb{P}$.

(b) $\iff$ (c). Follows by using the property $A \in \mathcal{A} \implies A^c \in \mathcal{A}$. $\square$

Corollary 3.5. *For any* $(A_k)_k \subseteq \mathcal{A}$ *we have*

$$\mathbb{P}\Big(\bigcup_k A_k\Big) \leq \sum_k \mathbb{P}(A_k)$$

if $\mathbb{P}$ *is* σ*-additive.*

Lemma 3.6 (*Borel-Cantelli lemma*)**.** *Let $(A_k)_k \subseteq \mathcal{A}$ and $A_\infty = \bigcap_{n\geq 1}\bigcup_{k\geq n} A_k$.*

(a) If $\sum_k \mathbb{P}(A_k) < \infty$ then $\mathbb{P}(A_\infty) = 0$.

(b) Under the additional condition that $(A_k)_k$ are (mutually) independent, it holds that

$$\sum_k \mathbb{P}(A_k) = \infty \implies \mathbb{P}(A_\infty) = 1.$$

Proof. (a) Consider the sets $B_n := \bigcup_{k\geq n} A_k$. Then $(B_n)_n$ is monotone decreasing and thus we have

$$\mathbb{P}\Big(\bigcap_n B_n\Big) = \lim_{n\to\infty} \mathbb{P}(B_n)$$

which implies

$$\mathbb{P}(A_\infty) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{k\geq n} A_k\Big) \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(A_k) = 0$$

where the last equality follows since $\sum_k \mathbb{P}(A_k) < \infty$ is assumed.

(b) Fix $n \geq 1$ and consider $B_m := \bigcap_{k=n}^m A_k^c$ for all $m \geq n$. Then $(B_m)_m$ is monotone decreasing and hence we have

$$\mathbb{P}\Big(\bigcap_{m\geq n} B_m\Big) = \lim_{m\to\infty} \mathbb{P}(B_m).$$

Furthermore

$$\bigcap_{m\geq n} B_m = \bigcap_{k\geq n} A_k^c$$

and thus

$$\mathbb{P}\Big(\bigcap_{k\geq n} A_k^c\Big) = \lim_{m\to\infty} \mathbb{P}\Big(\bigcap_{k=n}^m A_k^c\Big).$$

Now since $(A_k)_k$ are assumed to be independent, we get

$$\mathbb{P}\Big(\bigcap_{k\geq n} A_k^c\Big) = \lim_{m\to\infty} \prod_{k=n}^m \mathbb{P}(A_k^c) = \lim_{m\to\infty} \prod_{k=n}^m (1 - \mathbb{P}(A_k)).$$

Now note that $1 - x \leq e^{-x}$ holds for all $x \geq 0$. Hence

$$\prod_{k=n}^m (1 - \mathbb{P}(A_k)) \leq \prod_{k=n}^m e^{-\mathbb{P}(A_k)} = \exp\Big(-\sum_{k=n}^m \mathbb{P}(A_k)\Big)$$

and recall that by assumption $\sum_{k\geq 1} \mathbb{P}(A_k) = \infty$. Since n is fixed, this implies $\sum_{k\geq n} \mathbb{P}(A_k) = \infty$ and thus we can conclude that

$$\mathbb{P}\Big(\bigcap_{k\geq n} A_k^c\Big) \leq \lim_{m\to\infty} \exp\Big(-\sum_{k=n}^m \mathbb{P}(A_k)\Big) = 0$$

holds for all $n \in \mathbb{N}$ fixed. Now note that the sequence $(\bigcap_{k\geq n} A_k^c)_n$ is monotone increasing and thus

$$\mathbb{P}(A_\infty^c) = \mathbb{P}\Big(\bigcup_{n\geq 1}\bigcap_{k\geq n} A_k^c\Big) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{k\geq n} A_k^c\Big) = 0,$$

so

$$\mathbb{P}(A_\infty) = 1$$

and we can conclude.

□

Example 3.7. Let $X_1, X_2, \dots$ be independent outcomes of throwing $p_1, p_2, \dots$ coins with $p_i \in (0,1)$. Applying the Borel-Cantelli lemma to $A_k := \{X_k = 1\}$ implies the following results.

- If $\sum_{k\geq 1} p_i < \infty$ then $\mathbb{P}(X_k = 1 \text{ infinitely often}) = 0$.
- If $\sum_{k\geq 1} p_i = \infty$ then $\mathbb{P}(X_k = 1 \text{ infinitely often}) = 1$.

3.2 Transformations of Probability Spaces

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a measure space, $\tilde{\Omega} \neq \emptyset$ and $\tilde{\mathcal{A}}$ a σ-algebra on $\tilde{\Omega}$.

Definition 3.8. An application $\Phi : (\Omega, \mathcal{A}) \to (\tilde{\Omega}, \tilde{\mathcal{A}})$ is said to be *measurable* if for all $B \in \tilde{\mathcal{A}}$ we have $\Phi^{-1}(B) \in \mathcal{A}$.

Remarks 3.9.

- We can generate a σ-algebra with a given collection of subsets of Ω as follows: Given a collection $\mathcal{C} \subseteq 2^\Omega$, the *$\sigma$-algebra generated by* $\mathcal{C}$ is defined by

$$\sigma(\mathcal{C}) := \bigcap_{\substack{\mathcal{D}\supseteq\mathcal{C} \\ \mathcal{D}\ \sigma\text{-algebra}}} \mathcal{D}.$$

- If $\tilde{\mathcal{A}} = \sigma(\tilde{\mathcal{C}})$ with $\tilde{\mathcal{C}} \subseteq 2^{\tilde{\Omega}}$ then the application $\Phi : \Omega \to \tilde{\Omega}$ is measurable if and only if

$$\forall \tilde{C} \in \tilde{\mathcal{C}} : \Phi^{-1}(\tilde{C}) \in \mathcal{A}.$$

Theorem 3.10. *Let $\Phi : \Omega \to \tilde{\Omega}$ be a measurable application and define*

$$\tilde{\mathbb{P}} : \tilde{\mathcal{A}} \to [0,1], A \mapsto \mathbb{P}(\Phi^{-1}(A)).$$

Then $\tilde{\mathbb{P}}$ is a probability measure on $\tilde{\mathcal{A}}$ and $\tilde{\mathbb{P}}$ is called the image *of $\mathbb{P}$ under Φ or the* distribution of Φ *under $\mathbb{P}$ or the* induced probability measure *by Φ.*

Proof. We need to check that $\tilde{\mathbb{P}}$ is σ-additive and that $\tilde{\mathbb{P}}(\tilde{\Omega}) = 1$ holds. By definition, we have

$$\tilde{\mathbb{P}}(\tilde{\Omega}) = \mathbb{P}(\Phi^{-1}(\tilde{\Omega})) = \mathbb{P}(\Omega) = 1.$$

Now let $(A_k)_k \subseteq \tilde{\mathcal{A}}$ be pairwise disjoint. Then we have

$$\begin{aligned}\tilde{\mathbb{P}}\Big(\bigsqcup_k A_k\Big) &= \mathbb{P}\Big(\Phi^{-1}\Big(\bigsqcup_k A_k\Big)\Big) = \mathbb{P}\Big(\bigsqcup_k \Phi^{-1}(A_k)\Big) \\ &= \sum_k \mathbb{P}(\Phi^{-1}(A_k)) = \sum_k \tilde{\mathbb{P}}(A_k)\end{aligned}$$

which concludes the proof. $\square$

3.3 Real Random Variables

Let $\mathcal{C} := \{(-\infty, b] \mid b \in \mathbb{R}\}$, denote by $\mathcal{B} := \sigma(\mathcal{C})$ the the *Borel σ-algebra* and set $\tilde{\Omega} := \mathbb{R}$, $\tilde{\mathcal{A}} := \mathcal{B}$.

Definition 3.11. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. A *random variable* is a measurable application $X : \Omega \to \tilde{\Omega} = \mathbb{R}$. The *distribution* of X, denoted by μ_X, is equal to the image of $\mathbb{P}$ under X, so

$$\forall B \in \mathcal{B} : \quad \mu_X(B) = \mathbb{P}(X^{-1}(B)) =: \mathbb{P}(X \in B).$$

Examples 3.12.

- Consider $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ such that

  $$\mu_X(B) = \mathbb{P}(X \in B) = \int_B \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx$$

  holds for all $B \in \mathcal{B}$. Then X is said to be a *Gaussian* or *Normal* random variable. In this case, we write $X \sim \mathcal{N}(0,1)$ and we have

  $$\mathbb{E}[X] = 0, \quad \mathrm{Var}(X) = 1.$$

  In particular,

  $$\mu_X((-\infty, a]) = \mathbb{P}(X \le a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx.$$

- Consider $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ to be a random variable such that

  $$\mu_X(B) = \sum_{j \in \{0,1,\ldots,n\} \cap B} \binom{n}{j} p^j (1-p)^{n-j}$$

  for $p \in (0,1)$ and $n \in \mathbb{N}$. Here we recognize a *Binomial* random variable with success probability p and number of trials n. Note that in this case, we can replace $(\mathbb{R}, \mathcal{B})$ by $(\tilde{\Omega}, 2^{\tilde{\Omega}})$ for $\tilde{\Omega} = \{1, \ldots, n\}$.

3.4 Distribution Functions

Definition 3.13. Let $X : \Omega \to \mathbb{R}$ be a real random variable. The application

$$F_X : \mathbb{R} \to [0,1], b \mapsto \mu_X((-\infty, b]) = \mathbb{P}(X \leq b)$$

is called the *(cumulative) distribution function* of X or *CDF* in short.

Proposition 3.14. *Let μ be the distribution of X and F the CDF of X.*

- *For $a \leq b$ we have $\mu((a,b]) = \mathbb{P}(X \in (a,b]) = F(b) - F(a)$.*
- *For $a \in \mathbb{R}$ we have $\mu(\{a\}) = F(a) - F(a_-)$, where $F(a_-) := \lim_{x \uparrow a} F(a)$.*

Theorem 3.15. *For any distribution function F_X, we have the following properties:*

- monotonicity: $\forall x \leq y : F_X(x) \leq F_X(y)$,
- right-continuity: $F_X(x) = \lim_{h \downarrow 0} F_X(x+h)$,
- $\lim_{x \to -\infty} F_X(x) = 0$ *and* $\lim_{x \to \infty} F_X(x) = 1$.

Moreover, any function F with the properties above is the distribution function of some random variable.

Lemma 3.16. *Let F be a distribution function and define*

$$F^{-1}(t) := \inf\{x \in \mathbb{R} \mid F(x) \geq t\}$$

for all $t \in (0,1)$. Then

- *F^{-1} is monotone increasing,*
- *F^{-1} is left-continuous,*
- *$\forall x \in \mathbb{R} : F^{-1}(F(x)) \leq x$,*
- *$\forall t \in (0,1) : t \leq F(F^{-1}(t))$.*

Definition 3.17. For $t \in (0,1)$, the value $F^{-1}(t)$ is called the *t-quantile* of F. If $t = \frac{1}{2}$ then $F^{-1}(t)$ is called the *median* of F.

Definition 3.18. Let $\mathcal{A}$ be a σ-algebra of subsets of Ω.

- μ is called a *measure* if $\mu : \mathcal{A} \to [0, \infty]$ such that $\mu(\emptyset) = 0$ holds and μ is σ-additive. Then $(\Omega, \mathcal{A}, \mu)$ is called a *measure space*.
- A measure μ is called *finite* if $\mu(\Omega) < \infty$.

- A measure μ is called *σ-finite* if there exists a partition $\{F_k\}_{k\geq 1}$ of Ω such that $\mu(F_k) < \infty$ holds for all $k \geq 1$.

- Let μ_1 and μ_2 be two measures on $\mathcal{A}$. Then μ_2 is said to be *dominated* by μ_1 if

$$\mu_1(A) = 0 \implies \mu_2(A) = 0$$

holds for all $A \in \mathcal{A}$. In this case, we also say that μ_2 is *absolutely continuous* with respect to μ_1 and write $\mu_2 \ll \mu_1$.

Theorem 3.19 (*Radon-Nikodym*)**.** *Let $(\tilde{\Omega}, \tilde{\mathcal{A}}, \mu)$ be a measure space such that μ is σ-finite. Let ν be another measure on $\tilde{\mathcal{A}}$ such that ν is absolutely continuous with respect to μ, so $\nu \ll \mu$. Then there exists a measurable function f such that*

- $f \geq 0$,

- $\forall A \in \mathcal{A} : \nu(A) = \int_A f \, d\mu$.

The function f is called the *Radon-Nikodym derivative / density* of ν with respect to μ and we write $f = \frac{d\nu}{d\mu}$. If ν is a probability measure, i.e. $\nu(\tilde{\Omega}) = 1$, then

$$\int_{\tilde{\Omega}} f \, d\mu = 1$$

holds.

Examples 3.20.

- Consider $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathcal{X}, 2^{\mathcal{X}}, \mu)$ with $\mathcal{X} := X(\Omega)$ assumed to be countable and let μ be the counting measure, so $\mu(B) = |B|$. Write $\mathcal{X} = \{x_i\}_{i\in\mathbb{N}}$. Then μ is σ-finite and $\mu(B) = 0$ if and only if $B = \emptyset$. Hence we have

$$\forall B \in 2^{\mathcal{X}} : \mu(B) = 0 \implies \mu_X(B) = 0,$$

so $\mu_X \ll \mu$. Then by the Radon-Nikodym theorem there exists a measurable function $f_X \geq 0$ such that

$$\mathbb{P}(X \in B) = \mu_X(B) = \int_B f_X \, d\mu = \sum_{y\in B} f_X(y)$$

holds for all $B \in \tilde{\mathcal{A}} := 2^{\mathcal{X}}$. If $B = \{x\}$ then

$$\mathbb{P}(X = x) = f_X(x)$$

for any $x \in \mathcal{X}$. Hence we find that the Radon-Nikodym density of the distribution of X (in this discrete case) is given by the *probability mass function* $\mathbb{P}(X = x)$.

- Now consider $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}, \lambda)$ such that $\mu_X \ll \lambda$ where λ is the Lebesgue measure, so λ is σ-finite. Hence by the Radon-Nikodym theorem there exists a measurable function $f_X \geq 0$ such that

$$\mathbb{P}(X \in B) = \mu_X(B) = \int_B f_X \, d\lambda = \int_B f_X(x) \, dx$$
$$\overset{(1)}{\iff} \forall b \in \mathbb{R} : \mathbb{P}(X \leq b) = F_X(b) = \int_{-\infty}^{b} f_X(x) \, dx$$

for all $B \in \mathcal{B}$ where F_X is the distribution function of X. Note that (1) is a non-trivial fact that has to be shown. In this case, f_X also called the *probability density function* or *pdf* in short.

3.5 Standard Types of Distributions

3.5.1 Discrete Distributions

Let X be a random variable such that $X(\Omega) = \{x_k\}_{k \in \mathbb{N}}$ is countable with $f_X(x) = \mathbb{P}(X = x)$. Then

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{k \colon x_k \leq x} f_X(x_k)$$

holds for all $x \in \mathbb{R}$:

Examples 3.21.

- X has a *Dirac* distribution at a, written $X \sim \text{Dirac}(a)$, if

$$X : \Omega \to \mathbb{R}, \omega \mapsto a$$

holds with $\mathbb{P}(X = a) = 1$.

- X has a *Bernoulli* distribution with success probability $p \in (0, 1)$, written $X \sim \text{Bernoulli}(p)$, if we have $X(\Omega) = \{0, 1\}$ with

$$\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = p.$$

- X has a *Binomial* distribution with success probability p and number of trials $n \in \mathbb{N}$, written $X \sim \text{Binomial}(n, p)$, if we have $X(\Omega) = \{1, \ldots, n\}$ with

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

for all $x \in X(\Omega)$.

- X has a *Poisson* distribution with rate $\lambda \in (0, \infty)$, written $X \sim \text{Pois}(\lambda)$, if we have $X(\Omega) = \mathbb{N}_0$ with

$$\mathbb{P}(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

for all $x \in X(\Omega)$.

- X has a *Geometric* distribution with success probability $p \in (0,1)$, written $X \sim \text{Geo}(p)$, if we have $X(\Omega) = \mathbb{N}$ with
$$\mathbb{P}(X = x) = p(1-p)^x$$
for all $x \in X(\Omega)$.

3.5.2 Absolutely Continuous Distributions

A real random variable $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ is said to have an absolutely continuous distribution if there exists a measurable function $f_X \geq 0$ such that
$$\mu_X(B) = \int_B f_X(x)\, dx$$
for all $B \in \mathcal{B}$ and $\int_{\mathbb{R}} f_X(x)\, dx = 1$ holds.

Remarks 3.22.

(1) The CDF F_X is always continuous if X is absolutely continuous.

(2) If f_X is continuous at some $x_0 \in \mathbb{R}$ then F_X is differentiable at x_0 with $F'_X(x_0) = f_X(x_0)$.

(3) A density f_X is defined up to a set of measure 0.

(4) $F'_X = f_X$ holds almost everywhere.

Proof. *(of 1)* Let $x_0 \in \mathbb{R}$ and $h > 0$. Then
$$\begin{aligned} F_X(x_0+h) - F_X(x_0) &= \mathbb{P}(X \in (x_0, x_0+h]) \\ &= \int_{x_0}^{x_0+h} f_X(t)\, dt = \int_{\mathbb{R}} \mathbb{1}_{(x_0,x_0+h]}(t) f_X(t)\, dt. \end{aligned}$$
Now note that $0 \leq \mathbb{1}_{(x_0,x_0+h]} f_X \leq f_X$ holds and f_X is integrable, so by the dominated convergence theorem we have
$$\begin{aligned} \lim_{h\to 0}(F_X(x_0+h) - F_X(x_0)) &= \lim_{h\to 0} \int_{\mathbb{R}} \mathbb{1}_{(x_0,x_0+h]}(t) f_X(t)\, dt \\ &= \int_{\mathbb{R}} \lim_{h\to 0} \mathbb{1}_{(x_0,x_0+h]}(t) f_X(t)\, dt = 0. \end{aligned}$$
Hence F_X is continuous. $\square$

Examples 3.23.

- X has a *Uniform* distribution on $[a,b]$ for $a < b$, written $X \sim \mathcal{U}([a,b])$, if
$$f_X(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x)$$

holds with CDF

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

- X has an *Exponential* distribution with *intensity/rate* $\lambda > 0$, written $X \sim \mathrm{Exp}(\lambda)$, if

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

holds with CDF

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0. \end{cases}$$

- X has a *Gamma* distribution with *shape parameter* $\alpha > 0$ and *rate* $\lambda > 0$, written $X \sim \Gamma(\alpha, \lambda)$, if

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \mathbb{1}_{[0,\infty)}(x),$$

where

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

Note that in this case there is no closed form for the CDF.

- X has a *Normal/Gaussian* distribution with parameters $\mu \in \mathbb{R}$, $\sigma^2 > 0$, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if

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

for $x \in \mathbb{R}$. Again there is no closed form for the CDF.

3.5.3 Transformations of Random Variables

Let $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ be a random variable and $g : \mathbb{R} \to \mathbb{R}$ a measurable function. Then

$$Y = g(X) = g \circ X$$

is again a random variable with distribution

$$\mu_Y(B) = \mu_X(g^{-1}(B)).$$

Example 3.24. Let $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$. Then

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

is symmetric around 0. Hence

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{y}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}}\mathbb{1}_{[0,\infty)}(y) \\ &= \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\sqrt{\pi}}y^{\frac{1}{2}-1}e^{-\frac{y}{2}}\mathbb{1}_{[0,\infty)}(y) \\ &= \frac{\lambda^\alpha}{\sqrt{\pi}}y^{\alpha-1}e^{-\lambda y}\mathbb{1}_{[0,\infty)}(y) \end{aligned}$$

for $\alpha = \lambda = \frac{1}{2}$. Then $\sqrt{\pi} = \Gamma\left(\frac{1}{2}\right)$ and thus $Y \sim \Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$. The distribution of Y is better known under the name of *Chi-square* distribution with one degree of freedom and in this case we write $Y \sim \chi^2_{(1)}$.

Consider again a real random variable and let

$$g : U \to V$$

be bijective in C^1 and non-zero on U, where $U \subseteq \mathbb{R}$ is open. Now if $\mathbb{P}(X \in U) = 1$ then $Y = g \circ X$ is a random variable which has an absolutely continuous distribution with density

$$f_Y = \frac{1}{|g' \circ g^{-1}|} f_X \circ g^{-1}.$$

Examples 3.25.

- Let

  $$g : \mathbb{R} \to \mathbb{R}, x \mapsto ax + b$$

  for $a \neq 0, b \in \mathbb{R}$. Then $g \in C^1(\mathbb{R})$ and $g'(x) = a \neq 0$ and g is bijective with

  $$g^{-1}(y) = \frac{y-b}{a}.$$

  If X admits a density f_X then $Y = g \circ X$ admits the density

  $$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

  for $y \in \mathbb{R}$.

- Let $X \sim \mathcal{U}(0,1)$ with $f_X = \mathbb{1}_{[0,1]}$ and

  $$g : (0,\infty) \to \mathbb{R}, x \mapsto -\log x.$$

  Then again $g \in C^1((0,\infty))$, g is bijective and $g'(x) \neq 0$ for all $x \in (0,\infty)$. Hence $Y = g \circ X$ admits the density

  $$f_Y(y) = e^{-y}\mathbb{1}_{(0,\infty)}(y)$$

  and thus $Y \sim \text{Exp}(1)$.

3.6 Expectation (revisited)

Definition 3.26. Let $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$ be a real random variable such that $X \geq 0$. Then we define

$$\mathbb{E}[X] := \int_\Omega X(\omega)\, d\mathbb{P}(\omega) \overset{(1)}{=} \int_\mathbb{R} x\, d\mu_X(x)$$

where μ_X is the induced probability measure by X given by $\mu_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(X \in B)$. Note that (1) needs to be proven (was done in Analysis III).

Definition 3.27. For an arbitrary random variable X, we define

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

if $\mathbb{E}[X_-]$ and $\mathbb{E}[X_+]$ are not both infinite. In this case we still have

$$\mathbb{E}[X] = \int_\Omega X(\omega)\, d\mathbb{P}(\omega) = \int_\mathbb{R} x\, d\mu_X(x).$$

Recall 3.28. We defined $X_+ = \max(X, 0)$ and $X_- = \max(-X, 0)$.

Note 3.29. In the discrete case, we have

$$\int_\mathbb{R} x\, d\mu_X(x) = \sum_{i \in I} x_i \mathbb{P}(X = x_i)$$

where $\{x_i\}_i = X(\Omega)$. In the absolute continuous case, we have

$$\int_\mathbb{R} x\, d\mu_X(x) = \int_\mathbb{R} x f_X(x)\, dx$$

where f_X is the density of the distribution of X with respect to the Lebesque measure.

For $g : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ a measurable function such that $Y = g(X) = g \circ X$ is integrable, that is

$$\mathbb{E}(|Y|) < \infty \iff \mathbb{E}(Y) < \infty,$$

we have

$$\begin{aligned}
\mathbb{E}[Y] &= \int_\Omega g(X(\omega))\, d\mathbb{P}(\omega) \\
&= \int_\mathbb{R} g(x)\, d\mu_X(x) \\
&= \begin{cases} \sum_i g(x_i)\mathbb{P}(X = x_i) & \text{in the discrete case} \\ \int_\mathbb{R} g(x) f_X(x)\, dx & \text{in the absolute continuous case.} \end{cases}
\end{aligned}$$

Definition 3.30 (Moments of random variable)**.** Let X be a random variable. Then we define

- for $k \in \mathbb{N}$ the *k-th moment* of X by $\mathbb{E}[X^k]$.
- for $k \in (0, \infty)$ the *k-th absolute moment* of X by $\mathbb{E}[|X|^k]$.
- for $k \in \mathbb{N}$ the *k-th centered moment* of X by $\mathbb{E}[(X - \mathbb{E}[X])^k]$.
- for $k \in (0, \infty)$ the *k-th absolute centered moment* of X by $\mathbb{E}[|X - \mathbb{E}[X]|^k]$.

For $k = 2$ we call

$$\mathrm{Var}(X) := \mathbb{V}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}(|X - \mathbb{E}[X]|^2)$$

the *variance* of X.

Definition 3.31 (Standard deviation)**.** For a random variable X we define the *standard deviation* of X by $\sigma(X) := \sqrt{\mathbb{V}(X)}$

Proposition 3.32 (Properties of the variance)**.** *Let X be a random variable. Then*

(1) $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

(2) $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$.

(3) $\mathbb{V}(aX + b) = a^2\mathbb{V}(X)$ *($\mathbb{V}$ is translation invariant!).*

(4) $\sigma(aX + b) = |a|\sigma(X)$.

(5) Let X_1 and X_2 be two independent random variables such that $\mathbb{V}(X_1), \mathbb{V}(X_2) < \infty$. Then

$$\mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2)$$

holds.

Proof idea.

(1) Use linearity of $\mathbb{E}$ and that the expectation of a constant is the constant itself.

(2) Follows from (a) since $\mathbb{V}(X) \geq 0$ always holds.

(3) Use definition of $\mathbb{V}$ and linearity of $\mathbb{E}$.

(4) Follows from the definition of standard diviation using (c).

(5) Compute

$$\begin{aligned}
&\mathbb{V}(X_1 + X_2) = \mathbb{E}[(X_1 + X_2)^2] - \mathbb{E}[X_1 + X_2]^2 \\
&= \mathbb{E}[X_1^2 + 2X_1X_2 + X_2^2] - (\mathbb{E}[X_1]^2 + 2\mathbb{E}[X_1]\mathbb{E}[X_2] + \mathbb{E}[X_2]^2) \\
&= \mathbb{E}[X_1^2] + \cancel{\mathbb{E}[2X_1X_2]} + \mathbb{E}[X_2^2] - \mathbb{E}[X_1]^2 - \cancel{2\mathbb{E}[X_1]\mathbb{E}[X_2]} - \mathbb{E}[X_2]^2 \qquad \text{(a)} \\
&= \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 + \mathbb{E}[X_2^2] - \mathbb{E}[X_2]^2 \\
&= \mathbb{V}(X_1) + \mathbb{V}(X_2).
\end{aligned}$$

where at (a) we used the independence of X_1 and X_2. ∷

Examples 3.33.

- Let $X \sim \text{Bernoulli}(p), p \in (0,1)$. Then we know that $\mathbb{E}[X] = p$. Hence

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

Alternatively, we see that

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \overset{(1)}{=} \mathbb{E}[X] - \mathbb{E}[X]^2 = p - p^2 = p(1-p),$$

where at (1) we used the fact that $X(\omega) \in \{0,1\}$.

- Let $X \sim \text{Bin}(n,p)$ with $n \in \mathbb{N}$ and $p \in (0,1)$. Then $\mathbb{E}[X] = np$ and thus

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - n^2p^2$$

and

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_i g(x_i)\mathbb{P}(X = x_i) = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} \\ &= \ldots \text{(gets complicated)}\end{aligned}$$

where $X(\Omega) = \{0,1,\ldots,n\}$ and $\mathbb{P}(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$. But note that in this case we also have $X = X_1 + \ldots + X_n$ with $X_1, \ldots, X_n$ are independent outcomes of tossing a p-coin n times, so $X_1, \ldots, X_n$ are i.i.d. (independent identically distributed) $\sim \text{Bernoulli}(p)$. Hence

$$\mathbb{V}(X) = \sum_{i=1}^{n} \mathbb{V}(X_i) = np\mathbb{V}(X_1) = np(1-p).$$

- Let $X \sim \text{Pois}(\lambda)$ with $\lambda \in (0,\infty)$, so $\mathbb{E}[X] = \lambda$. Then

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{k=1}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{ke^{-\lambda}\lambda^k}{(k-1)!} \\ &= \sum_{k=0}^{\infty} \frac{(k+1)e^{-\lambda}\lambda^{k+1}}{k!} = \sum_{k=0}^{\infty} \frac{ke^{-\lambda}\lambda^{k+1}}{k!} \\ &= \lambda \sum_{k=0}^{\infty} \frac{ke^{-\lambda}\lambda^k}{k!} + \lambda \underbrace{\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!}}_{1} \\ &= \lambda\mathbb{E}[X] + \lambda = \lambda^2 + \lambda.\end{aligned}$$

Hence we have $\mathbb{V}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$.

- Let $X \sim \mathcal{U}(0,1)$, so $f_X(x) = \mathbb{1}_{[0,1]}(x)$ and

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\, dx = \int_0^1 x\, dx = \frac{1}{2}.$$

Furthermore

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} g(x) f_X(x)\, dx = \int_0^1 x^2\, dx = \frac{1}{3}$$

and thus

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

In general consider $Y \sim \mathcal{U}(a,b)$. Then $X := \frac{Y-a}{b-a} \sim \mathcal{U}(0,1)$ and thus

$$\mathbb{E}\left[\frac{Y-a}{b-a}\right] = \frac{1}{2} \implies \mathbb{E}[Y] = a + \frac{1}{2}(b-a) = \frac{a+b}{2}$$

and

$$\begin{aligned}
\frac{1}{12} = \mathbb{V}(X) &= \mathbb{V}\left(\frac{Y-a}{b-a}\right) = \mathbb{V}\left(\frac{Y}{b-a} - \frac{a}{b-a}\right) \\
&= \mathbb{V}\left(\frac{Y}{b-a}\right) = \frac{1}{(b-a)^2}\mathbb{V}(Y) \\
&\implies \mathbb{V}(Y) = \frac{(b-a)^2}{12}.
\end{aligned}$$

- Let $X \sim \mathcal{N}(0,1)$, so $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\, dx = \int_{\mathbb{R}} \underbrace{x \tfrac{1}{2\sqrt{\pi}} e^{-x^2/2}}_{\text{odd}}\, dx = 0.$$

Furthermore,

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_{\mathbb{R}} x^2 f_X(x)\, dx = \int_{\mathbb{R}} x^2 \tfrac{1}{2\sqrt{\pi}} e^{-x^2/2}\, dx \\
&= \tfrac{1}{2\sqrt{\pi}} \int_{\mathbb{R}} x \cdot x e^{-x^2/2}\, dx = \tfrac{1}{2\sqrt{\pi}} \left(\underbrace{[-x e^{-x^2/2}]_{-\infty}^{\infty}}_{0} + \int_{\mathbb{R}} e^{-x^2/2}\, dx \right) \\
&= \tfrac{1}{2\sqrt{\pi}} \int_{\mathbb{R}} e^{-x^2/2}\, dx = 1
\end{aligned}$$

and thus

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1.$$

In general, for $X \sim \mathcal{N}(m, \sigma^2)$ we have

$$\begin{aligned}
\mathbb{E}[X] &= m \\
\mathbb{E}[X^2] &= \sigma^2 + m^2 \\
\mathbb{V}(X) &= \sigma^2.
\end{aligned}$$

3.7 Inequalities

Theorem 3.34 (Jensen's inequality). *Let X be an integrable random variable and $g : \mathbb{R} \to \mathbb{R}$ be a convex function such that $g(X) = g \circ X$ is integrable. Then*

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

holds and is called Jensen's inequality*.*

Proof idea. Since g is convex, taking the tangent $l(x)$ of the graph of g at any point $x \in \mathbb{R}$ gives the inequality

$$g(x) \geq l(x) = g(\mathbb{E}[X]) + \alpha(x - \mathbb{E}[X]),$$

where α is the slope of the tangent, i.e. $\alpha = g'(\mathbb{E}[X])$. Then

$$\mathbb{E}[g(X)] \geq \mathbb{E}[l(X)] = g(\mathbb{E}[X]).$$

Remark 3.35. To remember the direction of Jensen's inequality, replace $g(x)$ by x^2 and recall that

$$\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$$

always holds.

Theorem 3.36 (Generalized Tchebychev's inequality). *Let $g : \mathbb{R} \to \mathbb{R}$ be a real measurable function such that*

$$g \geq 0$$

and g is non-decreasing on $\mathbb{R}$. Then for any $c \in \mathbb{R}$ such that $g(c) > 0$ we have that

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}[g(X)]}{g(c)}$$

holds. This is called the generalized Tchebychev inequality*.*

Proof. Let $c \in \mathbb{R}$ such that $g(c) > 0$. Then

$$\mathbb{1}_{[c,\infty)}(x) \leq \frac{g(x)}{g(c)}$$

holds for all $x \in \mathbb{R}$ since g is non-decreasing. Hence

$$\begin{aligned} \mathbb{E}[\mathbb{1}_{[c,\infty)}] \leq \mathbb{E}\left[\frac{g(X)}{g(c)}\right] &= \frac{\mathbb{E}[g(X)]}{g(c)} \\ \iff \mathbb{P}(X \geq c) &\leq \frac{\mathbb{E}[g(X)]}{g(c)} \end{aligned}$$

which concludes the proof. $\square$

Example 3.37 (*Markov's inequality*)**.** Let $g(x) = \max(x, 0) = x_+$. Replace the random variable X by $|X|$ to obtain

$$\mathbb{P}(|X| > c) \leq \frac{\mathbb{E}[g(X)]}{g(c)} = \frac{\mathbb{E}[|X|]}{c}$$

for all $c > 0$, which is called *Markov's inequality.* If X admits a finite variance $\mathbb{V}(X) < \infty$, then

$$\begin{aligned}\mathbb{P}(|X - \mathbb{E}[X]| > c) &= \mathbb{P}((X - \mathbb{E}[X])^2 > c^2)\\ &\leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{c^2}\\ &= \frac{\mathbb{V}(X)}{c^2}\end{aligned}$$

holds for all $c > 0$. For example, if $X \sim \mathcal{N}(\mu, \sigma^2)$ we obtain

$$\mathbb{P}(|X - \mu| > 3\sigma) \leq \frac{\sigma^2}{9\sigma^2} = \frac{1}{9}$$

since we have $\mathbb{E}[X] = \mu$ and $\mathbb{V}(X) = \sigma^2$.

3.8 Several Random Variables: Random Vectors

Definition 3.38 (Random vector)**.** Let $X_1, \ldots, X_n$ be n real random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Consider

$$\mathbf{X} := (X_1, \ldots, X_n) : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^n, \mathcal{B}^n),$$

where

$$\begin{aligned}\mathcal{B}^n &= \sigma\left(\left\{\prod_{i=1}^n (a_i, b_i] \mid -\infty < a_i < b_i < \infty\right\}\right)\\ &\overset{(1)}{=} \sigma(\{A_1 \times \ldots \times A_n \mid A_i \in \mathcal{B}\})\end{aligned}$$

is the Borel σ-algebra on $\mathbb{R}^n$ and (1) can be shown. Then $\mathbf{X}$ is a *random vector*, meaning that it is measurable with respect to $\mathcal{A}$ and $\mathcal{B}^n$. We can also define $\mu_{\mathbf{X}}$ to be the image of $\mathbb{P}$ under $\mathbf{X}$, that is the probability measure $(\mathbb{R}^n, \mathcal{B}^n)$ induced by $\mathbf{X}$. Hence

$$\forall B \in \mathcal{B}^n : \mu_{\mathbf{X}}(B) = \mathbb{P}(\mathbf{X} \in B) = \mathbb{P}((X_1, \ldots, X_n) \in B).$$

Furthermore, the (comulative) distribution function of $\mathbf{X}$ is given by

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = \mathbb{P}\left(\mathbf{X} \in \prod_{i=1}^n (-\infty, x_i]\right) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n).$$

Discrete case. Let $\mathbf{X}(\Omega) = X_1(\Omega) \times \ldots \times X_n(\Omega)$ with $X_i(\Omega)$ being coutable. Then $\mu_{\mathbf{X}}$ has density with respect to the counting measure, so

$$\mu_{\mathbf{X}}(B) = \sum_{(x_1,\ldots,x_n)\in B} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Absolutely continuous case. Then $\mu_{\mathbf{X}}$ has density $f_{\mathbf{X}}$ with respect to the Lebesgue measure λ on $\mathbb{R}^n$, so

$$\mu_{\mathbf{X}}(B) = \int_B f_{\mathbf{X}}(x_1, \ldots, x_n)\, d(x_1, \ldots, x_n)$$

is a measurable function from $(\mathbb{R}^n, \mathcal{B}^n)$ to $(\mathbb{R}, \mathcal{B})$ and

$$\int_{\mathbb{R}^n} f_{\mathbf{X}}(x_1, \ldots, x_n)\, d(x_1, \ldots, x_n) = 1$$

holds. **Marginal distributions.** "Individual" distribution of the components $X_1, \ldots, X_n$. Fix $i \in$

$\{1, \ldots, n\}$.

Question #1. How can we deduce the distribution of the component X_i from the (joint) distribution of $\mathbf{X}$?

$\longrightarrow$ For $B \in \mathcal{B}$ we have $\mu_{X_i}(B) = \mathbb{P}(X_i \in B)$ by definition. Now, note that

$$\{X_i \in B\} = \{X_1 \in \mathbb{R}, \ldots, X_{i-1} \in \mathbb{R}, X_i \in B, X_{i+1} \in \mathbb{R}, \ldots, X_n \in \mathbb{R}\}$$

and thus

$$\mu_{X_i}(B) = \mu_{\mathbf{X}}(\mathbb{R}^{i-1} \times B \times \mathbb{R}^{n-i}).$$

Question #2. If $\mathbf{X}$ has density $f_{\mathbf{X}}$ w.r.t. the Lebesgue measure on $\mathbb{R}^n$, is the distribution of X_i absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}$ and if yes, what is its density?

$\longrightarrow$ We have that

$$\begin{aligned} \mu_{X_i}(B) &= \int_{\mathbb{R}^{i-1}\times B\times\mathbb{R}^{n-i}} f_{\mathbf{X}}(x_1, \ldots, x_n)\, d(x_1, \ldots, x_n) \\ &= \int_B \underbrace{\int_{\mathbb{R}^{i-1}\times\mathbb{R}^{n-i}} f_{\mathbf{X}}(x_1, \ldots, x_n)\, d(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)}_{\text{density of the distribution of } X_i}\, dx_i \qquad (1) \end{aligned}$$

where at (1) we used Fubini's theorem. So the answer is yes and to obtain the density of X_i, wone needs to integrate the (joint) density $f_{\mathbf{X}}$ over the remaining components.

Remark 3.39. A similar result holds for discrete distributions:

$$\mathbb{P}(X_i = x_i) = \sum_{\substack{x_j \in X_j(\Omega) \\ j\in\{1,\ldots,n\}\setminus\{i\}}} \mathbb{P}(X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \ldots, X_n = x_n).$$

Example 3.40. Consider the random pair $\mathbf{Z} = (X, Y)$ with density

$$f_{\mathbf{Z}}(x, y) = ye^{-x}\mathbb{1}_{\{x>y>0\}}.$$

We now want to compute the marginal densities f_X and f_Y. Let's check that $f_{\mathbf{Z}}$ is a density. We have

$$\begin{aligned}\int_{\mathbb{R}^2} f_{\mathbf{Z}}\, d(x,y) &= \int_{\mathbb{R}}\int_{\mathbb{R}} ye^{-x} \underbrace{\mathbb{1}_{\{x>y>0\}}}_{\mathbb{1}_{\{y>0\}}\cdot\mathbb{1}_{\{x>y\}}} \,dx\,dy \\ &= \int_0^\infty \int_y^\infty ye^{-x}\,dx\,dy = \int_0^\infty y\int_y^\infty e^{-x}\,dx\,dy \\ &= \int_0^\infty y[-e^{-x}]_y^\infty\,dy = \int_0^\infty ye^{-y}\,dy \overset{(1)}{=} 1,\end{aligned}$$

where at (1) we used integration by parts. Hence

$$\begin{aligned}f_X(x) &= \int_{\mathbb{R}} f_{\mathbf{Z}}(x,y)\,dy = \int_{\mathbb{R}} ye^{-x}\mathbb{1}_{\{x>y>0\}}\,dy \\ &= e^{-x}\cdot\mathbb{1}_{\{x>0\}}\int_0^x y\,dy = \frac{x^2}{2}e^{-x}\mathbb{1}_{\{x>0\}} \\ &= \frac{1^3}{\Gamma(3)}x^{3-1}e^{-x}\mathbb{1}_{\{x>0\}}\end{aligned}$$

and thus $X \sim \Gamma(3, 1)$ holds. Similarly we have

$$\begin{aligned}f_Y(y) &= \int_{\mathbb{R}} f_{\mathbf{Z}}(x,y)\,dx \\ &= \int_{\mathbb{R}} ye^{-x}\mathbb{1}_{\{x>y>0\}}\,dx = \int_y^\infty ye^{-x}\,dx\cdot\mathbb{1}_{\{y>0\}} \\ &= y\cdot\mathbb{1}_{\{y>0\}}\int_y^\infty e^{-x}\,dx = ye^{-y}\mathbb{1}_{\{y>0\}} \\ &= \frac{1^2}{\Gamma(2)}y^{2-1}e^{-y}\mathbb{1}_{\{y>0\}}\end{aligned}$$

and thus $Y \sim \Gamma(2, 1)$.

Definition 3.41 (Independence)**.** Let $X_1, \ldots, X_n$ be real random variables defined on $(\Omega, \mathcal{A}, \mathbb{P})$. Then $X_1, \ldots, X_n$ are said to be *independent* if

$$\forall A_1, \ldots, A_n \in \mathcal{B} : \mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i)$$

holds.

Theorem 3.42. *Let $X_1, \ldots, X_n$ be independent random variables such that for all $i \in \{1, \ldots, n\}$ the distribution of X_i admits a density f_{X_i} w.r.t. the Lebesgue measure on $\mathbb{R}$. Then the distribution of*

$\boldsymbol{X} = (X_1, \ldots, X_n)$ is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^n$ with density

$$f_{\boldsymbol{X}}(x_1, \ldots, x_n) = \prod_{i=1}^n f_{X_i}(x_i).$$

Conversely, if $\boldsymbol{X}$ admits a density $f_{\boldsymbol{X}}$ of the form

$$f_{\boldsymbol{X}}(x_1, \ldots, x_n) = g_1(x_1) \cdots g_n(x_n)$$

for all $x_1, \ldots, x_n \in \mathbb{R}$ with measurable $g_i \geq 0$, then $X_1, \ldots, X_n$ are independent with

$$f_{X_i}(x_i) = \frac{g_i(x_i)}{\int_{\mathbb{R}} g_i(x)\, dx}$$

for all $x_i \in \mathbb{R}$ and $i \in \{1, \ldots, n\}$.

Example 3.43. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector with density

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2}.$$

We have that

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}}$$

and thus $X_1, \ldots, X_n$ are independent and we have $X_i \sim \mathcal{N}(0,1)$ for all $i \in \{1, \ldots, n\}$. In this case $\mathbf{X}$ is called a *Gaussian/Normal vector* with expectation $(0, \ldots, 0)$ and covariance matrix $\Sigma = I_{n \times n} \in \mathbb{R}^{n \times n}$.

3.9 Transformation of random vectors

Let $\mathbf{X} : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^n, \mathcal{B}^n)$ be a random vector and $g : (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}^m, \mathcal{B}^m)$ be a measurable function. Then

$$\mathbf{Y} := g(\mathbf{X}) = g \circ \mathbf{X}$$

is again a random vector with distribution given by

$$\mu_{\mathbf{Y}} = \mu_{\mathbf{X}} \circ g^{-1}.$$

Theorem 3.44. *Let*

$$g : (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}^n, \mathcal{B}^n), x \mapsto m + Sx$$

with $m \in \mathbb{R}^n$ and $S \in GL_n(\mathbb{R})$ fixed. If $\boldsymbol{X}$ is a random vector admitting a density $f_{\boldsymbol{X}}$, then $\boldsymbol{Y} = g(\boldsymbol{X}) = m + S\boldsymbol{X}$ admits the density

$$f_{\boldsymbol{Y}}(y) = \frac{1}{|\det(S)|} f_{\boldsymbol{X}}(S^{-1}(y - m))$$

for all $y \in \mathbb{R}^n$.

Exercise 3.45. Let X_1 and X_2 be two independent random variables with densities f_{X_1} anf f_{X_2} respectively and let $Z = X_1 + X_2$, called the *convolution* of X_1 and X_2. Show that Z admits the density

$$f_Z(z) = \int_{\mathbb{R}} f_{X_1}(x) f_{X_2}(z - x)\, dx.$$

HINT. Note that

$$\mathbf{Y} = \begin{pmatrix} Z \\ X_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

holds.

3.10 Covariance and Correlation

Definition 3.46 (Covariance)**.** Let X_1 and X_2 be two random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ satisfying

$$\mathbb{E}[|X_1 X_2|] < \infty,$$
$$\mathbb{E}[|X_1|],\ \mathbb{E}[|X_2|] < \infty.$$

Then, the *covariance* of X_1 and X_2 is defined by

$$\operatorname{cov}(X_1, X_2) := \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])].$$

Remark 3.47. Let X_1, X_2 be random variables as in the definition of covariance. Then

(1) $\operatorname{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$,

(2) $|\operatorname{cov}(X_1, X_2)| \le \sqrt{\mathbb{E}[X_1 - \mathbb{E}[X_1]]^2}\sqrt{\mathbb{E}[X_2 - \mathbb{E}[X_2]]} \stackrel{\text{Def.}}{=} \sqrt{\mathbb{V}(X_1)}\sqrt{\mathbb{V}(X_2)}$.

Hence if $\mathbb{V}(X_1), \mathbb{V}(X_2) > 0$, then we get

$$\frac{|\operatorname{cov}(X_1, X_2)|}{\sqrt{\mathbb{V}(X_1)}\sqrt{\mathbb{V}(X_2)}} \in [0, 1].$$

PROOF IDEA.

(1) Direct computation using linearity of $\mathbb{E}$.

(2) Follows from the Cauchy-Schwarz inequality. ∷

Theorem 3.48. *Let X_1, X_2, X_3 be random variables as in the definition of covariance. Then*

(a) $\operatorname{cov}(X, X) = \mathbb{V}(X)$.

(b) $\operatorname{cov}(X_1, X_2) = \operatorname{cov}(X_2, X_1)$.

(c) $\operatorname{cov}(X_1, \alpha X_2 + \beta) = \alpha \operatorname{cov}(X_1, X_2)$ *for all* $\alpha, \beta \in \mathbb{R}$.

(d) $\mathrm{cov}(X_1, X_2 + X_3) = \mathrm{cov}(X_1, X_2) + \mathrm{cov}(X_1, X_3)$.

(e) $\mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2) + 2\mathrm{cov}(X_1, X_2)$.

(f) If X_1 and X_2 are independent, then $\mathrm{cov}(X_1, X_2) = 0$ *holds.*

Proof idea.

(a)-(d) Clear.

(e) We have

$$\begin{aligned}\mathbb{V}(X_1 + X_2) &\overset{(a)}{=} \mathrm{cov}(X_1 + X_2, X_1 + X_2) \\ &= \mathrm{cov}(X_1, X_1) + \mathrm{cov}(X_1, X_2) + \mathrm{cov}(X_2, X_1) + \mathrm{cov}(X_2, X_2) \\ &= \mathbb{V}(X_1) + 2\mathrm{cov}(X_1, X_2) + \mathbb{V}(X_2).\end{aligned}$$

In general, it holds that

$$\begin{aligned}\mathbb{V}\Big(\sum_{i=1}^n a_i X_i\Big) &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \cdot \mathrm{cov}(X_i, X_j) \\ &= \sum_{i=1}^n a_i^2 \mathbb{V}(X_i) + \sum_{1 \le i \ne j \le n} a_i a_j \cdot \mathrm{cov}(X_i, X_j) \\ &= \sum_{i=1}^n a_i^2 \mathbb{V}(X_i) + 2 \sum_{1 \le i < j \le n} a_i a_j \cdot \mathrm{cov}(X_i, X_j)\end{aligned}$$

for all $a_1, \ldots, a_n \in \mathbb{R}$.

(f) If X_1 and X_2 are independent, then

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2] \overset{\text{by Def.}}{\Longrightarrow} \mathrm{cov}(X_1, X_2) = 0$$

holds. ∷

Definition 3.49 (Correlation)**.** Let X_1 and X_2 be two random variables defined on the same probability space such that $\mathbb{V}(X_1), \mathbb{V}(X_2) \in (0, \infty)$ holds. Then the *correlation* of X_1 and X_2 is defined by

$$\rho(X_1, X_2) := \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathbb{V}(X_1)}\sqrt{\mathbb{V}(X_2)}}.$$

Remark 3.50. Let X_1, X_2 be as in the definition of correlation. Then

(a) $|\rho(X_1, X_2)| \le 1$ (*Cauchy-Schwarz inequality*).

(b) We have $\rho(X_1, X_2) = 1$ if any only if there exists a $\alpha > 0$ with
$$\mathbb{P}(X_2 - \mathbb{E}[X_2]) = \alpha(X_1 - \mathbb{E}[X_1]) = 1.$$
In the same way $\rho(X_1, X_2) = -1$ holds if any only if there exists a $\alpha > 0$ with
$$\mathbb{P}(X_1 - \mathbb{E}[X_1]) = \alpha(X_2 - \mathbb{E}[X_2]) = 1.$$
This means that correlation is a *measure of linear dependence.*

3.11 Limit Theorems

Definition 3.51 (Modes of convergence)**.** Let $(Z_n)_n$ and Z be random variables defined on the same probability space.

(a) We say that the sequence $(Z_n)_n$ *converges in probability* to Z if
$$\forall \varepsilon > 0 : \lim_{n\to\infty} \mathbb{P}(|Z_n - Z| > \varepsilon) = 0$$
holds and we denote it by $Z_n \xrightarrow{\mathbb{P}} Z$.

(b) We say that $(Z_n)_n$ *converges to Z almost surely* (in short *a.s.*) if
$$\mathbb{P}(\lim_{n\to\infty} Z_n = Z) = 1$$
holds and we denote it by $Z_n \xrightarrow{\text{a.s.}} Z$.

Lemma 3.52. *We have $Z_n \xrightarrow{a.s.} Z$ if any only if*
$$\forall \varepsilon > 0 : \quad \lim_{n\to\infty} \mathbb{P}(|Z_k - Z| \le \varepsilon, \forall k \ge n) = 1$$
holds.

Proof. First set
$$A_{n,\varepsilon} := \{|Z_k - Z| \le \varepsilon \mid k \ge n\}$$
and note that
$$\begin{aligned} Z_n \xrightarrow{\text{a.s.}} Z &\iff \mathbb{P}(\forall \varepsilon > 0\ \exists n \ge 1\ \forall k \ge n : |Z_k - Z| \le \varepsilon) = 1 \\ &\iff \mathbb{P}\Big(\bigcap_{\varepsilon>0} \bigcup_{n\ge 1} A_{n,\varepsilon}\Big) = 1. \end{aligned}$$
Note that the sequence $(\bigcup_{n\ge1} A_{n,\varepsilon})_\varepsilon$ is decreasing when ε is decreasing. Indeed, if $\varepsilon_2 < \varepsilon_1$ then
$$\begin{aligned} \bigcup_{n\ge1} A_{n,\varepsilon_2} &= \{\exists n \ge 1\ \forall k \ge n : |Z_k - Z| \le \varepsilon_2\} \\ &\subseteq \{\exists n \ge 1\ \forall k \ge n : |Z_k - Z| \le \varepsilon_1\} \\ &= \bigcup_{n\ge1} A_{n,\varepsilon_1}. \end{aligned}$$

Hence we get

$$\mathbb{P}\Big(\bigcap_{\varepsilon>0}\bigcup_{n\geq 1} A_{n,\varepsilon}\Big) = \lim_{\varepsilon\to 0}\mathbb{P}\Big(\bigcup_{n\geq 1} A_{n,\varepsilon}\Big).$$

Claim. Now we have

$$\mathbb{P}\Big(\bigcap_{\varepsilon>0}\bigcup_{n\geq 1} A_{n,\varepsilon}\Big) = 1 \iff \forall\varepsilon>0:\ \mathbb{P}\Big(\bigcup_{n\geq 1} A_{n,\varepsilon}\Big) = 1.$$

" $\Longrightarrow$ ". Observe that

$$\begin{aligned}
&\bigcup_{n\geq 1} A_{n,\varepsilon} \supseteq \bigcap_{\varepsilon>0}\bigcup_{n\geq 1} A_{n,\varepsilon}\\
&\Longrightarrow \mathbb{P}\Big(\bigcup_{n\geq 1} A_{n,\varepsilon}\Big) \geq \mathbb{P}\Big(\bigcap_{\varepsilon>0}\bigcup_{n\geq 1} A_{n,\varepsilon}\Big) = 1\\
&\Longrightarrow \forall\varepsilon>0:\ \mathbb{P}\Big(\bigcup_{n\geq 1} A_{n,\varepsilon}\Big) = 1.
\end{aligned}$$

" $\Longleftarrow$ ". If $\varepsilon > 0$ $\mathbb{P}(\bigcup_{n\geq 1} A_{n,\varepsilon}) = 1$, then we have

$$\lim_{\varepsilon\to 0}\mathbb{P}\Big(\bigcup_{n\geq 1} A_{n,\varepsilon}\Big) = 1 \implies \mathbb{P}\Big(\bigcap_{\varepsilon>0}\bigcup_{n\geq 1} A_{n,\varepsilon}\Big) = 1.$$

This proves the 'Claim'. Now, note that the sequence $(A_{n,\varepsilon})_n$ is increasing and hence

$$\mathbb{P}\Big(\bigcup_{n\geq 1} A_{n,\varepsilon}\Big) = \lim_{n\to\infty}\mathbb{P}(A_{n,\varepsilon}).$$

Therefore, we can conclude that

$$\begin{aligned}
&Z_n \xrightarrow{\text{a.s.}} Z \iff \forall\varepsilon>0: \lim_{n\to\infty}\mathbb{P}(A_{n,\varepsilon}) + 1\\
&\iff \forall\varepsilon>0: \lim_{n\to\infty}\mathbb{P}(|Z_n - Z| \leq \varepsilon, \forall k \geq n) = 1
\end{aligned}$$

holds. □ □

Theorem 3.53.

(i) Almost sure convergence implies convergence in probability.

(ii) If we have

$$\forall\varepsilon>0:\quad \sum_{n=1}^{\infty}\mathbb{P}(|Z_n - Z| > \varepsilon) < \infty$$

then

$$Z_n \xrightarrow{a.s.} Z$$

holds.

Proof.

(i) We have that

$$\{|Z_k - Z| \le \varepsilon, \forall k \ge n\} \subseteq \{|Z_n - Z| \le \varepsilon\}.$$

Now, form Lemma 3.52 it follows that

$$\begin{aligned} Z_n \xrightarrow{\text{a.s.}} Z &\iff \forall \varepsilon > 0 : \lim_{n\to\infty} \mathbb{P}(|Z_n - Z| \le \varepsilon, \forall k \ge n) = 1 \\ &\implies \forall \varepsilon > 0 : \lim_{n\to\infty} \mathbb{P}(|Z_n - Z| \le \varepsilon) = 1 \qquad (1) \\ &\iff \forall \varepsilon > 0 : \lim_{n\to\infty} \mathbb{P}(|Z_n - Z| > \varepsilon) = 0 \\ &\iff Z_n \xrightarrow{\mathbb{P}} Z, \end{aligned}$$

where at (1) we used that

$$\mathbb{P}(|Z_n - Z| \le \varepsilon) \ge \mathbb{P}(|Z_n - Z| \le \varepsilon, \forall k \ge n)$$

holds.

(ii) For a fixed $\varepsilon > 0$ define

$$B_{n,\varepsilon} := \{|Z_n - Z| > \varepsilon\}.$$

By the first statement of the *Borel-Cantelli lemma*, we have

$$\sum_{n\ge1} \mathbb{P}(B_{n,\varepsilon}) < \infty \implies \mathbb{P}(B_{\infty,\varepsilon}) = 0,$$

where

$$B_{\infty,\varepsilon} := \bigcap_{n\ge1} \bigcup_{k\ge n} B_{k,\varepsilon} \quad (= \{B_{k,\varepsilon} \text{ infinitely often}\}).$$

Hence we have

$$\begin{aligned} \mathbb{P}(B^c_{\infty,\varepsilon}) = 1 &\iff \mathbb{P}\Big(\bigcup_{n\ge1} \underbrace{\bigcap_{k\ge n} B^c_{k,\varepsilon}}_{=A_{n,\varepsilon}}\Big) = 1 \\ &\implies \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{n\ge1} A_{n,\varepsilon}\Big) = 1 \\ &\iff \lim_{n\to\infty} \mathbb{P}(A_{n,\varepsilon}) = 1 \\ &\iff \lim_{n\to\infty} \mathbb{P}(|Z_k - Z| \le \varepsilon, \forall k \ge n) \end{aligned}$$

and thus again using Lemma 3.52 we can conclude that

$$Z_n \xrightarrow{\text{a.s.}} Z$$

holds since $\varepsilon > 0$ was arbitrary. $\square$

Example 3.54. Note that convergence in probability does not imply convergence almost surely. In fact, consider $(X_n)_n$ to be independent Bernoulli random variables such that

$$\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = \frac{1}{i}$$

holds for all $i \geq 1$. For $\varepsilon > 0$ we have

$$\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(X_n > \varepsilon) = \begin{cases} \mathbb{P}(X_n = 1) = \frac{1}{n} & \text{if } \varepsilon \in (0,1) \\ 0 & \text{if } \varepsilon \geq 1. \end{cases}$$

Hence

$$\forall \varepsilon > 0 : \lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \varepsilon) = 0$$

and thus $X_n \xrightarrow{\mathbb{P}} 0$ holds. Now observe that

$$\sum_{n\geq 1} \mathbb{P}(X_n = 1) = \sum_{n\geq 1} \frac{1}{n} = \infty$$

also holds and since $(X_n)_n$ are independent, it follows from the second statement of the *Borel-Cantelli lemma* that

$$\mathbb{P}\Big(\bigcap_{n\geq 1} \bigcup_{k\geq n} \{X_k = 1\} \Big) = 1.$$

Since $(\bigcup_{k\geq n}\{X_k = 1\})_n$ is decreasing, we have that

$$\mathbb{P}\Big(\bigcap_{n\geq 1} \bigcup_{k\geq n} \{X_k = 1\} \Big) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{k\geq n} \{X_k = 1\} \Big)$$

and thus

$$\lim_{n\to\infty} \mathbb{P}(X_i = 1 \text{ for some } k \geq n) = 1. \tag{1}$$

Suppose for a contradiction that $X_n \xrightarrow{\text{a.s.}} 0$ holds. Then

$$\forall \varepsilon > 0 : \lim_{n\to\infty} \mathbb{P}(X_k \leq \varepsilon, \forall k \geq n) = 1$$

must holds. For $\varepsilon \in (0,1)$ this means

$$\lim_{n\to\infty} \mathbb{P}(X_k = 0, \forall k \geq n) = 1 \iff \lim_{n\to\infty} \mathbb{P}(X_k = 1 \text{ for some } k \geq n) = 0$$

which contradicts (1). Hence $(X_n)_n$ does not converges almost surely to 0.

3.12 Weak Law of Large Numbers (W.L.L.N.)

Proposition 3.55 (W.L.L.N.). *Let $X_1, X_2, \dots, X_n$ be random variables defined on the same probability space. Assume that*

$$\begin{aligned} \forall i \in \{1, \dots, n\}: &\quad \mathbb{V}(X_i) < \infty, \\ \forall 1 \le i \neq j \le n: &\quad \operatorname{cov}(X_i, X_j) = 0, \\ \forall i \in \{1, \dots, n\}: &\quad \mathbb{E}[X_i] = m \in \mathbb{R}, \\ \sum_{i=1}^{n} \mathbb{V}(X_i) &= o(n^2) \ \textit{as } n \to \infty, \end{aligned}$$

hold. Then we have

$$\overline{X}_n \xrightarrow{\mathbb{P}} m$$

with

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

which is called the sample mean *or* empirical mean*.*

Proof. We have

$$\mathbb{E}[\overline{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} \underbrace{\sum_{i=1}^{n} \mathbb{E}[X_i]}_{n \cdot m} = m$$

and

$$\begin{aligned} \mathbb{V}(\overline{X}_n) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathbb{V}(X_i) + 2\sum_{1 \le i < j \le n} \underbrace{\operatorname{cov}(X_i, X_j)}_{=0}\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(X_i). \end{aligned}$$

By Theorem 3.36 we have

$$\begin{aligned} &\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| > \varepsilon) \le \frac{\mathbb{V}(\overline{X}_n)}{\varepsilon^2} \\ \iff &\mathbb{P}(|\overline{X}_n - m| > \varepsilon) \le \frac{1}{n^2} \underbrace{\sum_{i=1}^{n} \mathbb{V}(X_i)}_{=n\mathbb{V}(X_1)} \frac{1}{\varepsilon^2} = o(1)\frac{1}{\varepsilon^2} = o(1) \end{aligned}$$

as $n \to \infty$ and thus

$$\forall \varepsilon > 0 : \lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - m| > \varepsilon) = 0 \iff \overline{X}_n \xrightarrow{\mathbb{P}} m$$

which concludes the proof. □

SPECIAL CASE. Let $X_1, \ldots, X_n$ be i.i.d. random variables such that $\mathbb{E}[X_i] = m$ and $\mathbb{V}(X_i) = \sigma^2$ holds for every $i \in \{1, \ldots, n\}$. Then $\mathrm{cov}(X_i, X_j) = 0$ holds for $i \neq j$ and we have

$$\sum_{i=1}^{n} \mathbb{V}(X_i) = n\sigma^2 = o(n^2)$$

as $n \to \infty$. In this case, we have

$$\overline{X}_n \xrightarrow{\mathbb{P}} m.$$

Theorem 3.56 (S.L.L.N.). *Let $X_1, \ldots, X_n$ be i.i.d. random vairbales such that*

$$\mathbb{E}[X_i^2] = \mathbb{E}[X_1^2] < \infty.$$

Then $\overline{X}_n \xrightarrow{a.s.} \mathbb{E}[X_1] = m$ holds as $n \to \infty$.

Remark 3.57. We have

(a) $\mathbb{E}[X_1^2] < \infty \implies \mathbb{V}(X_1) < \infty$.

(b) The assumption that $\mathbb{E}[X_1^2] < \infty$ should hold is "too strong". Indeed, the SLLN holds under the weaker condition that $\mathbb{E}[|X_1|] < \infty$ holds but the proof for this will be more involved.

(c) Note that S.L.L.N. $\implies$ W.L.L.N. holds since convergence a.s. implies convergence in probability.

Proof. (of (a)) By Jensen's inequality 3.34 we have

$$\mathbb{E}[|X_1|] \leq \sqrt{\mathbb{E}[X_1^2]} < \infty$$

which implies

$$\mathbb{V}(X_1) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 < \infty.$$

□

3.13 Weak Convergence (Convergence in Law / Distribution)

Definition 3.58.

- Let μ_n for $n \geq 1$ and μ be probability measures on $(\mathbb{R}, \mathcal{B})$. We say that the sequence $(\mu_n)_n$ *converges weakly* to μ if
$$\int f \, d\mu_n \xrightarrow{n\to\infty} \int f \, d\mu$$
holds for all continuous and bounded functions f.

- Let Z_n for $n \geq 1$ and Z be random variables (not necessarily defined on the same probability space). We say that the sequence $(Z_n)_n$ *converges weakly* or *in law/distribution* to Z if $(\mu_{Z_n})_n$ converges weakly to μ_Z, where μ_{Z_n} and μ_Z are the distributions of Z_n and Z respectively. This means

$$\int f \, d\mu_{Z_n} \xrightarrow{n\to\infty} \int f \, d\mu_Z$$
$$\text{or equivalently} \quad \mathbb{E}[f(Z_n)] \xrightarrow{n\to\infty} \mathbb{E}[f(Z)]$$

should hold for all continuous and bounded functions f on $\mathbb{R}$. We denote this by

$$\mu_n \xrightarrow{d} \mu, \quad Z_n \xrightarrow{d} Z$$
$$\text{or} \quad \mu_n \xrightarrow{\mathcal{L}} \mu, \quad Z_n \xrightarrow{\mathcal{L}} Z.$$

Lemma 3.59. *Let μ_n and μ be probability measures on $(\mathbb{R}, \mathcal{B})$ with (commulative) distribution functions F_n and F respectively, that is $F_n(x) = \mu_n((-\infty, x])$ and $F(x) = \mu((-\infty, x])$ for $x \in \mathbb{R}$. Then the following statements are equivalent:*

- $\mu_n \xrightarrow{d} \mu$.
- $F_n(x) \xrightarrow{n\to\infty} F(x)$ *holds for any continuity point x of F.*
- $\int f \, d\mu_n \xrightarrow{n\to\infty} \int f \, d\mu$ *for any $f \in C_b^3(\mathbb{R})$, where*

$$C_b^3(\mathbb{R}) := \{f \in C^3(\mathbb{R}) \mid \exists M > 0 : \sup_{j\in\{0,1,2,3\}} |f^{(j)}(x)| < M\}.$$

Theorem 3.60 (Lévy's continuity theorem)**.** *Let Z_n and Z be random variables and define*

$$\varphi_{Z_n}(t) := \mathbb{E}[e^{itZ_n}]$$
$$\varphi_Z(t) := \mathbb{E}[e^{itZ}]$$

for $t \in \mathbb{R}$, which are called characteristic functions *of Z_n and Z. Then*

$$Z_n \xrightarrow{d} Z \iff \forall t \in \mathbb{R} : \; \varphi_{Z_n}(t) \xrightarrow{n\to\infty} \varphi_Z(t)$$

holds.

Example 3.61. Let $X_n \sim \text{Bin}\left(n, \frac{\lambda}{n}\right)$ for $\lambda \in (0, \infty)$ and $n \in \mathbb{N}$ such that $n > \lambda$. Then $X_n \xrightarrow{d} X \sim \text{Pois}(\lambda)$ holds.

Proof. We have

$$
\begin{aligned}
\varphi_{X_n}(t) = \mathbb{E}[e^{itX_n}] &= \sum_{k=0}^{n} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} e^{itk} \\
&= \sum_{k=0}^{n} \binom{n}{k} \left(\frac{\lambda}{n}e^{it}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \left(\frac{\lambda}{n}e^{it} + 1 - \frac{\lambda}{n}\right)^n = \left(1 + \frac{\lambda(e^{it}-1)}{n}\right)^n \xrightarrow{n\to\infty} e^{\lambda(e^{it}-1)}
\end{aligned}
$$

where in the last step we used $\lim_{n\to\infty}\left(1+\frac{\xi}{n}\right)^n = e^{\xi}$ holds for $\xi \in \mathbb{C}$. Similarly, we have

$$
\begin{aligned}
\varphi_X(t) = \mathbb{E}[e^{itx}] &= \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} e^{itk} \\
&= e^{-\lambda} \underbrace{\sum_{k=0}^{\infty} \frac{(e^{it}\lambda)^k}{k!}}_{=e^{\lambda e^{it}}} = e^{\lambda(e^{it}-1)}
\end{aligned}
$$

which proves the example by using Theorem 3.60. □

3.14 The Central Limit Theorem (C.L.T.)

Theorem 3.62 (CLT)**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables with $\mathbb{E}[X_i] = m \in \mathbb{R}$ and $\mathbb{V}(X_i) = \sigma^2 \in (0, \infty)$ for any $i \in \{1, \dots, n\}$. Then*

$$
\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \xrightarrow{d} Z \sim \mathcal{N}(0,1)
$$

holds, where again $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ as usual. (This means that

$$
\mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \le \xi\right) \xrightarrow{n\to\infty} \int_{-\infty}^{\xi} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx
$$

holds for any $\xi \in \mathbb{R}$.)

Example 3.63. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ with $p \in (0,1)$. Then $\mathbb{E}[X_i] = p$ and $\mathbb{V}(X_i) = p(1-p) \in (0,\infty)$ hold and thus by Theorem 3.62 we have

$$
\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0,1)
$$

or equivalently

$$
\sqrt{n}(\overline{X}_n - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)).
$$

Example 3.64. Suppose a load of 49 boxes is to be transported by an elevator. The weight of the boxes have expected value $m = 92$ kg and standard deviation $\sigma = 6$ kg.

QUESTION. What is the probability that the 49 boxes can be transported if we know that the maximal weight should not exeed 4410 kg?
$\longrightarrow$ ANSWER. Let $p =$ "the probability that the 49 boxes can be transported". Let $X_1, \ldots, X_n$ with $n = 49$ be the weights of the boxes. Then by the assumptions we have

$$\begin{aligned}
p &= \mathbb{P}(\sum_{i=1}^{49} X_i \le 4410) = \mathbb{P}\left(\overline{X}_{49} \le \frac{4410}{49}\right) \\
&= \mathbb{P}\left(\frac{\sqrt{49}(\overline{X}_{49} - m)}{\sigma} \le \frac{\sqrt{49}}{\sigma}\left(\frac{4410}{49} - m\right)\right) \\
&= \mathbb{P}\left(\frac{\sqrt{49}(\overline{X}_{49} - m)}{\sigma} \le \underbrace{\frac{7}{6}(90 - 92)}_{\approx -2.333}\right) \\
&\approx \mathbb{P}(Z \le -2.333) = 0.0098
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$ by using Theorem 3.62.

You made it!

Statistics

4 Introduction to Statistics

4.1 Notation

Notation. Let X be a random variable $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$.

- We know that X induces a probability measure, denoted by μ_X, that is

  $$\forall B \in \mathcal{B} : \mu_X(B) = \mathbb{P}(X \in B).$$

  Here, we will denote μ_X by P.

- We will write $X \sim P$ to mean that X has distribution equal to P.

  Problem. In statistical applications, the distribution P is *unknown*.

  $\longrightarrow$ Solution. We will estimate P based on i.i.d. "copies" of X, so $X_1, \ldots, X_n$.

- We write $\mathcal{X} := X(\Omega) =$ "the sample space (to which the values of X belong)" and

  $$\mathbf{X}_n := (X_1, \ldots, X_n) \in \mathcal{X}^n = \text{"the random sample of size } n\text{"}.$$

4.2 (Parametric) Statistical Models

Definition 4.1. A *(parametric) statistical model* stipulates that

$$P \in \mathcal{P} := \{P_\theta \mid \theta \in \Theta\},$$

where $\Theta \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$ is a *parametric space* and P_θ is a probability measure on $(\mathbb{R}, \mathcal{B})$ for all $\theta \in \Theta$. In particular, if $X \sim P = P_\theta$ for some $\theta \in \Theta$ and if X admits a finite expectation, we will write $\mathbb{E}_\theta[X] := \mathbb{E}[X]$. Also, if X admits a finite variance, then we will write $\mathrm{Var}_\theta(X) := \mathbb{V}(X)$.

Example 4.2. Let $X \sim \mathrm{Pois}(\theta)$ for some $\theta \in (0, \infty)$. This means that

$$P \in \{P_\theta \mid \theta \in \underbrace{(0, \infty)}_{=\Theta}\}$$

with

$$P_\theta(B) = \sum_{k \in B} \frac{e^{-\lambda} \lambda^k}{k!}$$

for all $B \in \mathcal{B}$. Suppose that $X \sim \mathcal{N}(\mu, \sigma^2)$ and put $\theta = (\mu, \sigma^2)$. Then

$$P \in \mathcal{P} = \{P_\theta \mid \theta \in \mathbb{R} \times (0, \infty)\}$$

and

$$\forall B \in \mathcal{B} : P_\theta(B) = \int_B \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx.$$

4.3 Parametric of Interest and Estimators

Let $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ be some (parametric) statistical model.

Definition 4.3. A *parameter of interest* is $\gamma = Q(P)$, where $Q : \mathcal{P} \to \Gamma \subseteq \mathbb{R}^k$ is some given map for $k \in \mathbb{N}$. For $\theta \in \Theta$ we will write $g(\theta) = Q(P_\theta)$ where $g : \Theta \to \Gamma$.

Examples 4.4.

- Consider again $X \sim \mathcal{N}(\mu, \sigma^2)$ and let

  $$\gamma = Q(P) = \int_{\mathbb{R}} x \, dP(x)$$

  be the parameter of interest. Then for $\theta = (\mu, \sigma^2)$ we have

  $$g(\theta) = \int_{\mathbb{R}} x \, dP_\theta(x) = \mathbb{E}_\theta[X] = \mu.$$

- Let $X \sim \text{Exp}(\lambda)$ for $\lambda \in (0, \infty)$, so $\theta = \lambda$ and $\Theta = (0, \infty)$. Set $g(\lambda) = \lambda$, which means that we are "interested" in the rate λ. Now compute

  $$\begin{aligned} \mathbb{E}_\lambda[X] &= \int_0^\infty x\lambda e^{-\lambda x} \, dx = \underbrace{[-xe^{-\lambda x}]_0^\infty}_{=0} + \int_0^\infty e^{-\lambda x} \, dx \\ &= \frac{1}{\lambda} \underbrace{\int_0^\infty \lambda e^{-\lambda x} \, dx}_{=1} = \frac{1}{\lambda}. \end{aligned}$$

  But this is equivalent to

  $$\lambda = \frac{1}{\mathbb{E}_\lambda[X]} = \left(\int_{\mathbb{R}} x \, dP_\lambda(x) \right)^{-1}$$

  and thus

  $$Q(P) = \left(\int_{\mathbb{R}} x \, dP(x) \right)^{-1}$$

  holds.

We consider a random sample $\mathbf{X}_n = (X_1, \ldots, X_n) \in \mathcal{X}^n = X(\Omega)^n$.

Definition 4.5 (Estimator)**.** An *estimator* T is a measurable map $T : \mathcal{X}^n \to \Gamma$. We will also call the value $T(X_1, \ldots, X_n)$ an *estimator* or a *statistic*.

Examples 4.6.

- Suppose we observe $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then consider

  $$T_1(X_1, \ldots, X_n) := X_1$$

and

$$T_2(X_1, \ldots, X_n) := \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Question. Which estimator is "better"?

- Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, \theta)$ with $\theta \in (0, \infty) =: \Theta$ and consider the estimators

$$\begin{aligned} T_1(X_1, \ldots, X_n) &:= 2\overline{X}_n \\ T_2(X_1, \ldots, X_n) &:= \max_{1 \le i \le n} X_i \\ T_3(X_1, \ldots, X_n) &:= \frac{n+1}{n} \max_{1 \le i \le n} X_i \end{aligned}$$

Question. Which estimator is "best"?

4.4 The L.L.N. and Constructing Estimators

Note. Recall that if $X_1, \ldots, X_n$ are i.i.d. random variables such that $\mathbb{E}[|X_1|] < \infty$, then by the W.L.L.N. (Proposition 3.55) we have

$$\overline{X}_n \xrightarrow{\mathbb{P}} m := \mathbb{E}[X_1] = \mathbb{E}[X_i]$$

for all $1 \le i \le n$. If we are interested in $\mu \ (= \gamma) = \int_{\mathbb{R}} x \, dP(x)$, then a sensible estimator is $\overline{X}_n$ (at least for n large enough).

Theorem 4.7 (Continuous mapping theorem)**.** *Let f be a real function with $C_f = \{$points of continuity of $f\}$. For a random variable Z such that $\mathbb{P}(Z \in C_f) = 1$, it holds that*

$$\begin{aligned} Z_n \xrightarrow{\mathbb{P}} Z &\implies f(Z_n) \xrightarrow{\mathbb{P}} f(Z), \\ Z_n \xrightarrow{a.s.} Z &\implies f(Z_n) \xrightarrow{a.s.} f(Z). \end{aligned}$$

This means that if the parameter of interest $g(\theta) = Q(P_\theta)$ takes the form $g(\theta) = f(\mathbb{E}_\theta[X])$ with X a random variable having the same distribution as $X_1, \ldots, X_n$ and f is a continuous function, then

$$T(X_1, \ldots, X_n) = f(\overline{X}_n)$$

is a sensible estimator of $g(\theta)$. We have

$$f(\overline{X}_n) \xrightarrow{\text{a.s.}/\mathbb{P}} f(\mathbb{E}_\theta[X]).$$

Example 4.8. Consider again $X, X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$ with $\lambda \in (0, \infty) = \Theta$ and $g(\lambda) = \lambda$. We have already shown that

$$g(\lambda) = \frac{1}{\mathbb{E}_\lambda[X]} = f(\mathbb{E}_\lambda[X])$$

holds for $f(x) = x^{-1}$ continuous on Θ. Then by the Continuous mapping theorem 4.7 we know that

$$T(X_1, \dots, X_n) = \frac{1}{\overline{X}_n} = f(\overline{X}_n)$$

is a good estimator of $g(\lambda) = \lambda$ since

$$f(\overline{X}_n) \xrightarrow{\text{a.s.}/\mathbb{P}} \lambda$$

holds. On the other hand, suppose that the parameter of interest $g(\theta)$ takes the form $\mathbb{E}_\theta[k(X)]$ with k such that $\mathbb{E}_\theta[|k(X)|] < \infty$. Then, by the L.L.N., a sensible estimator would be

$$T(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^n k(X_i).$$

Examples 4.9.

- Consider $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. We are interested in estimating σ^2, where $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. We have

$$\sigma^2 = \mathrm{Var}_\theta(X) = \mathbb{E}_\theta[X^2] - \mathbb{E}_\theta[X]^2.$$

  Now consider the estimator

$$\begin{aligned} T(x_1, \dots, x_n) &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2 \\ &= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2. \quad (\longrightarrow \textit{sample/empirical variance}) \end{aligned}$$

  In fact. We have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_i)^2 &= \frac{1}{n}\sum_{i=1}^n (X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2) \\ &= \frac{1}{n}\sum_{i=1}^n X_i^2 - 2\overline{X}_n \underbrace{\frac{1}{n}\sum_{i=1}^n X_i}_{=\overline{X}_n} + \overline{X}_n^2 \\ &= \frac{1}{n}\sum_{i=1}^n X_i^2 - 2\overline{X}_n + \overline{X}_n^2 \\ &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2. \end{aligned}$$

- Suppose that based on i.i.d. random variables $X, X_1, \dots, X_n$ we are interested in estimating the (common) comulative distribution function

$$F_{X_1}(t) = \mathbb{P}(X_1 \le t) = F_X(t) = \mathbb{P}(X \le t)$$

for $t \in \mathbb{R}$. Then

$$\begin{aligned} F_X(t) &= \int_\Omega \mathbb{1}_{\{X(\omega)\le t\}}\, d\mathbb{P}(\omega) \\ &= \int_\mathbb{R} \mathbb{1}_{\{x\le t\}}\, dP(x) \\ &= \mathbb{E}_P[\underbrace{\mathbb{1}_{\{X\le t\}}}_{=:k(X)}] \end{aligned}$$

with $P = \mu_X =$ the distribition of X. A sensible estimator is

$$\hat{F}_n(t) := \mathbb{F}_n(t) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i\le t\}}$$

where $\hat{F}_n = \mathbb{F}_n$ is called the *empirical/sample cumulative distribution function*. By the L.L.N. we have

$$\hat{F}_n(t) \xrightarrow{\text{a.s.}/\mathbb{P}} F_X(t)$$

for every $t \in \mathbb{R}$ and by the CLT 3.62 we also have

$$\sqrt{n}\frac{\hat{F}_n(t) - F_X(t)}{\sqrt{F_X(t)(1-F_X(t))}} \xrightarrow{d} \mathcal{N}(0,1).$$

for $t \in \mathbb{R}$ such that $F(t) \in (0,1)$. We also get

$$\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F_X(t)| \xrightarrow{\text{a.s.}} 0,$$

which is called the *Glivenko-Cantelli theorem.*

4.5 Mean Squared Error

Definition 4.10. The *mean squared error* (*MSE*) of some estimator T of $g(\theta)$ is the quantity

$$\begin{aligned} \text{MSE}_\theta(T) &:= \mathbb{E}_\theta[(T-g(\theta))^2] \\ &= \mathbb{E}_\theta[(T(X_1,\dots,X_n)-g(\theta))^2]. \end{aligned}$$

The *bias* of T is the quantity

$$\text{bias}_\theta(T) := \mathbb{E}_\theta[T] - g(\theta) = \mathbb{E}_\theta[T(X_1,\dots,X_n)] - g(\theta).$$

The estimator T is said to be *unbiased* if

$$\forall\theta\in\Theta : \mathbb{E}_\theta[T] = g(\theta)$$

holds, which is equivalent to $\text{bias}_\theta(T) = 0$ for every $\theta \in \Theta$.

Lemma 4.11. *We always have*

$$\mathrm{MSE}_\theta(T) = \mathrm{bias}_\theta(T)^2 + \mathrm{Var}_\theta(T).$$

Proof. We have

$$\begin{aligned}
&\mathrm{MSE}_\theta(T) = \mathbb{E}_\theta[(T - g(\theta))^2] \\
&= \mathbb{E}_\theta[((T - \mathbb{E}_\theta[T]) - (g(\theta) - \mathbb{E}_\theta[T]))^2] \\
&= \mathbb{E}_\theta[(T - \mathbb{E}_\theta[T])^2 - 2(T - \mathbb{E}_\theta[T])(g(\theta) - \mathbb{E}_\theta[(T)]) + (g(\theta) - \mathbb{E}_\theta[T])^2] \\
&= \mathbb{E}_\theta[(T - \mathbb{E}_\theta[T])^2] - 2\underbrace{\mathbb{E}_\theta[(T - \mathbb{E}_\theta[(T)])]}_{=\mathbb{E}_\theta[T] - \mathbb{E}_\theta[T] = 0}(g(\theta) - \mathbb{E}_\theta[T]) + (g(\theta) - \mathbb{E}_\theta[T])^2 \\
&= \mathrm{Var}_\theta(T) + \mathrm{bias}_\theta(T)^2
\end{aligned}$$

which proves the lemma. □

Examples 4.12.

- Let $X, X_1, \ldots, X_n$ be i.i.d. random variables with finite expectation μ and finite variance σ^2.

Then we have

$$\begin{aligned}\mathbb{E}[\overline{X}_n] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}n\mu = \mu\end{aligned}$$

and thus $\overline{X}_n$ is an unbiased estimator of μ for every $\mu \in \mathbb{R}$. We also have

$$\begin{aligned}\mathrm{MSE}_\mu(\overline{X}_n) &= \underbrace{\mathrm{bias}_\mu(\overline{X}_n)^2}_{=0} + \mathrm{Var}_\mu(\overline{X}_n)\\ &= \mathrm{Var}_\mu\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}_\mu\Big(\sum_{i=1}^{n} X_i\Big)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}_\mu(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

Now let $T_1(X_1, \ldots, X_n) := X_1$ as in a previous example. Then we have

$$\mathrm{MSE}_\mu(T_1) = \mathrm{MSE}_\mu(X_1) = \mathrm{Var}_\mu(X_1) = \sigma^2$$

and thus $\overline{X}_n$ is strictly better that T_1 in the sense of the MSE for all $n \geq 2$.

- Consider the same setting as above but now we are interested in estimating σ^2. For this, consider the estimator

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

for $n \geq 2$. We show here that S_n^2 is unbiased as follows.

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 &= \sum_{i=1}^{n}(X_i - \mu - (\overline{X}_n - \mu))^2\\ &= \sum_{i=1}^{n}[(X_i - \mu)^2 - 2(X_i - \mu)(\overline{X}_n - \mu) + (\overline{X}_n - \mu)^2]\\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X}_n - \mu)\cdot\underbrace{\sum_{i=1}^{n}(X_i - \mu)}_{=n(\overline{X}_n - \mu)} + n(\overline{X}_n - \mu)^2.\end{aligned}$$

Thus we have

$$\begin{aligned}
\mathbb{E}_\theta[S_n^2] &= \frac{1}{n-1}\mathbb{E}_\theta[\sum_{i=1}^n (X_i-\mu)^2 - n(\overline{X}_n-\mu)^2] \\
&= \frac{1}{n-1}(\sum_{i=1}^n \underbrace{\mathbb{E}_\theta[(X_i-\mu)^2]}_{=\mathrm{Var}_\theta(X_i)=\sigma^2} - n\mathbb{E}_\theta[(\overline{X}_n-\mu)^2]) \\
&= \frac{1}{n-1}\left(n\sigma^2 - n\frac{1}{n}\sigma^2\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2
\end{aligned}$$

and thus the estimator is unbiased. Note that this means that the sample variance

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$$

is a biased estimator of σ^2. Indeed, we have

$$\begin{aligned}
\hat{\sigma}_n^2 &= \frac{(n-1)S_n^2}{n} \\
\Longrightarrow \mathbb{E}[\hat{\sigma}_n^2] &= \frac{n-1}{n}\mathbb{E}_\theta[S_n^2] = \frac{n-1}{n}\sigma^2 \\
&= \sigma^2 - \frac{\sigma^2}{n}
\end{aligned}$$

and thus

$$\mathrm{bias}_\theta(\hat{\sigma}_n^2) = \mathbb{E}_\theta[\hat{\sigma}_n^2] - \sigma^2 = -\frac{\sigma^2}{n} \xrightarrow{n\to\infty} 0$$

but $\hat{\sigma}_n^2$ is always biased. Note that by the W.L.L.N 3.55 we have

$$\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 \xrightarrow{\mathbb{P}} \mathbb{E}[(X_1-\mu)^2] = \sigma^2 = \mathrm{Var}(X_1)$$

and

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu.$$

With the function $f(x) = (x-\mu)^2$ we have, by the Continuous mapping theorem 4.7, that

$$(\overline{X}_n - \mu)^2 = f(\overline{X}_n) \xrightarrow{\mathbb{P}} f(\mu) = 0$$

and thus

$$\hat{\sigma}_n^2 \xrightarrow{\mathbb{P}} \sigma^2 - 0 = \sigma^2.$$

We also have $S_n^2 = \frac{n}{n-1}\hat{\sigma}_n^2$ and thus

$$S_n^2 \xrightarrow{\mathbb{P}} \sigma^2$$

also follows. Now take $h(x) = \sqrt{x}$ and conclude that

$$\hat{\sigma}_n = h(\hat{\sigma}_n^2) \xrightarrow{\mathbb{P}} h(\sigma^2) = \sigma$$

again by the Continuous mapping theorem 4.7 and similarly $S_n \xrightarrow{\mathbb{P}} \sigma$.

4.6 The C.L.T. and Building Confidence Intervals

Recall that if $X_1, \dots, X_n$ are i.i.d. random variables such that $\mathbb{E}[X_i] = \mu \in \mathbb{R}$ and $\mathrm{Var}(X_i) = \sigma^2 \in (0, \infty)$ for every $i \in \{1, \dots, n\}$, then

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1).$$

Slutskey's Theorem 4.13. *If $Z_n \xrightarrow{d} Z$ and $A_n \xrightarrow{\mathbb{P}} a \in \mathbb{R}$, then $A_n Z_n \xrightarrow{d} aZ$ holds. Note that here the number a is not random.*

$\longrightarrow$ Consequence. By the CLT 3.62 and Slutskey's Theorem 4.13 we have

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\tilde{\sigma}_n} \xrightarrow{d} \mathcal{N}(0,1)$$

for any estimator $\tilde{\sigma}_n$ such that $\tilde{\sigma}_n \xrightarrow{\mathbb{P}} \sigma$ holds.
In fact. Consider the function $f(x) = \frac{\sigma}{x}$ for $x \in (0, \infty)$. By the CLT 3.62, we have that

$$f(\tilde{\sigma}_n) \xrightarrow{\mathbb{P}} 1$$

and thus

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\tilde{\sigma}_n} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \frac{\sigma}{\tilde{\sigma}_n} \xrightarrow{d} \mathcal{N}(0,1)$$

holds.

4.6.1 Application: Confidence Interval for the Expectation μ

For $a < b$ we have

$$\mathbb{P}\left(a < \frac{\sqrt{n}(\overline{X}_n - \mu)}{\tilde{\sigma}_n} \le b\right) \xrightarrow{n \to \infty} \mathbb{P}(a < Z \le b) \tag{1}$$

with $Z \sim \mathcal{N}(0,1)$ as a consequence of Lemma 3.59 and CLT 3.62. Indeed, we have

$$\mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\tilde{\sigma}_n} \le b\right) \xrightarrow{n \to \infty} \mathbb{P}(Z \le b)$$

and

$$\mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\tilde{\sigma}_n} \le a\right) \xrightarrow{n \to \infty} \mathbb{P}(Z \le a).$$

Now taking the difference shows (1). Note that (1) is also equivalent to saying

$$\begin{aligned}\mathbb{P}\left(\overline{X}_n - \frac{b\tilde{\sigma}_n}{\sqrt{n}} \le \mu < \overline{X}_n - \frac{a\tilde{\sigma}_n}{\sqrt{n}}\right) \xrightarrow{n \to \infty} \mathbb{P}(a < Z \le b) &= \mathbb{P}(Z \le b) - \mathbb{P}(Z \le a) \\ &= \Phi(b) - \Phi(a),\end{aligned}$$

where $\Phi(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} e^{-\frac{t^2}{2}}\, dt$ for $\xi \in \mathbb{R}$. We can now take a and b such that $\Phi(b) - \Phi(a) = 1 - \alpha$ with $\alpha \in (0, 1)$ small.
For example, we can take $a = \Phi^{-1}\left(\frac{\alpha}{2}\right)$ the $\frac{\alpha}{2}$-quantile of Φ of Z and $b = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ the $\left(1 - \frac{\alpha}{2}\right)$-quantile of Φ. We will write $a = \zeta_{\frac{\alpha}{2}}$ and $b = \zeta_{1-\frac{\alpha}{2}}$. It turns out that $a = -b$ in this case. To show this, it is enough to show that

$$\Phi\left(-\zeta_{1-\frac{\alpha}{2}}\right) = \frac{\alpha}{2}.$$

Let $\zeta \in \mathbb{R}$. Then

$$\begin{aligned}\Phi(-\zeta) &= \int_{-\infty}^{-\zeta} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt = \int_{\infty}^{\zeta} \frac{1}{\sqrt{2\pi}} e^{-\frac{(-t)^2}{2}} (-1)\, dt \\ &= 1 - \int_{-\infty}^{\zeta} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt = 1 - \Phi(\zeta).\end{aligned}$$

Therefore, we have

$$\Phi\left(-\zeta_{1-\frac{\alpha}{2}}\right) = 1 - \Phi\left(\zeta_{1-\frac{\alpha}{2}}\right) = 1 - \left(1 - \frac{\alpha}{2}\right) = \frac{\alpha}{2}.$$

Hence, with this choice of a and b, we have that

$$\begin{aligned} &\mathbb{P}_\mu\left(\overline{X}_n - \frac{\zeta_{1-\frac{\alpha}{2}}}{\sqrt{n}} \tilde{\sigma}_n \le \mu < \overline{X}_n + \frac{\zeta_{1-\frac{\alpha}{2}}}{\sqrt{n}} \tilde{\sigma}_n\right) \xrightarrow{n\to\infty} 1 - \alpha \\ \iff &\mathbb{P}_\mu\left(\mu \in [\overline{X}_n - \frac{\zeta_{1-\frac{\alpha}{2}} \tilde{\sigma}_n}{\sqrt{n}}, \overline{X}_n + \frac{\zeta_{1-\frac{\alpha}{2}}}{\sqrt{n}} \tilde{\sigma}_n)\right) \xrightarrow{n\to\infty} 1 - \alpha. \end{aligned}$$

Under some additional assumption, we can even show that

$$\mathbb{P}_\mu\left(\mu \in [\overline{X}_n - \frac{\zeta_{1-\frac{\alpha}{2}} \tilde{\sigma}_n}{\sqrt{n}}, \overline{X}_n + \frac{\zeta_{1-\frac{\alpha}{2}}}{\sqrt{n}} \tilde{\sigma}_n]\right) \xrightarrow{n\to\infty} 1 - \alpha$$

holds. Hence when n is large enough, we have

$$\mathbb{P}_\mu\left(\mu \in \underbrace{[\overline{X}_n - \frac{\zeta_{1-\frac{\alpha}{2}} \tilde{\sigma}_n}{\sqrt{n}}, \overline{X}_n + \frac{\zeta_{1-\frac{\alpha}{2}}}{\sqrt{n}} \tilde{\sigma}_n]}_{=: I_{\alpha,n}}\right) \approx 1 - \alpha,$$

where $I_{\alpha,n}$ is called a *two-sided symmetric confidence interval* for μ with asymptotic level $1 - \alpha$, so

$$\mathbb{P}_\mu(\mu \in I_{\alpha,n}) \approx 1 - \alpha$$

for large n.

Example 4.14. Suppose that $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$ for $\lambda \in (0, \infty)$. Then we have

$$\frac{\sqrt{n}(\overline{X}_n - \lambda)}{\tilde{\sigma}_n} \xrightarrow{d} \mathcal{N}(0, 1)$$

with

$$\tilde{\sigma}_n \xrightarrow{\mathbb{P}} \sigma = \sqrt{\lambda}$$

and $\lambda = \mathbb{E}_\lambda[X_1] = \mathrm{Var}_\lambda(X_1)$. We can either take

$$\tilde{\sigma}_n = \begin{cases} \hat{\sigma}_n & \text{or} \\ S_n \end{cases}$$

or we can also take $\tilde{\sigma}_n = \sqrt{\overline{X}_n}$. Indeed, by W.L.L.N 3.55 we have

$$\begin{aligned} \overline{X}_n &\xrightarrow{\mathbb{P}} \lambda \\ \implies \sqrt{\overline{X}_n} &\xrightarrow{\mathbb{P}} \sqrt{\lambda} \end{aligned}$$

by considering $f(x) = \sqrt{x}$ which is continuous on $(0, \infty)$. Hence for a confidence interval for λ we can take either of one of the following

$$\begin{aligned} I_1 &:= \left[\overline{X}_n - \frac{\zeta_{1-\alpha/2}\hat{\sigma}_n}{\sqrt{n}}, \overline{X}_n + \frac{\zeta_{1-\alpha/2}\hat{\sigma}_n}{\sqrt{n}}\right], \\ I_2 &:= \left[\overline{X}_n - \frac{\zeta_{1-\alpha/2}S_n}{\sqrt{n}}, \overline{X}_n + \frac{\zeta_{1-\alpha/2}S_n}{\sqrt{n}}\right], \\ I_3 &:= \left[\overline{X}_n - \frac{\zeta_{1-\alpha/2}\sqrt{\overline{X}_n}}{\sqrt{n}}, \overline{X}_n + \frac{\zeta_{1-\alpha/2}\sqrt{\overline{X}_n}}{\sqrt{n}}\right]. \end{aligned}$$

5 Estimators

5.1 The Method of Moments and the Maximum Likelihood Estimators

Let $k \in \mathbb{N}$ and recall that the k-th moment of a random variable X is given by $\mathbb{E}[X^k]$ provided that X^k is integrable, meaning that $\mathbb{E}[|X^k|] < \infty$ holds. A usual notation for the k-th moment is

$$\mu_k := \mathbb{E}[X^k].$$

If the distribution of X is P_{θ_0} for some $\theta_0 \in \Theta$, then we can also write

$$\mu_k(\theta_0) := \mathbb{E}_{\theta_0}[X^k] = \int_{\mathbb{R}} x^k \, dP_{\theta_0}(x).$$

Definition 5.1. The *k-th sample* or *(empirical) moment* is defined by

$$\hat{\mu}_k := \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

with $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta_0}$. Note that $\hat{\mu}_1 = \overline{X}_n$ holds.

In the following definition, we assume that $\Theta \subseteq \mathbb{R}^d$ for a $d \in \mathbb{N}$.

Definition 5.2. The *moment estimator* $\hat{\theta}$ is a solution to the system of equations

$$\mu_k(\theta) = \hat{\mu}_k \quad \text{for } k \in \{1, \ldots, d\}$$

subject to existence.

Question. Why will this be a good estimator?
$\longrightarrow$ By the W.L.L.N 3.55, we have

$$\hat{\mu}_k \xrightarrow{\mathbb{P}} \mathbb{E}_{\theta_0}[X^k] = \mu_k(\theta_0)$$

for $k \in \{1, \ldots, d\}$ for a random variable $X \sim P_{\theta_0}$. Assume that $\hat{\mu}_k = \mu_k(\hat{\theta})$, then

$$\mu_k(\hat{\theta}) \xrightarrow{\mathbb{P}} \mu_k(\theta_0)$$

and one expects that $\hat{\theta}$ is close to θ_0 as n grows.

Examples 5.3.

- Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(m_0, \sigma_0^2)$ for $\theta_0 = (m_0, \sigma_0^2) \in \Theta = \mathbb{R} \times (0, \infty) \subseteq \mathbb{R}^2$. For $X \sim P_\theta$ for $\theta = (m, \sigma^2)$, we have

$$\begin{aligned} \mu_1(\theta) &= \mu_1(m, \sigma^2) = \mathbb{E}_\theta[X] = m \\ \mu_2(\theta) &= \mu_2(m, \sigma^2) = \mathbb{E}_\theta[X^2] = \sigma^2 + m^2. \end{aligned}$$

To obtain the moment estimator $\hat{\theta}$, we need to solve

$$\begin{cases} \mu_1(\theta) = \hat{\mu}_1 = \overline{X}_n \\ \mu_2(\theta) = \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n X_i^2, \end{cases}$$

so we arrive at

$$\begin{cases} m = \overline{X}_n \\ \sigma^2 + m^2 = \frac{1}{n}\sum_{i=1}^n X_i^2. \end{cases}$$

Hence by using

$$\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \hat{\sigma}_n^2,$$

we see that a solution is given by

$$\hat{\theta} = (\overline{X}_n, \hat{\sigma}_n^2) \xrightarrow{\text{a.s.}/\mathbb{P}} (m_0, \sigma_0^2) = \theta_0.$$

- Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \Gamma(\alpha_0, \beta_0)$ for $\alpha_0, \beta_0 > 0$. The statistical model in this case is

$$\mathcal{P} = \{P_\theta \mid \theta \in (0, \infty)^2\}$$

with

$$P_\theta(B) = \int_B f_\theta(x)\, dx$$

for

$$f_\theta(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}_{\{x>0\}}.$$

We are interested in estimating α and β. For that, we compute the 2 first moments, $\mu_1(\theta)$ and $\mu_2(\theta)$ for some $\theta = (\alpha, \beta) \in (0, \infty)^2$. We have

$$\begin{aligned} \mu_1(\theta) &= \int_{\mathbb{R}} x f_\theta(x)\, dx = \int_0^\infty x \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\, dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+1-1} e^{-\beta x}\, dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\beta^{\alpha+1}} \underbrace{\int_0^\infty \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)} x^{\alpha+1-1} e^{-\beta x}\, dx}_{=1} = \frac{\alpha}{\beta} \end{aligned}$$

and

$$\begin{aligned} \mu_2(\theta) &= \int_{\mathbb{R}} x^2 f_\theta(x)\, dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+2-1} e^{-\beta x}\, dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+2)}{\beta^{\alpha+2}} \underbrace{\int_0^\infty \frac{\beta^{\alpha+2}}{\Gamma(\alpha+2)} x^{\alpha+2-1} e^{-\beta x}\, dx}_{=1} \\ &= \frac{\alpha(\alpha+1)}{\beta^2}. \end{aligned}$$

Now the moment estimator $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ solves the following system

$$\begin{cases} \mu_1(\theta) = \frac{\alpha}{\beta} = \overline{X}_n \\ \mu_2(\theta) = \frac{\alpha(\alpha+1)}{\beta^2} = \frac{1}{n}\sum_{i=1}^n X_i^2 \end{cases}$$
$$\iff \begin{cases} \frac{\alpha}{\beta} = \overline{X}_n \\ \frac{\alpha(\alpha+1)}{\beta^2} = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \hat{\sigma}_n^2. \end{cases}$$

Hance we get

$$\hat{\beta} = \frac{\overline{X}_n}{\hat{\sigma}_n^2}$$

and

$$\hat{\alpha} = \frac{\overline{X}_n^2}{\hat{\sigma}_n^2}$$

as a solution of the system and thus the moment estimator is given by $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$.

Remark 5.4. The moment estimator can be viewed as a "plug-in" estimator. This means that we raplace a theoretical quantity by its sample/empirical/observed couterpart.

5.2 Maximum Likelihood Estimator (MLE)

Assume we observe i.i.d. random variables $X_1, \dots, X_n \sim P_{\theta_0}$ where $\theta_0 \in \Theta$. We also assume that for all $\theta \in \Theta$, P_θ admits a density p_θ with respect to a σ-finite dominating measure μ.

- In the discrete case, μ is the counting measure and $p_\theta(x) = P_\theta(\{x\})$.
- In the absolutely continuous case, μ is Lebesgue measure and $P_\theta(B) = \int_B p_\theta(x)\,dx$ for $B \in \mathcal{B}$.

If f is some real function defined on a domain Z, we will denote by $\arg\max_{z\in Z} f(z)$ the location of a maximum of f (provided that it exists).

Definition 5.5. The *likelihood function* is given by

$$L_{\mathbb{X}} : \Theta \to \mathbb{R}, \theta \mapsto \prod_{i=1}^n p_\theta(X_i)$$

where $\mathbb{X} = (X_1, \dots, X_n) \in \mathcal{X}^n$. The *maximum likelihood estimator* (*MLE*) of θ_0 based on $X_1, \dots, X_n$ is defined by

$$\hat{\theta} := \arg\max_{\theta\in\Theta} L_{\mathbb{X}}(\theta),$$

subject to existence and uniqueness.

Remark 5.6. The function $x \mapsto \log x$ is strictly increasing on $(0, \infty)$. Hence

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta \in \Theta} \log(L_{\mathbb{X}}(\theta)) \\ &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log(p_\theta(X_i)).\end{aligned}$$

The function

$$\ell_{\mathbb{X}}(\theta) := \sum_{i=1}^{n} \log(p_\theta(X_i))$$

is called the *log-likelihood function*. To find the MLE of $\hat{\theta}$, we resort often to finding the solution(s) of the equation

$$\partial_\theta \Big(\sum_{i=1}^{n} \log(p_\theta(X_i)) \Big) = 0,$$

where $s_\theta(x) := \partial_\theta \log(p_\theta(x))$ is called the *score function*.

Question. Why does the MLE work? The hope is that the MLE $\hat{\theta} \approx \theta_0$ as $n \to \infty$. Note that

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log(p_\theta(X_i))$$

looks like the "average (sample mean)" of $\log(p_\theta(X_1)), \ldots, \log(p_\theta(X_n))$. This makes us think that

$$\theta_0 \stackrel{?}{=} \arg\max_{\theta \in \Theta} \mathbb{E}_{\theta_0}[\log(p_\theta(X))]$$

with $X \sim P_{\theta_0}$. The function $x \mapsto -\log x$ is convex on $(0, \infty)$. Then, by Jensen's inequality 3.34, we have for any random variable $Y \geq 0$ such that $\mathbb{E}[|\log(Y)|] < \infty$ we have

$$\begin{aligned}&\mathbb{E}[-\log Y] \geq -\log \mathbb{E}[Y] \\ \iff &\mathbb{E}[\log Y] \leq \log \mathbb{E}[Y].\end{aligned}$$

Suppose that for any $\theta \in \Theta$ we have $\mathbb{E}_{\theta_0}[|\log(p_\theta(X))|] < \infty$ with $X \sim P_{\theta_0}$. Then

$$\begin{aligned}\mathbb{E}_{\theta_0}[\log(p_\theta(X))] - \mathbb{E}_{\theta_0}[\log(p_{\theta_0}(X))] &= \mathbb{E}_{\theta_0}\Big[\log \underbrace{\left(\frac{p_\theta(X)}{p_{\theta_0}(X)}\right)}_{=:Y}\Big] \\ &\leq \log(\mathbb{E}_{\theta_0}[Y]) = \log\left(\mathbb{E}_{\theta_0}\left[\frac{p_\theta(X)}{p_{\theta_0}(X)}\right]\right),\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}_{\theta_0}\left[\frac{p_\theta(X)}{p_{\theta_0}(X)}\right] &= \int \frac{p_\theta(x)}{p_{\theta_0}(x)} p_{\theta_0}(x) \, d\mu(x) \\ &= \int p_\theta(x) \, d\mu(x) = 1.\end{aligned}$$

This implies that for all $\theta \in \Theta$ we have

$$\mathbb{E}_{\theta_0}[\log(p_\theta(X))] \leq \mathbb{E}_{\theta_0}[\log(p_{\theta_0}(X))]$$

and thus we get

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathbb{E}_{\theta_0}[\log(p_\theta(X))].$$

Examples 5.7.

- Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$ with μ_0 unknown and suppose that σ_0 is known, so the parameter of interest is μ_0. We want to compute the MLE of μ_0. We have $\mathcal{P} = \{P_\mu \mid \mu \in \mathbb{R}\}$ and P_μ admits the density with respect to Lebesgue measure

  $$p_\mu(x) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(x-\mu)^2}{2\sigma_0^2}}.$$

  The likelihood function is

  $$L_{\mathbb{X}}(\mu) = \prod_{i=1}^n p_\mu(X_i)$$

  for $\mu \in \mathbb{R} = \Theta$. Then

  $$\begin{aligned} L_{\mathbb{X}}(\mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2\sigma_0^2}(X_i-\mu)^2} \\ &= \frac{1}{(2\pi)^{n/2}\sigma_0^n} e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\mu)^2} \\ \overset{\text{take the log}}{\Longrightarrow} \ell_{\mathbb{X}}(\mu) &= -\frac{n}{2}\log(2\pi) - n\log(\sigma_0) - \frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i - \mu)^2, \\ \ell'_{\mathbb{X}}(\mu) &= \frac{1}{\sigma_0^2}\sum_{i=1}^n (X_i - n) = \frac{n}{\sigma_0^2}(\overline{X}_n - \mu) \overset{!}{=} 0 \\ \iff \mu &= \overline{X}_n \end{aligned}$$

  and thus $\mu = \overline{X}_n$ is the unique stationary/critical point of $\ell_{\mathbb{X}}$. Furthermore, we have

  $$\ell''_{\mathbb{X}}(\mu) = -\frac{n}{\sigma_0^2} < 0$$

  and thus $\ell_{\mathbb{X}}$ is strictly concave on $\mathbb{R}$, so

  $$\hat{\mu} = \overline{X}_n$$

  is the MLE.

- Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda_0)$ for some unknown $\lambda_0 \in \Theta = (0, \infty)$. Recall that

  $$p_\lambda(x) = \lambda e^{-\lambda x} \mathbb{1}_{\{x>0\}}$$

for $x \in \mathbb{R}$. Then

$$\begin{aligned} L_{\mathbb{X}}(\lambda) &= \prod_{i=1}^{n} p_\lambda(X_i) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} \mathbb{1}_{\{X_i > 0\}} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i} \underbrace{\prod_{i=1}^{n} \mathbb{1}_{\{X_i>0\}}}_{=\mathbb{1}_{\{X_1>0,\dots,X_n>0\}}}. \end{aligned}$$

Note that because of independence, we have

$$\mathbb{P}(X_1 > 0, \dots, X_n > 0) = \prod_{i=1}^{n} \mathbb{P}(X_i > 0) = \mathbb{P}(X_1 > 0)^n = 1,$$

because

$$\begin{aligned} \mathbb{P}(X_1 > 0) &= \int_{\mathbb{R}} \mathbb{1}_{\{x>0\}} \underbrace{f_{X_1}(x)}_{=p_{\lambda_0}(x)} \, dx \\ &= \int_0^\infty \lambda_0 e^{-\lambda_0 x} \, dx = 1. \end{aligned}$$

with f_{X_1} is the density of the distribution of X_1. This implies that

$$L_{\mathbb{X}}(\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}$$

holds a.s. Furthermore, we have

$$\begin{aligned} \ell_{\mathbb{X}}(\lambda) &= n \log(\lambda) - \lambda \sum_{i=1}^{n} X_i, \\ \ell'_{\mathbb{X}}(\lambda) = 0 &\iff \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0 \\ &\iff \lambda = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}_n} \\ \ell''_{\mathbb{X}}(\lambda) &= -\frac{n}{\lambda^2} < 0 \end{aligned}$$

and thus $\ell_{\mathbb{X}}$ is strictly concave on $(0, \infty)$, so the MLE is given by

$$\hat{\lambda} = \frac{1}{\overline{X}_n}.$$

- Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$, where μ_0 and σ_0 are both unknown. This means that

$$\mathcal{P} = \{P_\theta \mid \theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)\}$$

and P_θ has density

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

Then

$$\begin{aligned}
L_{\mathbb{X}}(\theta) &= \prod_{i=1}^n p_\theta(X_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(X_i-\mu)^2} \\
&= \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2}, \\
\ell_{\mathbb{X}}(\theta) &= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2.
\end{aligned}$$

The goal is to find the maximizer of $\ell_{\mathbb{X}}$ on $\Theta = \mathbb{R} \times (0, \infty)$. For this, let us fix $\sigma \in (0, \infty)$ and consider the function

$$f_\sigma(\mu) = \ell_{\mathbb{X}}(\mu, \sigma)$$

and let us maximize f_σ over $\mathbb{R}$. Since σ is fixed, we have

$$\begin{aligned}
f'_\sigma(\mu) &= \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu) = \frac{n}{\sigma^2}(\overline{X}_n - \mu) \overset{!}{=} 0 \\
\iff \; \mu &= \overline{X}_n.
\end{aligned}$$

and since again $f''_\sigma < 0$, we see that $\overline{X}_n$ is the maximizer of f_σ over $\mathbb{R}$. We conclude that

$$\ell_{\mathbb{X}}(\mu, \sigma^2) \leq \ell_{\mathbb{X}}(\overline{X}_n, \sigma^2)$$

for any $\sigma \in (0, \infty)$. Now put

$$g(\sigma) = \ell_{\mathbb{X}}(\overline{X}_n, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \overline{X}_n)^2$$

and let us maximize g on $(0, \infty)$. We have

$$g'(\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^n (X_i - \overline{X}_n)^2$$

and recall that

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$$

was the sample variance. Hence

$$\begin{aligned} g'(\sigma) &= -\frac{n}{\sigma} + \frac{n\hat{\sigma}_n^2}{\sigma^3} \\ &= \frac{n}{\sigma^3}(\hat{\sigma}_n^2 - \sigma^2) \stackrel{!}{=} 0 \\ \iff \sigma^2 &= \hat{\sigma}_n^2 \\ \iff \sigma = \hat{\sigma}_n &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}. \end{aligned}$$

Furthermore,

$$g''(\sigma) = \frac{n}{\sigma^2} - \frac{3n}{\sigma^4}\hat{\sigma}_n^2 = \frac{n}{\sigma^4}(\sigma^2 - 3\hat{\sigma}_n^2)$$

and

$$g''(\hat{\sigma}_n) = \frac{n}{\hat{\sigma}_n^4}(\hat{\sigma}_n^2 - 3\hat{\sigma}_n^2) = -\frac{2n}{\hat{\sigma}_n^2} < 0.$$

Hence $\hat{\sigma}_n$ is a local maximizer of g. But since $\hat{\sigma}_n$ was the only (unique) stationary point we find, this implies that $\hat{\sigma}_n$ has to be a global maximizer of g. Hence we conclude that

$$g(\sigma) \leq g(\hat{\sigma}_n)$$

holds for any $\sigma \in (0, \infty)$ and thus in total we get

$$\ell_{\mathbb{X}}(\mu, \sigma^2) \leq \ell_{\mathbb{X}}(\overline{X}_n, \hat{\sigma}_n^2)$$

for all $(\mu, \sigma^2) \in \Theta$. Thus the MLE is given by

$$\hat{\theta} = (\overline{X}_n, \hat{\sigma}_n^2).$$

6 Hypothesis Testing

Let $X_1, \dots, X_n$ be i.i.d. random variables with distribution P_θ, for some unknown $\theta \in \Theta$. To simplify the notation, we will write X to denote $(X_1, \dots, X_n) = \mathbb{X} = \mathbb{X}_n$.

PROBLEM. Let Θ_0 and Θ_1 be subsets of Θ with $\Theta_0 \cap \Theta_1 = \emptyset$. We want to decide between the two statements

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1$$

based on the observation X. This is called a *testing problem*.

- "$\theta \in \Theta_0$" is called the *null hypothesis* H_0.
- "$\theta \in \Theta_1$" is called the *alternative hypothesis* H_1.

Example 6.1. Suppose that $X \sim \text{Bin}(20, \theta)$ is the observed data for $\theta \in (0,1)$. Consider the testing problem

$$H_0 : \theta = \frac{1}{2} \qquad \text{versus} \qquad H_1 : \theta = \frac{3}{4}.$$

Suppose that $X = 14$ holds. Then we have

$$\mathbb{P}_{H_0}(X = 14) = \mathbb{P}_{1/2}(X = 14) = \binom{20}{14} 2^{-20} \approx 0.036$$

$$\mathbb{P}_{H_1}(X = 14) = \mathbb{P}_{3/4}(X = 14) = \binom{20}{14} \left(\frac{3}{4}\right)^{14} \left(\frac{1}{4}\right)^{20-14} \approx 0.168.$$

We now look at the ratio

$$\frac{\mathbb{P}_{H_1}(X = 14)}{\mathbb{P}_{H_0}(X = 14)} \approx 4.56.$$

QUESTION. Is 4.56 "big enough" to decide for H_1?

Definition 6.2. In any testing problem, we can describe the situation as follows:

Decision \ Truth	H_0	H_1
Reject H_0	Error of Type I	✓
Accept H_0	✓	Error of Type II

- *Error of Type I*: The error of rejecting H_0 while it is true.

- *Error of Type II*: The error of accepting H_0 while H_1 is true.

Definition 6.3. Consider the testing problem

$$H_0 : \ \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \ \theta \in \Theta_1.$$

A *(non-randomized) statistical test* at some given *level* $\alpha \in (0,1)$ is a measurable map

$$\Phi : \mathcal{X}^n \to \{0,1\}$$

such that

$$\Phi(x) = \begin{cases} 1 & \text{means that } H_0 \text{ is rejected} \\ 0 & \text{means that } H_0 \text{ is accepted} \end{cases}$$

and

$$\sup_{\theta_0 \in \Theta_0} \mathbb{P}_{\theta_0}(\Phi(X) = 1) \leq \alpha.$$

For $\theta_1 \in \Theta_1$ the quantity $\beta(\theta_1) := \mathbb{P}_{\theta_1}(\Phi(X) = 1)$ is called the *power* of the test Φ at θ_1.

Remark 6.4. Notice that

$$\begin{aligned} 1 - \beta(\theta_1) &= 1 - \mathbb{P}_{\theta_1}(\Phi(X) = 1) \\ &= \mathbb{P}_{\theta_1}(\Phi(X) = 0) \\ &= \text{Error of the } 2^{\text{nd}} \text{ kind at } \theta_1. \end{aligned}$$

Example 6.5 (Statistical test)**.** Let $X \sim \text{Bin}(20, \theta)$ with $\theta \in \Theta = (0,1)$ and consider the testing problem

$$H_0 : \ \theta \leq \frac{1}{2} \qquad \text{versus} \qquad H_1 : \ \theta > \frac{1}{2},$$

so here we have $\Theta_0 = (0, \frac{1}{2}]$ and $\Theta_1 = (\frac{1}{2}, 1)$. Set $\alpha := 0.05$ and consider the non-randomized test

$$\Phi(X) := \begin{cases} 1 & \text{if } X > c \\ 0 & \text{if } X \leq c \end{cases}$$

for some $c \in \mathbb{R}$ satisfying

$$\begin{aligned} &\sup_{\theta_0 \leq 1/2} \mathbb{P}_{\theta_0}(\Phi(X) = 1) \leq \alpha \\ \iff &\sup_{\theta \leq 1/2} \mathbb{P}_{\theta_0}(X > c) \leq \alpha. \end{aligned} \tag{1}$$

We can show that the function

$$\theta \mapsto \mathbb{P}_\theta(X > c) = \sum_{k=c+1}^{20} \binom{20}{k} \theta^k (1-\theta)^{n-k}$$

is non-decreasing on Θ. This then implies that

$$\sup_{\theta_0 \le 1/2} \mathbb{P}_{\theta_0}(X > c) = \mathbb{P}_{\theta_0 = 1/2}(X > c)$$

and thus c must satisfy

$$\begin{aligned}\mathbb{P}_{\theta_0=1/2}(X > c) \le \alpha \iff & \sum_{k=c+1}^{20} \binom{20}{k} \left(\frac{1}{2}\right)^{20} \le \alpha \\ \iff & \mathbb{P}_{\theta_0=1/2}(X \le c) \ge 1 - \alpha \\ \iff & F_{\theta_0=1/2}(c) \ge 1 - \alpha = 0.95,\end{aligned}$$

where $F_{\theta_0=1/2}$ is the CDF of $X \sim \mathrm{Bin}(20, \frac{1}{2})$. We have

$$F_{\theta_0=1/2}(13) \approx 0.942 < 0.95 < 0.979 \approx F_{\theta_0=1/2}(14)$$

and thus $c = 14$ is the first c such that (1) holds. Note that $c = 14$ is the 0.95-quantile of the distribution of $\mathrm{Bin}(20, \frac{1}{2})$. So the test is given by $\Phi(X) = \mathbb{1}_{\{X>14\}}$ and we have

$$\sup_{\theta \le 1/2} \mathbb{P}_{\theta_0}(\underbrace{X > 14}_{\text{reject } H_0}) = \mathbb{P}_{\theta_0=1/2}(X > 14) = 1 - F_{\theta_0=1/2}(14) \approx 0.02 < 0.05.$$

We can compute following values:

θ_1	0.6	0.75	0.85
$\beta(\theta_1)$	0.125	0.617	0.932

6.1 Randomized Tests

We still consider the testing problem

$$H_0 : \ \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \ \theta \in \Theta_1.$$

Definition 6.6. A *randomized statistical test* at *level* $\alpha \in (0, 1)$ is a measurable map

$$\Phi : \mathcal{X}^n \to [0, 1]$$

with

$$\Phi(X) = \begin{cases} 1 & \text{means that } H_0 \text{ is rejected} \\ q & \text{means that } H_0 \text{ is rejected with probability } q \\ 0 & \text{means that } H_0 \text{ is accepted} \end{cases}$$

and

$$\sup_{\theta_0 \in \Theta_0} \mathbb{E}_{\theta_0}[\Phi(X)] \le \alpha.$$

For $\theta_1 \in \Theta_1$ the quantity $\beta(\theta_1) := \mathbb{E}_{\theta_1}[\Phi(X)]$ is called the *power* of Φ at θ_1.

Remarks 6.7.

- Note that $\Phi(X)$ is always equal to the probability of rejecting H_0.
- If $\Phi(X) = q$, then this means that we toss a q-coin to decide whether we reject H_0 or not.

Examples 6.8 (Randomized statistical test)**.** Consider again $X \sim \mathrm{Bin}(20, \theta)$ and the testing problem

$$H_0 : \theta \leq \frac{1}{2} \qquad \text{versus} \qquad H_1 : \theta > \frac{1}{2}.$$

We have seen $\mathbb{P}_{\theta_0=1/2}(X > 14) \approx 0.02 < 0.05$, which means that there is room for the test to be less conservative. This motivates us to consider the randomized test

$$\Phi(X) = \begin{cases} 1 & \text{if } X > 14 \\ q & \text{if } X = 14 \\ 0 & \text{if } X < 14 \end{cases}$$

with $q \in [0, 1]$ such that

$$\sup_{\theta_0 \leq 1/2} \mathbb{E}_{\theta_0}[\Phi(X)] = \alpha.$$

We are going to admit that

$$\sup_{\theta_0 \leq 1/2} \mathbb{E}_{\theta_0}[\Phi(X)] = \mathbb{E}_{\theta_0=1/2}[\Phi(X)].$$

Then, q must satisfy

$$\begin{aligned}\mathbb{E}_{\theta_0}[\Phi(X)] &= 1 \cdot \mathbb{P}_{\theta_0=1/2}(\Phi(X)=1) + q \cdot \mathbb{P}_{\theta_0=1/2}(\Phi(X)=q) \\ &\quad + \ 0 \cdot \mathbb{P}_{\theta_0=1/2}(\Phi(X)=0) \\ &= \mathbb{P}_{\theta_0=1/2}(X>14) + q\mathbb{P}_{\theta_0=1/2}(X=14) = 0.5.\end{aligned}$$

This means that

$$q = \frac{0.05 - \mathbb{P}_{\theta_0=1/2}(X>14)}{\mathbb{P}_{\theta_0=1/2}(X=14)} \approx 0.79$$

and thus the randomized test is given by

$$\Phi(X) = \begin{cases} 1 & \text{if } X > 14 \\ 0.79 & \text{if } X = 14 \\ 0 & \text{if } X < 14 \end{cases}$$

and the error of type I is exactly equal to α. We can compute the following values:

θ_1	0.6	0.75	0.85
$\beta(\theta_1)$	0.224	0.75	0.968

Observe that this test is now “more powerful” than the test in Example 6.5.

6.2 The Neyman-Pearson Test

Definition 6.9. A hypothesis H_0 is said to be *simple*, if the corresponding parameter subspace contains only one element, i.e. $\Theta_0 = \{\theta_0\}$. If $|\Theta_0| > 1$, then H_0 is said to be *composite*.

In the following, we will consider testing a simple H_0 versus a simple alternative H_1. In general, if p is the (unknown) density of $X \in \mathcal{X}^n$ with respect to some σ-finite dominating measure μ and if $p \in \{p_0, p_1\}$ for some known densities p_0 and p_1, we can consider the testing problem

$$H_0: \ p = p_0 \qquad \text{versus} \qquad H_1: \ p = p_1. \tag{1}$$

This formulation can be put in the previous context by writing

$$p = (1-\theta)p_0 + \theta p_1$$

for $\theta \in \{0,1\}$, $\Theta_0 = \{0\}$ and $\Theta_1 = \{1\}$. Then (1) is equivalent to

$$H_0: \ \theta = 0 \qquad \text{versus} \qquad H_1: \ \theta = 1. \tag{2}$$

Definition 6.10. A *Neyman-Pearson test* at *level* $\alpha \in (0,1)$ for the testing problem (1) is a randomized test of the form

$$\Phi_{\mathrm{NP}}(X) = \begin{cases} 1 & \text{if } \frac{p_1(X)}{p_0(X)} > k_\alpha \\ q_\alpha & \text{if } \frac{p_1(X)}{p_0(X)} = k_\alpha \\ 0 & \text{if } \frac{p_1(X)}{p_0(X)} < k_\alpha \end{cases}$$

with $k_\alpha > 0$ and $q_\alpha \in [0,1]$ such that

$$\mathbb{E}_{p_0}[\Phi_{\mathrm{NP}}(X)] = \alpha.$$

NP-lemma 6.11. *Let $\alpha \in (0,1)$ and k_α, q_α be such that*

$$\mathbb{E}_{p_0}[\Phi_{NP}(X)] = \alpha$$

holds. Then for any other test $\tilde{\Phi}$ such that

$$\mathbb{E}_{p_0}[\tilde{\Phi}(X)] \leq \alpha$$

we have

$$\mathbb{E}_{p_1}[\Phi_{NP}(X)] \geq \mathbb{E}_{p_1}[\tilde{\Phi}(X)].$$

Remark 6.12. We say that Φ_{NP} is *uniformly most powerful* (in short *UMP*).

Proof. We first show that

$$I := \int_{\mathcal{X}^n} \underbrace{(\Phi_{\mathrm{NP}}(x) - \tilde{\Phi}(x))(p_1(x) - k_\alpha p_0(x))}_{=: f(x)} \, d\mu(x) \geq 0$$

holds, where μ is the σ-finite dominating measure of the problem (i.e. either the counting measure or Lebesgue measure). Observe that

$$\begin{aligned} I = & \int_{\{x \mid p_1(x) > k_\alpha p_0(x)\}} f(x)\, d\mu(x) + \int_{\{x \mid p_1(x) < k_\alpha p_0(x)\}} f(x)\, d\mu(x) \\ & \quad + \int_{\{x \mid p_1(x) = k_\alpha p_0(x)\}} \underbrace{f(x)}_{=0} \, d\mu(x) \geq 0. \\ = & \int_{\{x \mid p_1(x) > k_\alpha p_0(x)\}} \underbrace{(1 - \tilde{\Phi}(x))}_{\geq 0} \underbrace{(p_1(x) - k_\alpha p_0(x))}_{>0} \, d\mu(x) \\ + & \int_{\{x \mid p_1(x) < k_\alpha p_0(x)\}} \underbrace{(0 - \tilde{\Phi}(x))}_{\leq 0} \underbrace{(p_1(x) - k_\alpha p_0(x))}_{<0} \, d\mu(x) \geq 0 \end{aligned}$$

This means that

$$\begin{aligned}
\int_{\mathcal{X}^n} (\Phi_{\mathrm{NP}}(x) - \tilde{\Phi}(x)) p_1(x)\, d\mu(x) &\geq k_\alpha \int_{\mathcal{X}^n} (\Phi_{\mathrm{NP}}(x) - \tilde{\Phi}(x)) p_0(x)\, d\mu(x) \\
\iff \mathbb{E}_{p_1}[\Phi_{\mathrm{NP}}(x) - \tilde{\Phi}(x)] &\geq k_\alpha \mathbb{E}_{p_0}[\Phi_{\mathrm{NP}}(x) - \tilde{\Phi}(x)] \\
\iff \mathbb{E}_{p_1}[\Phi_{\mathrm{NP}}(x)] - \mathbb{E}_{p_1}[\tilde{\Phi}(x)] &\geq k_\alpha (\mathbb{E}_{p_0}[\Phi_{\mathrm{NP}}(x)] - \mathbb{E}_{p_0}[\tilde{\Phi}(x)]) \\
&= \underbrace{k_\alpha}_{>0} (\underbrace{\alpha - \mathbb{E}_{p_0}[\tilde{\Phi}(x)]}_{\geq 0}) \geq 0
\end{aligned}$$

and thus we get

$$\mathbb{E}_{p_1}[\Phi_{\mathrm{NP}}(x)] \geq \mathbb{E}_{p_1}[\tilde{\Phi}(x)]$$

as claimed. □

Remark 6.13. What are k_α and q_α?
$\longrightarrow$ It can be shown that k_α can be always taken to be equal to the $(1-\alpha)$-quantile of the distribution of $Y = \frac{p_1(X)}{p_0(X)}$ under H_0 (so $p = p_0$). This means that if we denote by F_0 the CDF of Y under $X \sim p_0$ then

$$k_\alpha = \inf\{y \in \mathbb{R} \mid F_0(y) \geq 1 - \alpha\}$$

holds. On the other hand, we know that q_α satisfies

$$\begin{aligned}
\mathbb{P}_{p_0}(Y > k_\alpha) + q_\alpha \mathbb{P}_{p_0}(Y = k_\alpha) &= \alpha \\
\iff 1 - F_0(k_\alpha) + q_\alpha (F_0(k_\alpha) - F_0(k_{\alpha-})) &= \alpha,
\end{aligned}$$

where $F_0(k_{\alpha-}) = \lim_{y \to k_{\alpha-}} F_0(y)$. Hence

$$q_\alpha = \begin{cases} \frac{\alpha - (1 - F_0(k_\alpha))}{F_0(k_\alpha) - F_0(k_{\alpha-})} & \text{if } F_0(k_\alpha) > F_0(k_{\alpha-}) \\ 0 & \text{if } F_0(k_\alpha) = F_0(k_{\alpha-}), \end{cases}$$

so the value of q_α depends on whether F_0 is continuous at k_α or has a jump.

Example 6.14. Let $X \sim \mathrm{Bin}(n, \theta)$ with $n \in \mathbb{N}$ and $\theta \in (0,1)$. We want to test

$$H_0:\ \theta = \theta_0 \qquad \text{versus} \qquad H_1:\ \theta = \theta_1$$

for $\theta_1 > \theta_0$ using the NP-test. Note that

$$p_\theta(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

is the density of $X \sim \mathrm{Bin}(n, \theta)$ with respect to the counting measure and define $p_0 = p_{\theta_0}$ and $p_1 := p_{\theta_1}$. Then

$$\frac{p_1(x)}{p_0(x)} = \frac{\theta_1^x (1-\theta_1)^{n-x}}{\theta_0^x (1-\theta_0)^{n-x}} = \Big(\underbrace{\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}}_{>1}\Big)^x \Big(\frac{1-\theta_1}{1-\theta_0}\Big)^n =: g(x)$$

and g is strictly increasing and bijective. This means that the NP-test can be rewritten as

$$\Phi_{\mathrm{NP}}(X) = \begin{cases} 1 & \text{if } X > c_\alpha \\ q_\alpha & \text{if } X = c_\alpha \\ 0 & \text{if } X < c_\alpha \end{cases}$$

where c_α is the $(1-\alpha)$-quantile of the distribution of X under H_0, so c_α is the $(1-\alpha)$ quantile of $\mathrm{Bin}(n, \theta_0)$ and q_α as in the remark.

Example 6.15. Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ (so $\mathcal{X} = \mathbb{R}$) where $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma_0^2)$ with $\mu \in \mathbb{R}$ unknown and $\sigma_0 > 0$ is known. Consider the testing problem

$$H_0 : \ \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \ \mu = \mu_1$$

with $\mu_0 \neq \mu_1$. We want to determine the NP-test of level α. For $\mu \in \mathbb{R}$, we have

$$\begin{aligned} p_\mu(x_1, \ldots, x_n) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(x_i - \mu)^2 \right) \\ &= \frac{1}{\sqrt{2\pi}^n \sigma_0^n} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2 \right) \end{aligned}$$

and thus

$$\begin{aligned} \frac{p_1(x_1, \ldots, x_n)}{p_0(x_1, \ldots, x_n)} &= \exp\left(\frac{1}{2\sigma_0^2} \Big(\sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \mu_1)^2 \Big) \right) \\ &\overset{(1)}{=} \exp\left(\frac{n(\mu_1 - \mu_0)}{\sigma_0^2} (\overline{x}_n - \mu_0) - \frac{n(\mu_1 - \mu_0)^n}{2\sigma_0^2} \right) \end{aligned}$$

where (1) follows by inserting $-\mu_0 + \mu_0$ into the second sum and computing it. Hence

$$\begin{aligned} & \frac{p_1(x_1, \ldots, x_n)}{p_0(x_1, \ldots, x_n)} > \text{“something”} \\ \iff & (\mu_1 - \mu_0)(\overline{x}_n - \mu_0) > \text{“something”}. \end{aligned}$$

For the *right-sided* testing problem $\mu_1 > \mu_0$ this means that

$$\overline{x}_n - \mu_0 > \text{“something”} \iff \overline{x}_n > \text{“something”}$$

since here $\mu_1 - \mu_0 > 0$. Then, the NP-test is given by

$$\Phi_{\mathrm{NP}}(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if } \overline{X}_n > c_\alpha \\ q_\alpha & \text{if } \overline{X}_n = c_\alpha \\ 0 & \text{if } \overline{X}_n < c_\alpha \end{cases}$$

with $c_\alpha \in \mathbb{R}$ and $q_\alpha \in [0,1]$ such that

$$\mathbb{E}_{\mu_0}[\Phi_{\mathrm{NP}}(X_1,\ldots,X_n)] = \alpha.$$

In the following, we will use the fact that if $X_1,\ldots,X_n$ are independent random variables such that $X_i \sim \mathcal{N}(\mu_i,\sigma_i^2)$ then

$$\sum_{i=1}^n a_i X_i \sim \mathcal{N}(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2)$$

holds for every $a_i \in \mathbb{R}$. In particular, if $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu,\sigma_0^2)$ then

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}\left(\mu, \tfrac{\sigma_0^2}{n}\right)$$

holds which implies that

$$\mathbb{P}_{\mu_0}(\overline{X}_n = c_\alpha) = 0$$

because of the continuity of the normal distribution. Hence

$$\Phi_{\mathrm{NP}}(X_1,\ldots,X_n) = \begin{cases} 1 & \text{if } \overline{X}_n > c_\alpha \\ 0 & \text{otherwise} \end{cases}$$

where c_α is the $(1-\alpha)$-quantile of the distribution of $\overline{X}_n$ under H_0. Note that by using

$$\overline{X}_n > c_\alpha \iff \underbrace{\frac{\overline{X}_n - \mu_0}{\sqrt{\sigma_0^2/n}}}_{\sim\mathcal{N}(0,1)} > \frac{c_\alpha - \mu_0}{\sqrt{\sigma_0^2/n}}$$

we get

$$\begin{aligned} &\mathbb{P}_{\mu_0}(\overline{X}_n > c_\alpha) = \alpha \\ \iff &\mathbb{P}_{\mu_0}\left(\tfrac{\overline{X}_n - \mu_0}{\sqrt{\sigma_0^2/n}} > \tfrac{c_\alpha - \mu_0}{\sqrt{\sigma_0^2/n}}\right) = \alpha \\ \iff &\mathbb{P}\left(Z > \tfrac{c_\alpha - \mu_0}{\sqrt{\sigma_0^2/n}}\right) = \alpha \\ \iff &1 - \mathbb{P}\left(Z \le \tfrac{c_\alpha - \mu_0}{\sqrt{\sigma_0^2/n}}\right) = 1-\alpha \end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$, so $\frac{c_\alpha - \mu_0}{\sqrt{\sigma_0^2/n}} = \zeta_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\mathcal{N}(0,1)$. Using this approach, we also get

$$\Phi_{\mathrm{NP}}(X_1,\ldots,X_n) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma_0} > \zeta_{1-\alpha} \\ 0 & \text{otherwise.} \end{cases}$$

For the *left-sided* testing problem, we assume that $\mu_0 > \mu_1$. Recall that the NP-test is based on the ratio

$$\begin{aligned}\frac{p_{\mu_1}(x_1,\dots,x_n)}{p_{\mu_0}(x_1,\dots,x_n)} &= \exp\left(\frac{n(\mu_1-\mu_0)}{\sigma_0^2}(\overline{x}_n-\mu_0)\right)\exp\left(\frac{-n}{2\sigma_0^2}(\mu_1-\mu_0)^2\right) > \text{“something”}\\ &\iff \overline{x}_n < \text{“something”}.\end{aligned}$$

Using similar arguments as for the right-sided problem, we can show that the NP-test of level α is given by

$$\Phi_{\mathrm{NP}}(X_1,\dots,X_n) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma_0} < \zeta_\alpha \\ 0 & \text{otherwise.}\end{cases}$$

Now consider the *two-sided* testing problem

$$H_0:\ \mu=\mu_0 \qquad \text{versus} \qquad H_1:\ \mu\neq\mu_0,$$

so $\Theta_1 = \mathbb{R}\setminus\{\mu_0\}$ which gives rise to the name. Note that here we cannot apply the NP-test, because H_1 is not simple. However, we can show that the following test

$$\Phi(X_1,\dots,X_n) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}|\overline{X}_n-\mu_0|}{\sigma_0} > \zeta_{1-\alpha/2} \\ 0 & \text{otherwise}\end{cases}$$

has good properties and is of level α. Let us show that it is indeed of level α.

Proof. We need to show that

$$\mathbb{E}_{\mu_0}[\Phi(X_1,\dots,X_n)] \le \alpha$$

holds or equivalently

$$\mathbb{P}_{\mu_0}\left(\sqrt{n}\frac{|\overline{X}_n-\mu_0|}{\sigma_0} > \zeta_{1-\alpha/2}\right) \le \alpha$$

under H_0, so $\overline{X}_n \sim \mathcal{N}(\mu_0,\sigma_0^2)$ and we also have

$$\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma_0} \sim \mathcal{N}(0,1).$$

Thus we get

$$\begin{aligned}\mathbb{P}_{\mu_0}\left(\sqrt{n}\frac{|\overline{X}_n-\mu_0|}{\sigma_0} > \zeta_{1-\alpha/2}\right) &= \mathbb{P}(|Z| > \zeta_{1-\alpha/2})\\ &= \mathbb{P}(Z > \zeta_{1-\alpha/2} \text{ or } Z < -\zeta_{1-\alpha/2})\\ &= \mathbb{P}(Z > \zeta_{1-\alpha/2}) + \mathbb{P}(Z < -\zeta_{1-\alpha/2})\\ &= 2\mathbb{P}(Z > \zeta_{1-\alpha/2})\\ &= 2(1-(1-\tfrac{\alpha}{2})) = \alpha\end{aligned}$$

for $Z \sim \mathcal{N}(0,1)$ by using the symmetry of the normal distribution. $\square$

7 One Sample Tests

SETTING. We observe i.i.d. random variables $X_1, \dots, X_n$ whose distribution results from shifting some "baseline" distribution by some amount $\theta \in \Theta$. This θ is called the *shift* or *location parameter*.

Examples 7.1.

- Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$ for $\theta \in \mathbb{R}$ and $\sigma \in (0, \infty)$, where σ can be known or unknown. The "baseline" distribution is $\mathcal{N}(0, \sigma^2)$ and the location parameter is

$$\theta = \underbrace{\mathbb{E}_\theta[X_1]}_{\text{expectation}} = \underbrace{F_\theta^{-1}(\tfrac{1}{2})}_{\text{median}}$$

  with F_θ the CDF of $\mathcal{N}(\theta, \sigma^2)$. Note that here the expectation and the median are equal because $\mathcal{N}(\theta, \sigma^2)$ is symmetric around θ (generally this does not hold).

- Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{U}(\theta, \theta + 1)$. The "baseline" distribution is $\mathcal{U}(0, 1)$ and the location parameter θ is

$$\theta = \mathbb{E}_\theta[X_1] - \tfrac{1}{2} = F_\theta^{-1}(\tfrac{1}{2}) - \tfrac{1}{2}.$$

We can consider the following testing problems:

- *Right-sided* given by

$$\begin{array}{lllll} & H_0 : \theta = \theta_0 & \text{versus} & H_1 : \theta > \theta_0 \\ \text{or} & H_0 : \theta \leq \theta_0 & \text{versus} & H_1 : \theta > \theta_0. \end{array}$$

- *Left-sided* given by

$$\begin{array}{lllll} & H_0 : \theta = \theta_0 & \text{versus} & H_1 : \theta < \theta_0 \\ \text{or} & H_0 : \theta \geq \theta_0 & \text{versus} & H_1 : \theta < \theta_0. \end{array}$$

- *Two-sided* given by

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0.$$

7.1 The Student's Test

We assume that $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$. For $\gamma \in (0, 1)$ set

$$\zeta_\gamma = \gamma\text{-quantile of } \mathcal{N}(0, 1)$$

and let us first assume that $\sigma = \sigma_0$ is known.

- Consider the (simplified) right-sided testing problem

$$H_0:\ \theta=\theta_0 \qquad \text{versus} \qquad H_1:\ \theta=\theta_1$$

  with $\theta_1>\theta_0$. We know that the NP-test

$$\Phi_{\mathrm{NP}}(X_1,\ldots,X_n)=\mathbb{1}_{\left\{\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sigma_0}>\zeta_{1-\alpha}\right\}}$$

  is UMP of level α by the NP-lemma 6.11.

- For the (simplified) right-sided testing problem

$$H_0:\ \theta=\theta_0 \qquad \text{versus} \qquad H_1:\ \theta=\theta_1$$

  with $\theta_0>\theta_1$ we know that the NP-test

$$\Phi_{\mathrm{NP}}(X_1,\ldots,X_n)=\mathbb{1}_{\left\{\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sigma_0}<\zeta_{\alpha}\right\}}$$

  is UMP of level α, again by the NP-lemma 6.11.

- For the two-sided testing problem

$$H_0:\ \theta=\theta_0 \qquad \text{versus} \qquad H_1:\ \theta\neq\theta_0$$

  the test

$$\Phi(X_1,\ldots,X_n)=\mathbb{1}_{\left\{\frac{\sqrt{n}|\overline{X}_n-\theta_0|}{\sigma_0}>\zeta_{1-\alpha/2}\right\}}$$

  is of level α and has some "good" properties.

Note that if σ is unknown, the previous tests cannot be used. In a way, we need to estimate σ. In this case, σ is called a "nuisance parameter".

Definition 7.2 (The student distribution)**.** A random variable Y is said to have a *Student distribution* if

$$Y=\frac{Z}{\sqrt{X/m}}$$

with $Z\sim\mathcal{N}(0,1)$ and $X\sim\chi^2_{(m)}=$ "Chi-square distribution with m degrees of freedom", that is

$$X=X_1^2+X_2^2+\ldots+X_m^2,$$

where $X_1,\ldots,X_m\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,1)$ and Z and X are independent. The Student distribution is also called the t-distribution and denoted by

$$Y\sim\mathcal{T}_{(m)}.$$

Remark 7.3. It can be shown that the Student distribution with m degrees of freedom is absolutely continuous with density

$$f(x) = \frac{\Gamma(\frac{m+1}{2})}{\sqrt{\pi m}\,\Gamma_{(m)}} \frac{1}{(1+\frac{x^2}{m})^{\frac{m+1}{2}}}$$

for $x \in \mathbb{R}$. Note that f is symmetric around 0. If $m = 1$, then

$$\mathcal{T}_{(1)} = \text{ Cauchy distribution.}$$

Theorem 7.4. *Let* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$. *Then we have*

$$\frac{\sqrt{n}(\overline{X}_n - \theta)}{S_n} \sim \mathcal{T}_{(n-1)},$$

where

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

The Student's test. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$ and for simplicity, consider the (simpler) testing problem

$$H_0:\ \theta = \theta_0 \qquad \text{versus} \qquad H_1:\ \theta = \theta_1$$

with $\theta_1 > \theta_0$ and σ is unknown. Consider the following test

$$\Phi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}(\overline{X}_n - \theta_0)}{S_n} > t_{n-1,1-\alpha} \\ 0 & \text{otherwise,} \end{cases}$$

where

$$t_{n-1,1-\alpha} := (1-\alpha)\text{-quantile of } \mathcal{T}_{(n-1)}.$$

The test defined above is of level α. Indeed, we have

$$\begin{aligned} \mathbb{E}_{\theta_0}[\Phi(X_1, \ldots, X_n)] &= \mathbb{P}_{\theta_0}\left(\tfrac{\sqrt{n}(\overline{X}_n - \theta_0)}{S_n} > t_{n-1,1-\alpha}\right) \\ &= \mathbb{P}(T_{n-1} > t_{n-1,1-\alpha}) = 1 - (1-\alpha) = \alpha \end{aligned}$$

by Theorem 7.4 for $T_{n-1} \sim \mathcal{T}_{(n-1)}$.

Since the test Φ does not involve the particular value of θ_1, it can be used again for testing

$$H_0:\ \theta = \theta_0 \qquad \text{versus} \qquad H_1:\ \theta > \theta_0.$$

Now let $\theta < \theta_0$ and compute

$$\begin{aligned}
\mathbb{E}_\theta[\Phi(X_1,\dots,X_n)] &= \mathbb{P}_\theta\left(\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{S_n} > t_{n-1,1-\alpha}\right) \\
&= \mathbb{P}_\theta\left(\frac{\sqrt{n}(\overline{X}_n-\theta+\theta-\theta_0)}{S_n} > t_{n-1,1-\alpha}\right) \\
&= \mathbb{P}_\theta\left(\frac{\sqrt{n}(\overline{X}_n-\theta)}{S_n} - \underbrace{\frac{\sqrt{n}(\theta_0-\theta)}{S_n}}_{\geq 0} > t_{n-1,1-\alpha}\right) \\
&\leq \mathbb{P}_\theta\left(\frac{\sqrt{n}(\overline{X}_n-\theta)}{S_n} > t_{n-1,1-\alpha}\right).
\end{aligned}$$

This means that

$$\begin{aligned}
\mathbb{P}_\theta\left(\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{S_n} > t_{n-1,1-\alpha}\right) &\leq 1-\mathbb{P}_\theta\left(\underbrace{\frac{\sqrt{n}(\overline{X}_n-\theta)}{S_n}}_{\sim\mathcal{T}_{(n-1)}} \leq t_{n-1,1-\alpha}\right) \\
&= 1-(1-\alpha) = \alpha
\end{aligned}$$

and the calculation holds for every $\theta \leq \theta_0$. Hence we get

$$\sup_{\theta\leq\theta_0} \mathbb{P}_\theta\left(\tfrac{\sqrt{n}(\overline{X}_n-\theta_0)}{S_n} > t_{n-1,1-\alpha}\right) = \sup_{\theta\in\Theta_0} \mathbb{E}_\theta[\Phi(X_1,\dots,X_n)] \leq \alpha$$

with $\Theta_0 = (-\infty,\theta_0]$, so Φ has level α for the testing problem

$$H_0:\ \theta \leq \theta_0 \qquad \text{versus} \qquad H_1:\ \theta > \theta_0.$$

For the analoguous left-sided testing problem

$$H_0:\ \theta \geq \theta_0 \qquad \text{versus} \qquad H_1:\ \theta < \theta_0$$

we can show that the test

$$\Phi(X_1,\dots,X_n) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}(\overline{X}_n-\theta_0)}{S_n} < t_{n-1,\alpha} \\ 0 & \text{otherwise} \end{cases}$$

is of level α, meaning that

$$\sup_{\theta\geq\theta_0} \mathbb{E}_\theta[\Phi(X_1,\dots,X_n)] \leq \alpha,$$

where

$$t_{n-1,\alpha} := \alpha\text{-quantile of } \mathcal{T}_{(n-1)}.$$

Now consider the two-sided testing problem

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0.$$

Here, the test

$$\Phi(X_1, \dots, X_n) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}|\overline{X}_n - \theta_0|}{S_n} > t_{n-1,1-\alpha/2} \\ 0 & \text{otherwise} \end{cases}$$

is of level α. Indeed, put

$$T_{n-1} := \frac{\sqrt{n}(\overline{X}_n - \theta_0)}{S_n} \overset{\text{under } H_0}{\sim} \mathcal{T}_{(n-1)}$$

and compute

$$\begin{aligned} \mathbb{P}_{\theta_0}(|T_{n-1}| > t_{n-1,1-\alpha/2}) &\overset{(1)}{=} \mathbb{P}(|T_{n-1}| > t_{n-1,1-\alpha/2}) \\ &= \mathbb{P}(T_{n-1} > t_{n-1,1-\alpha/2}) + \mathbb{P}(T_{n-1} < -t_{n-1,1-\alpha/2}) \\ &= 2 \cdot \mathbb{P}(T_{n-1} > t_{n-1,1-\alpha/2}) \\ &\overset{(2)}{=} 2(1 - (1 - \tfrac{\alpha}{2})) = \alpha. \end{aligned}$$

where (1) works because $\mathcal{T}_{(n-1)}$ does not depend on θ_0 and at (2) we used the symmetry of the student's distribution.

Example 7.5. We observe the following values sampled from five i.i.d. random variables with distribution $\mathcal{N}(\theta, \sigma^2)$:

$$0.926,\ 0.513,\ 1.272,\ 1.359,\ -0.038$$

We want to know whether $\theta = 0$ is a plausible assumption. Formally, we want to test

$$H_0 : \theta = 0 \qquad \text{versus} \qquad H_1 : \theta \neq 0.$$

Since σ is unknown, we may use the two-sided student test in this case for $n = 5$. We take $\alpha = 0.05$, put

$$T_4 := \frac{\sqrt{5}\,\overline{X}_5}{S_5}$$

and we know that

$$t_{4,0.975} = 2.776$$

is the 0.975-quantile of $\mathcal{T}_{(4)}$. Using this, we find that

$$\begin{gathered} \overline{X}_5 = 0.806,\ S_5 = 0.577 \\ \Longrightarrow |T_4| = T_4 = 3.121 > t_{4,0.975} \end{gathered}$$

and thus we reject H_0 at the level $\alpha = 0.05$. Let us now take $\alpha = 0.01$. The only thing that changes in the test is the quantile of $\mathcal{T}_{(4)}$, so we compute

$$t_{4,0.995} = 4.604 > T_4.$$

Hence at this level we cannot reject H_0.

7.2 The Sign Test

Let $X_1, \ldots, X_n$ be i.i.d. random variables from an unknown distribution. Let m denote the (unknown) median of distribution, so

$$m = F^{-1}(\tfrac{1}{2})$$

where F is the CDF of the distribution. Consider the testing problem

$$H_0 : \; m = m_0 \qquad \text{versus} \qquad H_1 : \; m \neq m_0$$

for some fixed value m_0. We assume that the CDF F is continuous at m. This means that

$$\mathbb{P}(X_i < m) = \underbrace{\mathbb{P}(X_i \leq m)}_{=F(m)} = \frac{1}{2}$$

holds for every $i \in \{1, \ldots, n\}$. Consider the statistic

$$T_n = |\{i \mid X_i > m_0\}| = \sum_{i=1}^{n} \mathbb{1}_{\{X_i > m_0\}} \overset{\text{under } H_0}{\sim} \text{Bin}(n, \tfrac{1}{2}).$$

We want to reject H_0 if $|T_n - \frac{n}{2}|$ is "too big". Consider the test

$$\Phi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if } |T_n - \frac{n}{2}| > c_\alpha \\ 0 & \text{otherwise,} \end{cases}$$

where c_α is chosen such that

$$\mathbb{E}_{m_0}[\Phi(X_1, \ldots, X_n)] = \mathbb{P}_{m_0}(|T_n - \tfrac{n}{2}| > c_\alpha) \leq \alpha.$$

Now using

$$\{|T_n - \tfrac{n}{2}| > c_\alpha\} = \{T_n > \tfrac{n}{2} + c_\alpha\} \sqcup \{T_n < \tfrac{n}{2} - c_\alpha\}$$

we get

$$\begin{aligned} \mathbb{P}_{m_0}(|T_n - \tfrac{n}{2}| > c_\alpha) &= \mathbb{P}_{m_0}(T_n > \tfrac{n}{2} + c_\alpha) + \mathbb{P}_{m_0}(T_n < \tfrac{n}{2} - c_\alpha) \\ &= \mathbb{P}_{m_0}(T_n > \tfrac{n}{2} + c_\alpha) + \mathbb{P}_{m_0}(n - T_n > \tfrac{n}{2} + c_\alpha). \end{aligned}$$

Now observe that

$$\begin{aligned} n - T_n = n - \sum_{i=1}^{n} \mathbb{1}_{\{X_i > m_0\}} &= \sum_{i=1}^{n} (1 - \mathbb{1}_{\{X_i > m_0\}}) \\ &= \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq m_0\}} \overset{\text{under } H_0}{\sim} \text{Bin}(n, \tfrac{1}{2}). \end{aligned}$$

This implies that

$$\begin{aligned}&\mathbb{P}_{m_0}(|T_n - \tfrac{n}{2}| > c_\alpha) = 2 \cdot \mathbb{P}_{m_0}(T_n > \tfrac{n}{2} + c_\alpha) \leq \alpha \\ \iff &\mathbb{P}_{m_0}(T_n > \tfrac{n}{2} + c_\alpha) \leq \tfrac{\alpha}{2}\end{aligned}$$

and thus we take c_α such that

$$\tfrac{n}{2} + c_\alpha = (1 - \tfrac{\alpha}{2})\text{-quantile of } \mathrm{Bin}(n, \tfrac{1}{2}).$$

Example 7.6. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$. We want to test

$$H_0 : \ \theta = 0 \qquad \text{versus} \qquad H_1 : \ \theta \neq 0.$$

Note that here θ is the expectation and also the median. This means that we can use one of the following tests:

- The Student test

$$\Phi_1(X_1, \dots, X_n) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}\,|\overline{X}_n|}{S_n} > t_{n-1,1-\alpha/2} \\ 0 & \text{otherwise.} \end{cases}$$

- The sign test

$$\Phi_2(X_1, \dots, X_n) = \begin{cases} 1 & \text{if } |T_n - \frac{n}{2}| > c_\alpha \\ 0 & \text{otherwise,} \end{cases}$$

  where T_n and c_α are as above.

Note that the Student test uses some knowledge about the distribution while the sign test does not, so we may expect the first test to be better (i.e. to have a higher power).

7.3 Two Sample Tests

Setting. We observe $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_1, \sigma^2)$ and $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_2, \sigma^2)$ for $\theta_1, \theta_2 \in \mathbb{R}$ and $\sigma \in (0, \infty)$ such that $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ are independent. We want to test

$$H_0 : \ \theta_1 = \theta_2 \qquad \text{versus} \qquad H_1 : \ \theta_1 \neq \theta_2.$$

Remark 7.7 (Some facts about the Gaussian distribution).

(1) For any random variable Z we have $Z \sim \mathcal{N}(\theta, \sigma^2)$ if any only if for all $t \in \mathbb{R}$ we have $\mathbb{E}[e^{itZ}] = e^{it\theta - \frac{1}{2}t^2\sigma^2}$. Here $\mathbb{E}[e^{itz}]$ is called the *characteristic function*.

(2) $\mathbf{Z} = (Z_1, \dots, Z_k)^T \in \mathbb{R}^k$ for some $k \in \mathbb{N}$ is a Gaussian vector with expectation $\boldsymbol{\theta}$ and covariance matrix Σ, so $Z \sim \mathcal{N}(\boldsymbol{\theta}, \Sigma)$, if and only if

$$\mathbb{E}[e^{it^T\mathbf{Z}}] = e^{it^T\boldsymbol{\theta} - \frac{1}{2}t^T\Sigma t}$$

holds for all $t \in \mathbb{R}^k$. Here $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T$ with $\theta_i = \mathbb{E}[Z_i]$ and

$$\Sigma_{ij} = \operatorname{cov}(Z_i, Z_j) = \mathbb{E}[(Z_i - \theta_i)(Z_j - \theta_j)] = \mathbb{E}[Z_i Z_j] - \theta_i \theta_j.$$

(3) If $Z_1, \dots, Z_k$ are independent random variables such that $Z_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$ for $i \in \{1, \dots, k\}$, then $\mathbf{Z} = (Z_1, \dots, Z_k)^T$ is a Gaussian vector with parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ and $\Sigma = \operatorname{diag}(\sigma_1^2, \dots, \sigma_k^2) \in \mathbb{R}^{k \times k}$. Let $t = (t_1, \dots, t_k) \in \mathbb{R}^k$. Then we have

$$\begin{aligned} \mathbb{E}[e^{it^T \mathbf{Z}}] &= \mathbb{E}[e^{i \sum_{j=1}^k t_j Z_j}] = \mathbb{E}\left[\prod_{j=1}^k e^{it_j Z_j}\right] \\ &\overset{(1)}{=} \prod_{j=1}^k \mathbb{E}[e^{it_j Z_j}] = \prod_{j=1}^k e^{it_j \theta_j - \frac{1}{2} t_j^2 \sigma_j^2} \\ &= e^{i \sum_{j=1}^k t_j \theta_j - \frac{1}{2} \sum_{j=1}^k t_j^2 \sigma_j^2} \\ &= e^{it^T \boldsymbol{\theta} - \frac{1}{2} t^T \Sigma t}, \end{aligned}$$

where at (1) we used independence.

(4) If $\mathbf{Z} \sim \mathcal{N}(0, I_k) \in \mathbb{R}^n$, then for any $\theta \in \mathbb{R}^k$ and $A \in \mathbb{R}^{k \times n}$ we have

$$\mathbf{Y} := \underbrace{\theta + A\mathbf{Z}}_{\in \mathbb{R}^k} \sim \mathcal{N}(\theta, AA^T).$$

(5) Any linear combination of the components of a Gaussian vector is a Gaussian random variable. Indeed, let $\mathbf{Z} = (Z_1, \dots, Z_k)^T \sim \mathcal{N}(\theta, \Sigma)$ and $a_1, \dots, a_k \in \mathbb{R}$. Put $X := \sum_{j=1}^k a_j Z_j$ and $a = (a_1, \dots, a_k)^T$, then

$$X = a^T \mathbf{Z}$$

holds. For $t \in \mathbb{R}$ put $\alpha := ta$ and observe that

$$\begin{aligned} \mathbb{E}[e^{itX}] &= \mathbb{E}[e^{ita^T \mathbf{Z}}] = \mathbb{E}[e^{i\alpha^T \mathbf{Z}}] \\ &= e^{i\alpha^T \theta - \frac{1}{2} \alpha^T \Sigma \alpha} \\ &= e^{ia^T \theta t - \frac{1}{2} a^T \Sigma a t^2}, \end{aligned}$$

so $X \sim \mathcal{N}(a^T \theta, a^T \Sigma a)$.

(6) If $\mathbf{Z} = (Z_1, \dots, Z_k)^T$ is a Gaussian vector, then $Z_1, \dots, Z_k$ are independent if any only if

$$\operatorname{cov}(Z_i, Z_j) = 0 \tag{1}$$

holds for all $1 \leq i < j \leq k$. Indeed, observe the following.

"$\Longrightarrow$". If $Z_1, \ldots, Z_k$ are independent, then Z_i and Z_j are independent for any fixed $1 \leq i < j \leq k$ and thus (1) holds.

"$\Longleftarrow$". Suppose that $\Sigma_{ij} = \mathrm{cov}(Z_i, Z_j) = 0$ holds for all $1 \leq i < j \leq k$. Then we have

$$\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_k^2)$$

with $\sigma_i^2 = \mathrm{cov}(Z_i, Z_i) = \mathrm{var}(Z_i)$. Let $\theta = \mathbb{E}[\mathbf{Z}]$ and $t \in \mathbb{R}^k$. Then

$$\begin{aligned}\mathbb{E}[e^{it^T\mathbf{Z}}] &= e^{it^T\theta - \frac{1}{2}t^T\Sigma t} = e^{i\sum_{j=1}^k t_j\theta_j - \frac{1}{2}\sum_{j=1}^k t_j^2\sigma_j^2} \\ &= \prod_{j=1}^k e^{it_j\theta_j - \frac{1}{2}t_j^2\sigma_j^2} = \prod_{j=1}^k \mathbb{E}[e^{it_jZ_j}]\end{aligned}$$

which is equivalent to

$$\mathbb{E}\left[\prod_{j=1}^k e^{it_jZ_j}\right] = \prod_{j=1}^k \mathbb{E}[e^{it_jZ_j}],$$

so $Z_1, \ldots, Z_k$ are independent.

(7) If $\mathbf{Z} = (Z_1, \ldots, Z_k)^T$ and $\mathbf{W} = (W_1, \ldots, W_m)^T$ are such that $(\mathbf{Z}, \mathbf{W})^T$ is a Gaussian vectors, then $\mathbf{Z}$ and $\mathbf{W}$ are independent if any only if

$$\mathrm{cov}(Z_i, W_j) = 0$$

holds for all $1 \leq j \leq k$ and $1 \leq j \leq m$.

Remark 7.8. Recall that if $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$ then we have

$$\frac{\sqrt{n}(\overline{X}_n - \theta)}{S_n} \sim \mathcal{T}_{(n-1)}.$$

Now note that

$$\frac{\sqrt{n}(\overline{X}_n - \theta)}{S_n} = \frac{\sqrt{n}\frac{\overline{X}_n - \theta}{\sigma}}{\frac{S_n}{\sigma}} = \frac{\sqrt{n}\frac{\overline{X}_n - \theta}{\sigma}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n \left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2}}.$$

Then $\frac{\sqrt{n}(\overline{X}_n - \theta)}{\sigma} \sim \mathcal{N}(0,1)$ is independent of $\frac{1}{n-1}\sum_{i=1}^n \left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2$ and $\sum_{i=1}^n (\frac{X_i - \overline{X}_n}{\sigma})^2 \sim \chi^2_{(n-1)}$ holds.

Proof.

- We already know that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$. From above we also know:
  - Fact 3: $\mathbf{X} = (X_1, \ldots, X_n)^T$ is a Gaussian vector.

- Fact 5: We know that $\overline{X}_n$ is a Gaussian random variable with $\overline{X}_n \sim \mathcal{N}(\mathbb{E}[\overline{X}_n], \mathrm{Var}(\overline{X}_n))$ which implies that
$$\frac{\overline{X}_n - \theta}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0,1).$$

- We now show that $\frac{\sqrt{n}(\overline{X}_n-\theta)}{\sigma}$ is independent of $\frac{1}{n-1}\sum_{i=1}^n \left(\frac{X_i-\overline{X}_n}{\sigma}\right)^2$. To show this, we show that $\overline{X}_n$ is independent of $(X_1-\overline{X}_n, \dots, X_n-\overline{X}_n)^T$. We have that $(\overline{X}_n, X_1-\overline{X}_n, \dots, X_n-\overline{X}_n)^T$ is a Gaussian vector since it is a linear transformation of $(X_1,\dots,X_n)^T$ (see fact 4 above). We also have

$$\begin{aligned}
\mathrm{cov}(\overline{X}_n, X_i-\overline{X}_n) &= \mathrm{cov}(\overline{X}_n, X_i) - \mathrm{cov}(\overline{X}_n, \overline{X}_n) \\
&= \mathrm{cov}(\overline{X}_n, X_i) - \mathrm{var}(\overline{X}_n) \\
&= \mathrm{cov}\left(\frac{1}{n}\sum_{j=1}^n X_j, X_i\right) - \frac{\sigma^2}{n} \\
&= \frac{1}{n}\mathrm{cov}(X_i, X_i) - \frac{\sigma^2}{n} \\
&= \frac{1}{n}\mathrm{var}(X_i) - \frac{\sigma^2}{n} = 0
\end{aligned}$$

which concludes the proof. □

Back to the testing problem. Recall the testing problem

$$H_0:\ \theta_1 = \theta_2 \qquad \text{versus} \qquad H_1:\ \theta_1 \neq \theta_2.$$

The vector $\mathbf{Z} = (X_1,\dots,X_n,Y_1,\dots,Y_m)^T \in \mathbb{R}^{n+m}$ is a Gaussian vector with expectation

$$\theta = (\underbrace{\theta_1,\dots,\theta_1}_{n}, \underbrace{\theta_2,\dots,\theta_2}_{m})^T \in \mathbb{R}^{n+m}$$

and covariance $\Sigma = \mathrm{diag}(\sigma^2,\dots,\sigma^2) = \sigma^2 I_{n+m}$. Then $\overline{X}_n - \overline{Y}_m$ is a Gaussian random vector because it is a linear combination of $\mathbf{Z}$. Its parameters are

$$\mathbb{E}[\overline{X}_n - \overline{Y}_m] = \mathbb{E}[\overline{X}_n] - \mathbb{E}[\overline{Y}_m] = \theta_1 - \theta_2$$

and

$$\mathrm{Var}(\overline{X}_n - \overline{Y}_m) = \mathrm{var}(\overline{X}_n) + \mathrm{var}(\overline{Y}_m) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \frac{n+m}{nm}\sigma^2,$$

hence $\overline{X}_n - \overline{Y}_m \sim \mathcal{N}\left(\theta_1-\theta_2, \frac{n+m}{nm}\sigma^2\right)$ and $\overline{X}_n - \overline{Y}_m \overset{\text{under } H_0}{\sim} \mathcal{N}\left(0, \frac{n+m}{nm}\sigma^2\right)$.

The idea. If $|\overline{X}_n - \overline{Y}_m|$ is "too big", then we reject H_0. But what is "too big"?

CASE 1. $\sigma = \sigma_0$ is known. Then under H_0 we have

$$\overline{X}_n - \overline{Y}_m \sim \mathcal{N}\left(0, \frac{n+m}{nm}\sigma_0^2\right)$$
$$\iff \sqrt{\frac{nm}{n+m}}\frac{\overline{X}_n - \overline{Y}_m}{\sigma_0} \sim \mathcal{N}(0,1).$$

The test of level α is given by

$$\Phi(X_1, \dots, X_n, Y_1, \dots, Y_m) = \begin{cases} 1 & \text{if } \sqrt{\frac{nm}{n+m}}\frac{|\overline{X}_n - \overline{Y}_m|}{\sigma_0} > \zeta_{1-\alpha/2} \\ 0 & \text{otherwise,} \end{cases}$$

where $\zeta_{1-\alpha/2}$ is the $(1-\frac{\alpha}{2})$-quantile of $\mathcal{N}(0,1)$.

CASE 2. σ is unknown. It is a "nuisance" parameter which needs to be estimated. Consider

$$S_{n,m}^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^n (X_i - \overline{X}_n)^2 + \sum_{j=1}^m (Y_j - \overline{Y}_m)^2\right)$$

and

$$T_{n,m} = \sqrt{\frac{nm}{n+m}} \cdot \frac{\overline{X}_n - \overline{Y}_m}{S_{n,m}}.$$

It can be shown that under H_0 we have

$$T_{n,m} \sim \mathcal{T}_{(n+m-2)}.$$

In this case, the test of level α is given by

$$\Phi(X_1, \dots, X_n, Y_1, \dots, Y_m) = \begin{cases} 1 & \text{if } |T_{n,m}| > t_{n+m-2,1-\alpha/2} \\ 0 & \text{otherwise,} \end{cases}$$

where $t_{n+m-2,1-\alpha/2}$ is the $\left(1-\frac{\alpha}{2}\right)$-quantile of $\mathcal{T}_{(n+m-2)}$.

Appendix

A Convergence Results for Random Variables

Definition A.1. Let $p \geq 1$ and let $(X_n)_n$ be a sequence of real-valued random variables. We say that X_n *converges in* L^p to a random variable X, denoted by $X_n \xrightarrow{L^p} X$, if

$$\mathbb{E}[|X_n - X|^p] \xrightarrow{n\to\infty} 0$$

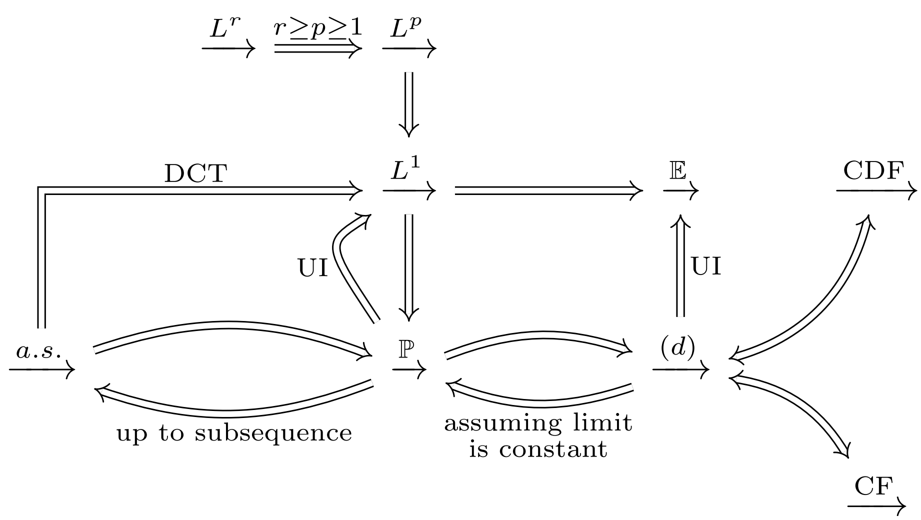
holds.

Definition A.2. A family $(X_i)_{i\in I}$ of real-valued random variables is said to be *uniformly integrable*, or *UI* in short, if

$$\sup_{i\in I} \mathbb{E}[|X_i| \mathbb{1}_{|X_i|\geq K}] \xrightarrow{K\to\infty} 0$$

holds.

If not indicated otherwise, the following implications hold for any sequence of random variables and $\xrightarrow{\mathbb{E}}$ denotes convergence of the mean, $\xrightarrow{\text{CDF}}$ denotes pointwise convergence of the CDF's at the continuity points and $\xrightarrow{\text{CF}}$ denotes pointwise convergence of the characteristic functions.



B Summary of Distributions

Definition B.1. For a random variable $X : \Omega \to \mathbb{R}$ the *support* of X is defined to be the smallest closed set $R_X \subseteq \mathbb{R}$ with $\mathbb{P}(X \in R_X) = 1$.

Definition B.2. For any random variable $X : \Omega \to \mathbb{R}$ we define its *skewness* by

$$\gamma_X := \frac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{\sigma_X^3} = \frac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{\mathbb{E}[(X - \mathbb{E}[X])^2]^{3/2}}.$$

Definition B.3. For $t, c > 0$ we define the *gamma function* by

$$\Gamma(t) := \int_0^\infty x^{t-1} e^{-x}\, dx$$

and the *lower incomplete gamma function* by

$$\mathcal{G}(t, c) := \int_0^c x^{t-1} e^{-x}\, dx.$$

B.1 Discrete Distributions

Random Variable X	Uniform $\sim \mathcal{U}(E)$	Bernoulli	Binomial	Geometric	Poisson	Negative Binomial
Parameters	E	$p \in [0,1]$	$n \in \mathbb{N}, p \in [0,1]$	$p \in (0,1]$	$\lambda > 0$	$\gamma > 0, p \in [0,1]$
Support R_X	finite set $E \subseteq \mathbb{R}$	$\{0,1\}$	$\{0,1,\ldots,n\}$	$\{0,1,2,\ldots\}$	$\{0,1,2,\ldots\}$	$\{0,1,2,\ldots\}$
PMF $\mathbb{P}(X = k),\ k \in R_X$	$\frac{1}{n}$	$(1-p)\mathbb{1}_{k=0} + p\mathbb{1}_{k=1}$	$\binom{n}{k} p^k (1-p)^{n-k}$	$(1-p)^k p$	$e^{-\lambda} \frac{\lambda^k}{k!}$	$\binom{k+\gamma-1}{k} p^k (1-p)^\gamma$
CDF $\mathbb{P}(X \le t),\ t \in R_X$	$\frac{\lvert(-\infty,t] \cap E\rvert}{n}$	$(1-p)\mathbb{1}_{t\ge0} + p\mathbb{1}_{t\ge1}$	$I_{1-p}(n-t, 1+t)$	$1 - (1-p)^{\lfloor t\rfloor+1}$	$\frac{\Gamma(\lfloor t+1\rfloor, \lambda)}{\lfloor t\rfloor!}$	$I_{1-p}(\gamma, k+1)$
Characteristic function $\varphi_X(t)$	$\frac{e^{iat} - e^{i(b+1)t}}{(b-a+1)(1-e^{it})}$, for $E = \{a,\ldots,b\}$	$1 - p + pe^{it}$	$(1-p+pe^{it})^n$	$\frac{p}{1-(1-p)e^{it}}$	$\exp\{\lambda(e^{it}-1)\}$	$\left(\frac{1-p}{1-pe^{it}}\right)^\gamma$
Expectation	$\frac{1}{n}\sum_{x\in E} x$	p	np	$\frac{1-p}{p}$	λ	$\frac{\gamma p}{1-p}$
Variance	$\frac{(b-a+1)^2-1}{12}$, for $E = \{a,\ldots,b\}$	$p(1-p)$	$np(1-p)$	$\frac{1-p}{p^2}$	λ	$\frac{\gamma p}{(1-p)^2}$
Skewness	0	$\frac{1-2p}{\sqrt{p(1-p)}}$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\frac{2-p}{\sqrt{1-p}}$	$\frac{1}{\sqrt{\lambda}}$	$\frac{1+p}{\sqrt{p\gamma}}$

B.2 Continuous Distributions

Random Variable X	Uniform	Exponential	Gaussian	Beta	Gamma	χ_k^2
Parameters	$[a,b]$	$\lambda > 0$	$\mu \in \mathbb{R}, \sigma^2 > 0$	$\alpha, \beta > 0$	$\gamma, c > 0$	$k \in \mathbb{N}$
Support R_X	$[a,b] \subseteq \mathbb{R}$	$[0,\infty)$	$\mathbb{R}$	$[0,1]$	$[0,\infty)$	$[0,\infty)$
Density $f_X(x),\ x \in R_X$	$\frac{1}{b-a}$	$\lambda e^{-\lambda x}$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}}$	$\frac{c^\gamma}{\Gamma(\gamma)} x^{\gamma-1} e^{-cx}$	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$
CDF $\mathbb{P}(X \leq t),\ t \in R_X$	$\frac{x-a}{b-a} \mathbb{1}_{x \in [a,b]} + \mathbb{1}_{x>b}$	$1 - e^{-\lambda t}$	$\Phi\left(\frac{t-\mu}{\sigma}\right)$	$I_t(\alpha, \beta)$	$\frac{\mathcal{G}(\gamma, ct)}{\Gamma(\gamma)}$	$\frac{\mathcal{G}(k/2, t/2)}{\Gamma(k/2)}$
Characteristic function $\varphi_X(t)$	$\frac{e^{itb}-e^{ita}}{t(b-a)}$, for $t \neq 0$ 1, for $t = 0$	$\frac{\lambda}{\lambda - it}$	$\exp\left\{\mu it - \frac{\sigma^2 t^2}{2}\right\}$	$F_1(\alpha; \alpha+\beta; it)$	$\left(\frac{c}{c-it}\right)^\gamma$	$(1-2it)^{-k/2}$
Expectation	$\frac{a+b}{2}$	$\frac{1}{\lambda}$	μ	$\frac{\alpha}{\alpha+\beta}$	$\frac{\gamma}{c}$	k
Variance	$\frac{(b-a)^2}{12}$	$\frac{1}{\lambda^2}$	σ^2	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\frac{\gamma}{c^2}$	$2k$
Skewness	0	2	0	$\frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$	$\frac{2}{\sqrt{\gamma}}$	$\sqrt{\frac{8}{k}}$

References

[1] H. Föllmer and H. Künsch, *Wahrscheinlichkeitsrechnung und Statistik*, Lecture notes, (2013).

[2] S. van de Geer, *Statistics for Mathematics (Stat4Math)*, Lecture notes, (2021).