Stochastics

Bachelor Applied Artificial Intelligence (AAI-B3)

André Herzwurm

TH Rosenheim, WiSe 2022/23

Literature

H.-O. Georgii, *Stochastics*, DeGruyter, 2nd Edition, 2013.

A. Steland, *Basiswissen Statistik*, Springer, 4. Auflage, 2016. (German)

S.M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*, Academic Press, 2014.

Contents

Chapter I

Introduction

We consider births in Germany. Question (two variants):

(i) What is the probability of having a girl?

(ii) Is it more likely to have a boy than a girl?

Empirical data from statistical sample yields the gender of N births:

$$k \text{ girls } (1), \quad N-k \text{ boys } (0).$$

The relative frequencies are given by

$$\widehat{p}(1) = \frac{k}{N}, \quad \widehat{p}(0) = \frac{N-k}{N} = 1 - \widehat{p}(1)$$

and serve as a metric of the data (descriptive statistics).
Naive answers to the above questions read:

(i) The sought probability is given by $\widehat{p}(1)$.

(ii) Yes if and only if $\widehat{p}(0) > \widehat{p}(1)$.

Criticism:

- The answers are based on only one sample or data set.
- The sample size and the variability of the data are not taken into account.

Therefore we will consider

inferential statistics (see Chapters IV and V)

to infer properties of an underlying probability distribution. This, in turn, requires a

mathematical model (see Chapters II and III)

of the underlying random mechanism.

Chapter II

Probability Theory – Discrete Case

In this section we aim at modelling and analyzing random experiments with at most countably infinite outcomes (discrete case).

In the sequel let Ω be a finite or a countably infinite set. The set Ω is a model for the possible results of a random experiment.

1 Discrete Probability Spaces

Definition 1. The set Ω is called *sample space*. Its elements $\omega \in \Omega$ are called *outcomes*. Any subset $A \subseteq \Omega$ of Ω is called an *event*.

Example 2. Consider a fair coin with head (1) and tail (0) that is tossed twice. The corresponding sample space is given by

$$\Omega = \{(1,1),(1,0),(0,1),(0,0)\} = \{1,0\}^2.$$

The event $A = \{(1,1),(1,0)\}$ contains all outcomes with 1 in the first toss.

Clearly, we may also use this sample space as a model for two-stage processes having two outcomes (e.g., yes/no) per stage, e.g., a two-stage production process where in each stage some tolerance level is satisfied or not.

Terminology: We say that the event $A \subseteq \Omega$ occurs if $\omega \in A$.

Notation: For sets A, B we write $A \subseteq B$ if every element of A is also an element of B, and we write $A = B$ if $A \subseteq B$ and $B \subseteq A$. The empty set containing no elements is denoted by $\varnothing$. The cardinality of a set A is denoted by $|A|$.

Definition 3. Let $A, B \subseteq \Omega$ be events. The *union*, the *intersection*, and the *difference* of A and B as well as the *complement* of A are defined by

$$\begin{aligned} A \cup B &= \{\omega \in \Omega \colon \omega \in A \text{ or } \omega \in B\}, \\ A \cap B &= \{\omega \in \Omega \colon \omega \in A \text{ and } \omega \in B\}, \\ A \setminus B &= \{\omega \in \Omega \colon \omega \in A \text{ and } \omega \notin B\}, \\ A^{\mathsf{c}} &= \Omega \setminus A = \{\omega \in \Omega \colon \omega \notin A\}, \end{aligned}$$

respectively.

Remark 4. Foe events $A, B, C \subseteq \Omega$ we have

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C), \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C), \\ (A \cap B)^{\mathsf{c}} &= A^{\mathsf{c}} \cup B^{\mathsf{c}}, \\ (A \cup B)^{\mathsf{c}} &= A^{\mathsf{c}} \cap B^{\mathsf{c}}, \\ A \setminus B &= A \cap B^{\mathsf{c}}. \end{aligned}$$

Definition 5. The *(countable) union* of events $A_1, A_2, A_3, \ldots \subseteq \Omega$ is defined by

$$\bigcup_{i=1}^{\infty} A_i = \big\{\omega \in \Omega \colon (\exists i \in \mathbb{N} \colon \omega \in A_i)\big\}.$$

The events $A_1, A_2, A_3, \ldots \subseteq \Omega$ are *pairwise disjoint* if $A_i \cap A_j = \varnothing$ for all $i, j \in \mathbb{N}$ with $i \neq j$.

Remark 6. Clearly, we can express the union of finitely many events as a countable union, i.e., for $n \in \mathbb{N}$ and events $A_1, \ldots, A_n \subseteq \Omega$ we have

$$\bigcup_{i=1}^{n} A_i = A_1 \cup \ldots \cup A_n = \bigcup_{i=1}^{\infty} A_i$$

if we put $A_i = \varnothing$ for $i > n$.

We are now going to assign probabilities $P(A)$ to events $A \subseteq \Omega$.

Definition 7. The set

$$\mathcal{P}(\Omega) = \{A \colon A \subseteq \Omega\}$$

of all subsets of Ω is called *power set* of Ω.

Example 8. The power set of $\Omega = \{0, 1\}$ is given by

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

Moreover, we have $|\mathcal{P}(\Omega)| = 2^{|\Omega|}$.

Definition 9. A function $P \colon \mathcal{P}(\Omega) \to \mathbb{R}$ is a *probability measure (on Ω)* (or *probability distribution (on Ω))* if

(i) $0 \leq P(A) \leq 1$ for all $A \subseteq \Omega$,

(ii) $P(\Omega) = 1$,

(iii) for all pairwise disjoint events $A_1, A_2, \ldots \in \mathcal{P}(\Omega)$ we have

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \qquad \textit{(σ-additivity)}$$

In such a case the triple $(\Omega, \mathcal{P}(\Omega), P)$ is called a *discrete probability space*.

Proposition 10. Let P be a probability measure on Ω and let $A, B \in \mathcal{P}(\Omega)$. Then we have

(i) $A \cap B = \varnothing \;\Rightarrow\; P(A \cup B) = P(A) + P(B)$ (*additivity*),

(ii) $A \subseteq B \;\Rightarrow\; P(B \setminus A) = P(B) - P(A)$,

(iii) $A \subseteq B \;\Rightarrow\; P(A) \leq P(B)$ (*monotonicity*),

(iv) $P(A^{\mathsf{c}}) = 1 - P(A)$,

(v) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof. ad (i): Since $A \cap B = \varnothing$, we have

$$A \cup B = A \cup B \cup \varnothing \cup \varnothing \cup \ldots$$

for pairwise disjoint events and thus the σ-additivity of P shows

$$P(A \cup B) = P(A) + P(B) + \sum_{i=3}^{\infty} P(\varnothing).$$

Since $P(A \cup B) \in [0, 1]$, we have $P(\varnothing) = 0$.
ad (ii)-(v): See `Exercise` 1.1. □

2 Probability Mass Functions

Definition 1. Let P be a probability measure on Ω. The function $p\colon \Omega \to \mathbb{R}$ defined by

$$p(\omega) = P(\{\omega\})$$

is called *probability mass function (PMF) (associated to P)*.

Proposition 2.

(i) For P and p according to Definition 1 and for every $A \subseteq \Omega$ we have

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

In particular, P is uniquely determined by p.

(ii) Every probability mass function $p\colon \Omega \to \mathbb{R}$ satisfies

$$\forall\, \omega \in \Omega\colon 0 \leq p(\omega) \leq 1 \quad \wedge \quad \sum_{\omega \in \Omega} p(\omega) = 1. \tag{1}$$

(iii) Every function $p\colon \Omega \to \mathbb{R}$ satisfying (1) defines a probability measure P on Ω by

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

for $A \subseteq \Omega$.

Proof. Since Ω is countable, every $A \subseteq \Omega$ is countable and can thus be expressed as a countable union $A = \bigcup_{\omega \in A} \{\omega\}$ of pairwise disjoint sets.
ad (i): The σ-additivity of P yields for $A \subseteq \Omega$

$$P(A) = P\left(\bigcup_{\omega \in A} \{\omega\}\right) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} p(\omega).$$

ad (ii): Use part (i) with $A = \Omega$ and note that $P(\Omega) = 1$.
ad (iii): Verify the conditions of Definition 1.9. □

Example 3. Let Ω be finite with $|\Omega| = n \in \mathbb{N}$. For $\omega \in \Omega$ we put

$$p(\omega) = \frac{1}{n}.$$

Then p satisfies (1) and the associated probability measure P is given by

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

for every $A \subseteq \Omega$. Hence the calculation of probabilities is based on counting elements.

Definition 4. The probability measure P according to Example 3 is called the *discrete uniform distribution (on the finite set Ω).*

Example 5. Consider a fair coin that is tossed twice, cf. Example 1.2. We model this random experiment by using the discrete uniform distribution P on

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

The events

$$\begin{aligned} A &= \{(1,1), (1,0)\} = \text{“1 in the first toss”}, \\ B &= \{(1,1), (0,1)\} = \text{“1 in the second toss”} \end{aligned}$$

satisfy

$$\begin{aligned} P(A) = \frac{2}{4} = \frac{1}{2} = P(B), \\ P(A \cap B) = P(\{(1,1)\}) = \frac{1}{4} = P(A) \cdot P(B), \\ P(A \cup B) = P(\{(1,1), (1,0), (0,1)\}) = \frac{3}{4}. \end{aligned}$$

Example 6. Consider a two-stage production process where in each process a tolerance level is satisfied (1) or not (0). We model this by

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

and the (fictional) probability mass function

ω	$(1,1)$	$(1,0)$	$(0,1)$	$(0,0)$
$p(\omega)$	0.8	0.09	0.01	0.1

.

The events

$$A = \{(1,1),(1,0)\} = \text{“tolerance level satisfied in first stage”},$$
$$B = \{(1,1),(0,1)\} = \text{“tolerance level satisfied in second stage”}$$

satisfy

$$P(A) = 0.89,$$
$$P(B) = 0.81,$$
$$P(A \cap B) = 0.8 \neq 0.7209 = P(A) \cdot P(B).$$

3 Conditional Probability and Independence

In the sequel let $(\Omega, \mathcal{P}(\Omega), P)$ be a discrete probability space.

Question: How do we change the probability measure P if we know that a certain event $B \subseteq \Omega$ has occured.

Definition 1. For $A, B \subseteq \Omega$ with $P(B) > 0$ the *conditional probability of A given B* is defined by

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

Example 2. Let P be the discrete uniform distribution on a finite set Ω. For $A, B \subseteq \Omega$ with $B \neq \varnothing$ we have $P(B) > 0$ and

$$P(A \,|\, B) = \frac{|A \cap B|}{|\Omega|} \cdot \frac{|\Omega|}{|B|} = \frac{|A \cap B|}{|B|}.$$

In Example 2.5 (fair coin tossed twice “independently”) we thus have

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

for $A = \{(1,1),(1,0)\}$ and $B = \{(1,1),(0,1)\}$.

Example 3. In Example 2.6 (two-stage production process) we have

$$P(A \,|\, B) = \frac{P(\{(1,1)\})}{P(\{(1,1),(0,1)\})} = \frac{0.8}{0.81} \approx 0.9876$$

and

$$P(B \,|\, A) = \frac{P(\{(1,1)\})}{P(\{(1,1),(1,0)\})} = \frac{0.8}{0.89} \approx 0.8988$$

for $A = \{(1,1),(1,0)\}$ and $B = \{(1,1),(0,1)\}$.

Remark 4. Let p be the probability mass function associated to P, and let $B \subseteq \Omega$ with $P(B) > 0$. For $A \subseteq \Omega$ we have

$$P(A \mid B) = \frac{1}{P(B)} \cdot \sum_{\omega \in A \cap B} p(\omega) = \sum_{\omega \in A} q(\omega),$$

where

$$q(\omega) = \begin{cases} \dfrac{p(\omega)}{P(B)}, & \text{if } \omega \in B, \\ 0, & \text{else.} \end{cases}$$

Then $Q \colon \mathcal{P}(\Omega) \to \mathbb{R}$ given by

$$Q(A) = P(A \mid B)$$

defines a probability measure on Ω with probability mass function q.

Definition 5. Two events $A, B \subseteq \Omega$ are *independent* if

$$P(A \cap B) = P(A) \cdot P(B).$$

Remark 6. Let $P(B) > 0$. Then we have

$$A, B \text{ independent } \Leftrightarrow \; P(A \mid B) = P(A).$$

Example 7. In Example 2.5 (fair coin tossed twice "independently") the events A and B are independent. In Example 2.6 (two-stage production process) the events A and B are dependent.

Proposition 8 (Bayes's law)**.** Let $n \in \mathbb{N}$ and $\{B_1, \ldots, B_n\} \subseteq \mathcal{P}(\Omega)$ be a partition[1] of Ω with $P(B_i) > 0$ for all $i = 1, \ldots, n$.

(i) For all $A \subseteq \Omega$ we have

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i).$$

(ii) For all $A \subseteq \Omega$ with $P(A) > 0$ and for all $k = 1, \ldots, n$ we have

$$P(B_k \mid A) = \frac{P(A \mid B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)}.$$

Proof. ad (i): Since $\{B_1, \ldots, B_n\}$ is a partition, we obtain for every $A \subseteq \Omega$

$$A = A \cap \left(\bigcup_{i=1}^{n} B_i \right) = \bigcup_{i=1}^{n} (A \cap B_i)$$

with pairwise disjoint sets $A \cap B_1, \ldots, A \cap B_n \subseteq \Omega$. Hence we get

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i).$$

[1] A partition of Ω is a family of non-empty pairwise disjoint sets $B_1, B_2, \ldots \subseteq \Omega$ with $\bigcup_{i=1}^{n} B_i = \Omega$.

Since $P(B_i) > 0$ for all $i = 1, \dots, n$, the conditional probabilities $P(A \mid B_i)$ are well-defined such that $P(A \cap B_i) = P(A \mid B_i) \cdot P(B_i)$.
ad (ii): If $P(A) > 0$, the conditional probability $P(B_k \mid A)$ is well-defined and we have

$$P(B_k \mid A) \cdot P(A) = P(B_k \cap A) = P(A \cap B_k) = P(A \mid B_k) \cdot P(B_k)$$

for all $k = 1, \dots, n$. Use part (i). □

4 Random Variables

In the sequel let $(\Omega, \mathcal{P}(\Omega), P)$ be a discrete probability space.
Question: How can we describe particular aspects of a random experiment (rather than the full experiment)?

Example 1. Consider a fair die that is rolled twice independently, i.e., we consider a discrete uniform distribution P on $\Omega = \{1, \dots, 6\}^2$. Particular aspects could be:

(i) "number of pips in first die roll",

(ii) "number of pips in second die roll",

(iii) "sum of pips".

In the sequel let $\mathfrak{X} \subseteq \mathbb{R}$ be finite or countably infinite. We will typically consider $\mathfrak{X} \subseteq \mathbb{N}_0$.

Definition 2. A function $X \colon \Omega \to \mathfrak{X}$ is called a *random variable (with values in $\mathfrak{X}$)*. Its function values $x = X(\omega) \in \mathfrak{X}$ are called *realizations* of X.

Example 3 (Continuation of Example 1)**.** The first two aspects are described by $\mathfrak{X} = \{1, \dots, 6\}$ and the random variables $X_1, X_2 \colon \Omega \to \mathfrak{X}$ given by

$$X_1(\omega) = \omega_1, \quad X_2(\omega) = \omega_2$$

for $\omega = (\omega_1, \omega_2) \in \Omega$. The third aspect is described by $\mathfrak{S} = \{2, \dots, 12\}$ and the random variable $S \colon \Omega \to \mathfrak{S}$ given by

$$S(\omega) = X_1(\omega) + X_2(\omega) = \omega_1 + \omega_2.$$

In the sequel let $X \colon \Omega \to \mathfrak{X}$ be a random variable. In many cases one is just interested in the probabilities

$$P_X(A) = P(\{\omega \in \Omega \colon X(\omega) \in A\})$$

for $A \subseteq \mathfrak{X}$ and in particular in

$$p_X(x) = P_X(\{x\}) = P(\{\omega \in \Omega : X(\omega) = x\})$$

for $x \in \mathfrak{X}$.

Example 4 (Continuation of Example 3)**.** For $x \in \{1, \dots, 6\}$ we have

$$p_{X_1}(x) = p_{X_2}(x) = \frac{1}{6},$$

and p_S is given by

s	2	3	4	5	6	7	8	9	10	11	12
$p_S(s)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Proposition 5. The following holds true:

(i) $p_X \colon \mathfrak{X} \to \mathbb{R}$ is a probability mass function,

(ii) P_X is the associated probability measure on $\mathfrak{X}$, i.e., for all $A \subseteq \mathfrak{X}$ we have

$$P_X(A) = \sum_{x \in A} p_X(x).$$

Definition 6. P_X and p_X are called *distribution* and *probability mass function of* X, respectively.

Example 7 (Continuation of Example 4)**.** A *rod graph* can be used to illustrate the probability mass functions p_{X_1} and p_S, see Figure 4.1.

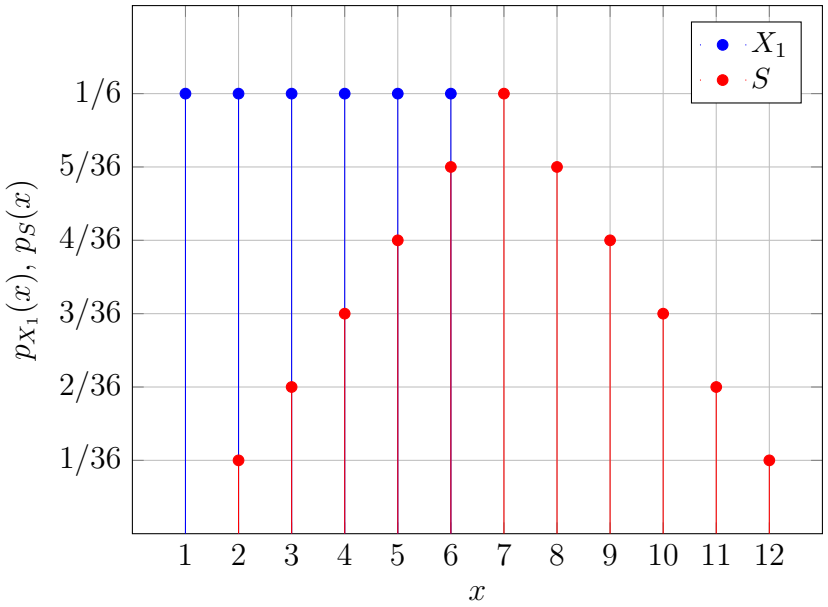


Figure 4.1: Probability mass functions p_{X_1} and p_S from Example 4.

Definition 8. Two random variables $X_1, X_2 \colon \Omega \to \mathfrak{X}$ are *identically distributed* if

$$P_{X_1}(A) = P_{X_2}(A)$$

for all $A \subseteq \mathfrak{X}$.

Example 9 (Continuation of Example 4)**.** P_{X_1} and P_{X_2} are both the discrete uniform distribution on $\{1, \ldots, 6\}$. Hence X_1 and X_2 are identically distributed.

Proposition 10 (Criterion for identical distributions)**.** $X_1, X_2 \colon \Omega \to \mathfrak{X}$ are identically distributed if and only if

$$p_{X_1}(x) = p_{X_2}(x)$$

for all $x \in \mathfrak{X}$.

Proof. Use Proposition 5. □

Remark 11. The case of a multivariate random variable (random vector)

$$X = (X_1, \ldots, X_n) \colon \Omega \to \mathfrak{X}$$

with an at most countably infinite set $\mathfrak{X} \subseteq \mathbb{R}^n$ is treated analogously. The components X_i of X are random variables.

Example 12 (Continuation of Example 3)**.**

(i) For $\mathfrak{X} = \{1, \ldots, 6\}^2$ and $x = (x_1, x_2) \in \mathfrak{X}$ we have

$$p_{(X_1,X_2)}(x) = P(\{\omega \in \Omega \colon \omega = x\}) = P(\{x\}) = \frac{1}{36}.$$

Hence $P_{(X_1,X_2)}$ is the discrete uniform distribution on $\mathfrak{X}$.

(ii) For $\mathfrak{X} = \{(x_1, s) \in \mathbb{N}^2 \colon 1 \leq x_1 \leq 6 \wedge x_1 + 1 \leq s \leq x_1 + 6\}$ and $(x_1, s) \in \mathfrak{X}$ we have

$$p_{(X_1,S)}(x_1, s) = P(\{\omega \in \Omega \colon \omega_1 = x_1 \wedge \omega_1 + \omega_2 = s\}) = \frac{1}{36}.$$

Hence $P_{(X_1,S)}$ is the discrete uniform distribution on $\mathfrak{X}$.

In the sequel we use

$$\{X = x\} = \{\omega \in \Omega \colon X(\omega) = x\},$$
$$\{X \in A\} = \{\omega \in \Omega \colon X(\omega) \in A\},$$

and we consider random variables $X_1, \ldots, X_n$ on $(\Omega, \mathcal{P}(\Omega), P)$ taking values in an at most countably infinite set $\mathfrak{X}$.

Definition 13. $X_1, \ldots, X_n$ are *independent* if for all $A_1, \ldots, A_n \subseteq \mathfrak{X}$ we have

$$P\left(\bigcap_{i=1}^{n} \{X_i \in A_i\}\right) = \prod_{i=1}^{n} P(\{X_i \in A_i\}).$$

Example 14 (Continuation of Example 3)**.**

(i) For $\mathfrak{X} = \{1, \dots, 6\}$ and $A_1, A_2 \subseteq \mathfrak{X}$ we have

$$\begin{aligned} P(\{X_1 \in A_1\} \cap \{X_2 \in A_2\}) &= P(\{\omega \in \Omega \colon \omega_1 \in A_1 \wedge \omega_2 \in A_2\}) \\ &= \frac{|A_1| \cdot |A_2|}{|\Omega|} = \frac{|A_1| \cdot |A_2|}{36} \end{aligned}$$

and according to Example 9 for $i = 1, 2$

$$P(\{X_i \in A_i\}) = \frac{|A_i|}{6}.$$

Hence X_1 and X_2 are independent.

(ii) For $\mathfrak{X} = \{1, \dots, 12\}$, $A_1 = \{6\}$ and $B = \{2\}$ we have

$$\begin{aligned} P(\{X_1 \in A_1\} \cap \{S \in B\}) &= P(\{\omega \in \Omega \colon \omega_1 = 6 \wedge \omega_1 + \omega_2 = 2\}) \\ &= P(\varnothing) = 0 \end{aligned}$$

as well as $P(\{X_1 \in A_1\}) > 0$ and $P(\{S \in B\}) > 0$. Hence X_1 and S are not independent.

Proposition 15 (Criterion for independence)**.** $X_1, \dots, X_n$ are independent if and only if for all $x_1, \dots, x_n \in \mathfrak{X}$ we have

$$P\left(\bigcap_{i=1}^{n} \{X_i = x_i\}\right) = \prod_{i=1}^{n} p_{X_i}(x_i). \tag{2}$$

Proof. "⇒": Take $A_i = \{x_i\}$.
"⇐": Since $\mathfrak{X}$ is countable, every $A_i \subseteq \mathfrak{X}$ can be expressed as a countable union $A_i = \bigcup_{x_i \in A_i} \{x_i\}$ of pairwise disjoint sets. Use the σ-additivity of P and (2). [...] □

Remark 16. Consider the special case of $n = 2$ and $\mathfrak{X} = \{0, \dots, k\}$ for some $k \in \mathbb{N}$. Put

$$p_{i,j} = P(\{X_1 = i\} \cap \{X_2 = j\})$$

and

$$p_{i,\bullet} = P(\{X_1 = i\}) \quad \text{and} \quad p_{\bullet,j} = P(\{X_2 = j\})$$

for $i, j \in \{0, \dots, k\}$. Clearly, we have

$$p_{i,\bullet} = \sum_{j=0}^{k} p_{i,j} \quad \text{and} \quad p_{\bullet,j} = \sum_{i=0}^{k} p_{i,j}$$

for all $i, j \in \{0, \dots, k\}$. The corresponding *contingency table* is given by Table II.1. Proposition 15 shows that

$$X_1, X_2 \text{ independent } \Leftrightarrow \forall\, i, j \in \{0, \dots, k\} \colon p_{i,j} = p_{i,\bullet} \cdot p_{\bullet,j}.$$

X_1 \ X_2	0	$\cdots$	k	$\sum$
0	$p_{0,0}$	$\cdots$	$p_{0,k}$	$p_{0,\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
k	$p_{k,0}$	$\cdots$	$p_{k,k}$	$p_{k,\bullet}$
$\sum$	$p_{\bullet,0}$	$\cdots$	$p_{\bullet,k}$	1

Table II.1: Contingency table from Remark 16.

5 Empirical Distribution

In the sequel let $(\Omega, \mathcal{P}(\Omega), P)$ be a discrete probability space and let $X\colon \Omega \to \mathfrak{X}$ be a random variable with a countable set $\mathfrak{X} \subseteq \mathbb{R}$.

We consider the case where the set $\mathfrak{X}$ of possible realizations is known but the distribution P_X is unknown. Instead, a *sample* (data set)

$$x_1, \ldots, x_N \in \mathfrak{X} \tag{3}$$

consisting of realizations of N independent repetitions of the random experiments is available.

Question: Can we approximately determine the probability mass function p_X of X?

Definition 1. Given the sample (3) the *relative frequency* of $x \in \mathfrak{X}$ is given by

$$\begin{aligned}\widehat{p}(x) &= \frac{|\{\ell \in \{1, \ldots, N\}\colon x_\ell = x\}|}{N} \\ &= \frac{\text{number of elements of the sample equal to } x}{N}.\end{aligned}$$

Remark 2. For $p = \widehat{p}$ we have (1) in Proposition 2.2(ii). According to Proposition 2.2(iii), see also Proposition 4.5, we obtain a probability distribution $\widehat{P}$ on $\mathfrak{X}$ satisfying

$$\begin{aligned}\widehat{P}(A) = \sum_{x \in A} \widehat{p}(x) &= \frac{|\{\ell \in \{1, \ldots, N\}\colon x_\ell \in A\}|}{N} \\ &= \frac{\text{number of elements of the sample with values in } A}{N}\end{aligned}$$

for $A \subseteq \mathfrak{X}$.

Definition 3. $\widehat{P}$ is called *empirical distribution* of the sample (3).

Remark 4. A rod graph can be used to illustrate empirical distributions in terms of relative frequencies. Based on the scale of measure we distinguish the following types:

(i) Nominal scale: Elements of $\mathfrak{X}$ indicate a name (operations $=, \neq$). Example: gender (male (0), female (1)).

X_1 \ X_2	0	$\cdots$	k	$\sum$
0	$\widehat{p}_{0,0}$	$\cdots$	$\widehat{p}_{0,k}$	$\widehat{p}_{0,\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
k	$\widehat{p}_{k,0}$	$\cdots$	$\widehat{p}_{k,k}$	$\widehat{p}_{k,\bullet}$
$\sum$	$\widehat{p}_{\bullet,0}$	$\cdots$	$\widehat{p}_{\bullet,k}$	1

Table II.2: Contingency table for empirical distribution from Remark 6.

(ii) Ordinal scale: Elements of $\mathfrak{X}$ allow for a ranking (operations $<, >, =$). Example: rank order.

(iii) Metric (cardinal) scale: Elements of $\mathfrak{X}$ are numeric and allow for arithmetic operations. Examples: time, temperature, sales.

Remark 5. The case of a multivariate random variable (random vector)

$$X = (X_1, \ldots, X_n) \colon \Omega \to \mathfrak{X}$$

with an at most countably infinite set $\mathfrak{X} \subseteq \mathbb{R}^n$ is treated analogously.

Remark 6 (Counterpart of Remark 4.16)**.** Consider the special case of two random variables $(X_1, X_2) \colon \Omega \to \mathfrak{X}$ and $\mathfrak{X} = \{0, \ldots, k\}^2$ with $k \in \mathbb{N}$. Put

$$\widehat{p}_{i,j} = \frac{|\{\ell \in \{1, \ldots, N\} \colon x_\ell = (i,j)\}|}{N}$$

and

$$\widehat{p}_{i,\bullet} = \frac{|\{\ell \in \{1, \ldots, N\} \colon x_{\ell,1} = i\}|}{N}, \quad \widehat{p}_{\bullet,j} = \frac{|\{\ell \in \{1, \ldots, N\} \colon x_{\ell,2} = j\}|}{N}$$

for $i, j \in \{0, \ldots, k\}$. We clearly have

$$\widehat{p}_{i,\bullet} = \sum_{j=0}^{k} \widehat{p}_{i,j} \quad \text{and} \quad \widehat{p}_{\bullet,j} = \sum_{i=0}^{k} \widehat{p}_{i,j}$$

for all $i, j \in \{0, \ldots, k\}$. The corresponding contingency table is given by Table II.2. The random variables X_1 and X_2 are assumed to be independent if and only if

$$\widehat{p}_{i,j} \approx \widehat{p}_{i\,\bullet} \cdot \widehat{p}_{\bullet,j}$$

for all $i, j \in \{0, \ldots, k\}$.

Example 7. The relative frequencies of tossing two fair coins, see Example 2.5 (one coin tossed twice independently), resulting from a computer simulation with sample size $N = 10^3$ is shown in Figure 5.1. The corresponding contingency table for

$$X_1(\omega) = \omega_1, \quad X_2(\omega) = \omega_2$$

with $\omega = (\omega_1, \omega_2) \in \Omega$ is shown in Table II.3.

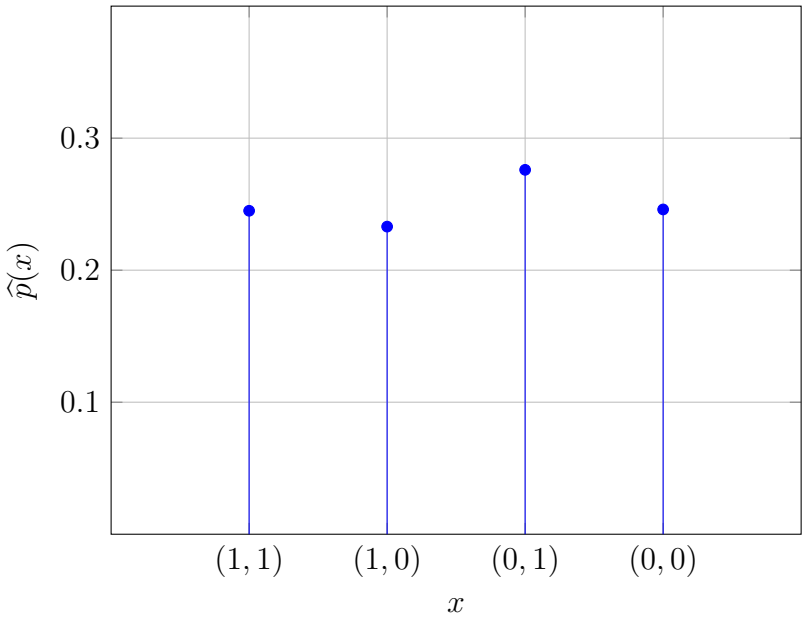


Figure 5.1: Relative frequencies of tossing two fair coins based on 10^3 repetitions.

X_1 \ X_2	0	1	$\sum$
0	0.246	0.276	0.522
1	0.233	0.245	0.478
$\sum$	0.479	0.521	1

Table II.3: Contingency table for empirical distribution in Figure 5.1.

6 Special Discrete Distributions

In this section we discuss several discrete probability distributions that serve as standard models for special random experiments.

Binomial Distribution

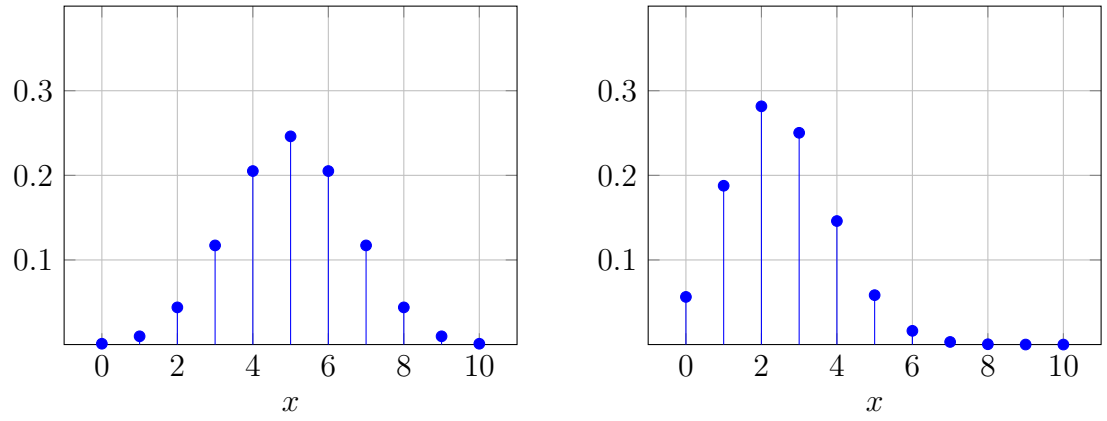
Consider a random experiment with the two outcomes 1 (success) and 0 (failure) that is repeated n times independently. This can be modeled by

(i) parameters $n \in \mathbb{N}$ (number of repetitions) and $p \in [0, 1]$ (probability of success),

(ii) independent and identically distributed (*i.i.d.*) random variables $X_1, \ldots, X_n$ satisfying

$$p = P(\{X_i = 1\}) = 1 - P(\{X_i = 0\})$$

for all $i = 1, \ldots, n$.



(a) $p = 1/2$. (b) $p = 1/4$.

Figure 6.1: Probability mass functions of $X \sim \mathrm{B}(10, p)$ with $p = 1/2$ and $p = 1/4$.

Put

$$S_n = \sum_{i=1}^{n} X_i$$

and note that for $\omega \in \Omega$ we have

$$S_n(\omega) = |\{i \in \{1, \dots, n\} \colon X_i(\omega) = 1\}|,$$

i.e., the number of successes is given by S_n.

Proposition 1. For $k \in \{0, \dots, n\}$ we have

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

Definition 2. A random variable X is *binomially distributed* with parameters $n \in \mathbb{N}$ und $p \in [0, 1]$ if

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

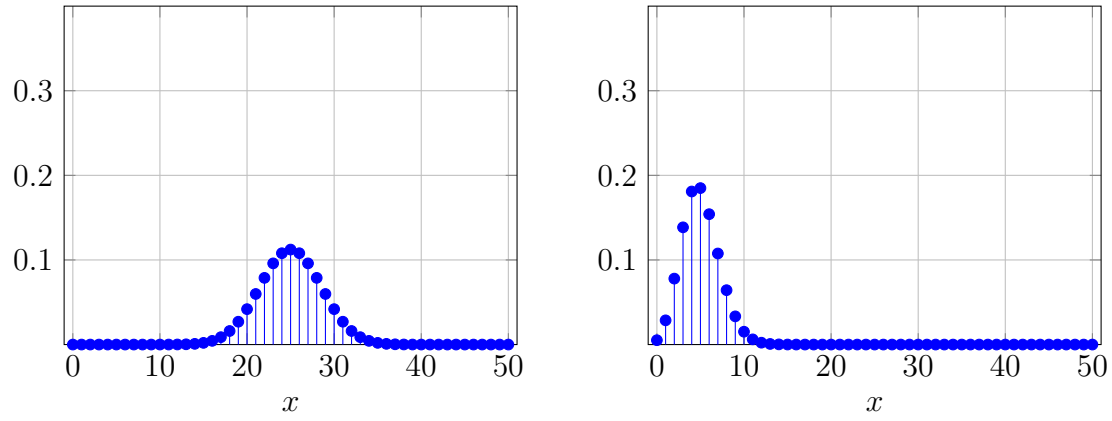
for all $k \in \{0, \dots, n\}$.
Notation: $X \sim \mathrm{B}(n, p)$.

Statistical Problem. Given n and k (by a sample), estimate p.

Example 3. The probability mass functions of $X \sim \mathrm{B}(n, p)$ for $n \in \{10, 50\}$ and different values of $p \in [0, 1]$ are illustrated in Figure 6.1 and Figure 6.2.

Proposition 4. Let X and Y be independent with $X \sim \mathrm{B}(n, p)$ and $Y \sim \mathrm{B}(m, p)$ for $m, n \in \mathbb{N}$ and $p \in [0, 1]$. Then we have

$$X + Y \sim \mathrm{B}(n + m, p).$$



(a) $p = 1/2$. (b) $p = 1/10$.

Figure 6.2: Probability mass functions of $X \sim \mathrm{B}(50, p)$ with $p = 1/2$ and $p = 1/10$.

Multinomial Distribution

Consider a random experiment with outcomes $0, \dots, m-1$ that is repeated n times independently. This can be modeled by

(i) parameters $n \in \mathbb{N}$ (number of repetitions), $m \in \mathbb{N} \setminus \{1\}$ (number of outcomes), and $p_0, \dots, p_{m-1} \in [0,1]$ such that $\sum_{j=0}^{m-1} p_j = 1$ (probabilities of m outcomes),

(ii) i.i.d. random variables $X_1, \dots, X_n$ satisfying

$$P(\{X_i = j\}) = p_j$$

for all $i = 1, \dots, n$ and $j = 0, \dots, m-1$.

Put

$$S_j(\omega) = |\{i \in \{1, \dots, n\} \colon X_i(\omega) = j\}|$$

for $\omega \in \Omega$ and $j \in \{0, \dots, m-1\}$, i.e., S_j is the number of random experiments with outcome j, and put

$$S = (S_0, \dots, S_{m-1}).$$

Remark 5. (i) For $j = 0, \dots, m-1$ we have $S_j \sim \mathrm{B}(n, p_j)$.

(ii) The random variables $S_0, \dots, S_{m-1}$ are not independent in general. We have

$$\sum_{j=0}^{m-1} S_j(\omega) = n$$

for all $\omega \in \Omega$.

Proposition 6. For $k = (k_0, \dots, k_{m-1}) \in \mathbb{N}_0^m$ with $\sum_{j=0}^{m-1} k_j = n$ we have

$$P(\{S = k\}) = \frac{n!}{k_0! \cdots k_{m-1}!} \cdot p_0^{k_0} \cdots p_{m-1}^{k_{m-1}}.$$

Definition 7. A random vector X follows a *multinomial distribution* with parameters $n \in \mathbb{N}$ and $p_0, \dots, p_{m-1} \in [0,1]$ with $\sum_{j=0}^{m-1} p_j = 1$ if

$$P(\{X = k\}) = \frac{n!}{k_0! \cdots k_{m-1}!} \cdot p_0^{k_0} \cdots p_{m-1}^{k_{m-1}}$$

for all $k = (k_0, \dots, k_{m-1}) \in \mathbb{N}_0^m$ with $\sum_{j=0}^{m-1} k_j = n$.
Notation: $X \sim \mathrm{M}(n, p_0, \dots, p_{m-1})$.

Hypergeometric Distribution

Consider a sample of size $n \in \mathbb{N}$ drawn without replacement from a set with $N \in \mathbb{N}$ elements consisting of K elements of type "success" and $N - K$ elements of type "failure". This can be modeled by

(i) parameters $N, K, n \in \mathbb{N}$ with $n \le N$ and $K \le N$,

(ii) the uniform distribution P on

$$\Omega = \{\omega \subseteq \{1, \dots, N\} \colon |\omega| = n\}.$$

Put

$$X(\omega) = |\omega \cap \{1, \dots, K\}|$$

for $\omega \in \Omega$, i.e., $\omega \subseteq \{1, \dots, N\}$ with $|\omega| = n$, to count the number of successes contained in the subset ω.
A typical application is given by quality control.

Remark 8. Proposition A.?? shows

$$|\Omega| = \binom{N}{n}.$$

Proposition 9. For $k \in \mathbb{N}_0$ with

$$n - (N - K) \le k \le \min(n, K) \tag{4}$$

we have

$$P(\{X = k\}) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}}. \tag{5}$$

Definition 10. A random variable X is *hypergeometrically distributed* with parameters $N, K, n \in \mathbb{N}$ with $n \le N$ and $K \le N$ if (5) holds for all $k \in \mathbb{N}_0$ with (4).
Notation: $X \sim \mathrm{H}(N, K, n)$.

Statistical Problem. (i) Given N, n, and k (by a sample), estimate K.

(ii) Given K, n, and k (by a sample), estimate N.

Proposition 11. Let $X_N \sim \mathrm{H}(N, K_N, n)$ for $N \in \mathbb{N}$ such that

$$\lim_{N\to\infty} \frac{K_N}{N} \in]0,1[\,.$$

Put $p = \lim_{N\to\infty} \frac{K_N}{N}$. Then we have

$$\lim_{N\to\infty} P(\{X_N = k\}) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

for all $k \in \{0, \dots, n\}$.

Proof. By assumption, we have $\lim_{N\to\infty} K_N = \infty$. Hence $k \in \mathbb{N}_0$ together with condition (4) becomes $k \in \{0, \dots, n\}$ if $N \in \mathbb{N}$ is large enough. In this case we have

$$\begin{aligned}
P(\{X_N = k\}) &= \frac{\binom{K_N}{k} \cdot \binom{N-K_N}{n-k}}{\binom{N}{n}} \\
&= \frac{K_N \cdot (K_N - 1) \cdot \ldots \cdot (K_N - k + 1)}{1 \cdot 2 \cdot \ldots \cdot k} \\
&\quad \cdot \frac{(N - K_N) \cdot (N - K_N - 1) \cdot \ldots \cdot (N - K_N - (n-k) + 1)}{1 \cdot 2 \cdot \ldots \cdot (n-k)} \\
&\quad \cdot \frac{1 \cdot 2 \cdot \ldots \cdot n}{N \cdot (N-1) \cdot \ldots \cdot (N-k+1) \cdot (N-k) \cdot \ldots \cdot (N-n+1)} \\
&= \binom{n}{k} \cdot \frac{K_N \cdot (K_N - 1) \cdot \ldots \cdot (K_N - k + 1)}{N \cdot (N-1) \cdot \ldots \cdot (N-k+1)} \\
&\quad \cdot \frac{(N - K_N) \cdot (N - K_N - 1) \cdot \ldots \cdot (N - K_N - (n-k) + 1)}{(N-k) \cdot \ldots \cdot (N-n+1)}
\end{aligned}$$

for $k \in \{0, \dots, n\}$. Moreover, the orange and the green factor tend to p^k and $(1-p)^{n-k}$, respectively, as $N \to \infty$. □

Example 12. The probability mass functions of $X \sim \mathrm{H}(100, 20, 10)$ and $Y \sim \mathrm{B}(10, 1/5)$ are illustrated in Figure 6.3. As indicated by Proposition 11, the distance

$$\max_{k\in\{0,\dots,n\}} |p_X(k) - p_Y(k)|$$

with $n = 10$ is rather small.

Poisson Distribution

Proposition 13 (Poisson limit theorem)**.** Let $X_n \sim \mathrm{B}(n, p_n)$ for $n \in \mathbb{N}$ such that

$$\lim_{n\to\infty} n \cdot p_n \in]0,\infty[\,.$$

Put $\lambda = \lim_{n\to\infty} n \cdot p_n$. Then we have

$$\lim_{n\to\infty} P(\{X_n = k\}) = \exp(-\lambda) \cdot \frac{\lambda^k}{k!}$$

for all $k \in \mathbb{N}_0$.

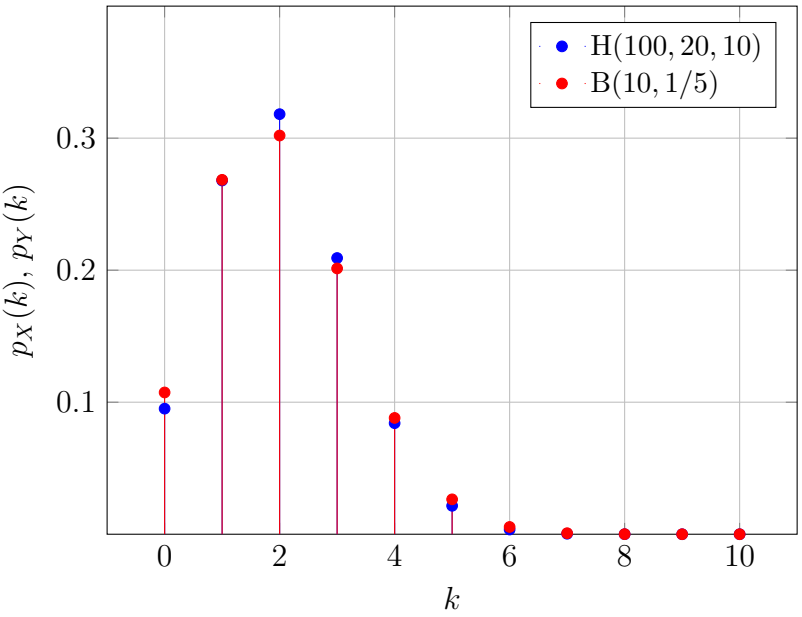


Figure 6.3: Probability mass functions of $X \sim \mathrm{H}(100, 20, 10)$ and $Y \sim \mathrm{B}(10, 1/5)$.

Proof. For $k \in \mathbb{N}_0$ and $n \in \mathbb{N}$ with $n \geq k$ we have

$$\begin{aligned} P(\{X_n = k\}) &= \binom{n}{k} \cdot p_n^k \cdot (1 - p_n)^{n-k} \\ &= \underbrace{\frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{n^k}}_{\to 1} \cdot \underbrace{\frac{(n \cdot p_n)^k}{k!}}_{\to \lambda^k / k!} \cdot \left(1 - \frac{n \cdot p_n}{n}\right)^n \cdot (1 - p_n)^{-k}. \end{aligned}$$

By assumption, we have $\lim_{n\to\infty} p_n = 0$ and thus $\lim_{n\to\infty} (1 - p_n)^{-k} = 1$. Moreover, every convergent sequence $(\lambda_n)_{n\in\mathbb{N}}$ with $\lambda = \lim_{n\to\infty} \lambda_n$ satisfies

$$\lim_{n\to\infty} \left(1 + \frac{\lambda_n}{n}\right)^n = \exp(\lambda).$$

Hence we get

$$\lim_{n\to\infty} P(\{X_n = k\}) = \frac{\lambda^k}{k!} \cdot \exp(-\lambda). \qquad \square$$

Remark 14. For all $\lambda \in \mathbb{R}$ we have $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \exp(\lambda)$.

Definition 15. A random variable X follows a *Poisson distribution* with parameter $\lambda \in]0, \infty[$ if

$$P(\{X = k\}) = \exp(-\lambda) \cdot \frac{\lambda^k}{k!}$$

for all $k \in \mathbb{N}_0$.

Notation: $X \sim \mathrm{Poi}(\lambda)$.

Example 16. The probability mass functions of $X \sim \mathrm{Poi}(\lambda)$ with $\lambda \in \{1/2, 3\}$ are illustrated in Figure 6.4. Moreover, the probability mass functions of $Y \sim \mathrm{B}(50, 1/10)$

and $Z \sim \text{Poi}(5)$ are illustrated in Figure 6.5. As indicated by Proposition 13, the distance

$$\max_{k\in\{0,\dots,n\}} |p_Y(k) - p_Z(k)|$$

with $n = 50$ is rather small.

Typical examples where a Poisson distribution serves as the stochastic model include

(i) number of decay events from a radioactive source within a certain time interval,

(ii) incoming calls in a call center per hour.

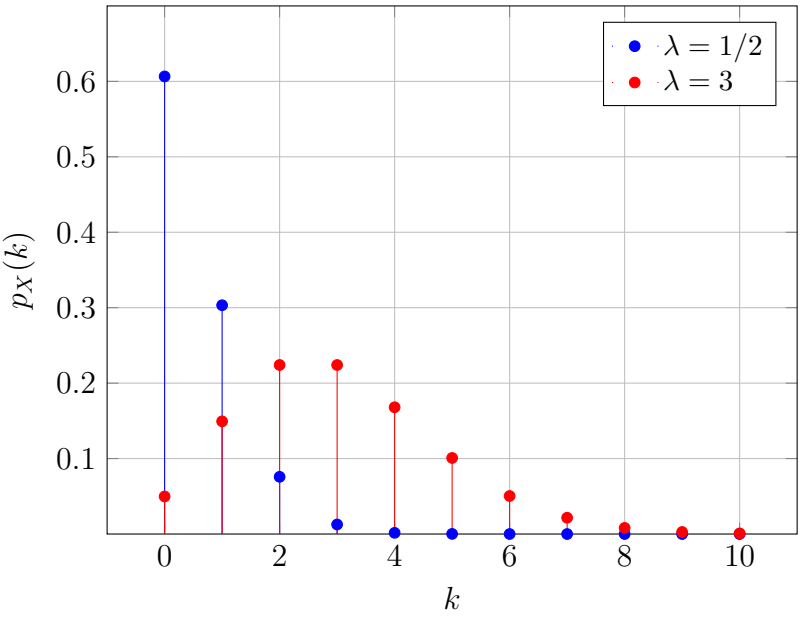


Figure 6.4: Probability mass functions of $X \sim \text{Poi}(\lambda)$ with $\lambda = 1/2$ and $\lambda = 3$.

Geometric Distribution

Consider a random experiment with the two outcomes 1 (success) and 0 (failure) that is repeated n times independently. This can be modeled by

(i) parameters $n \in \mathbb{N}$ (number of repetitions) and $p \in [0, 1]$ (probability of success),

(ii) i.i.d. random variables $X_1, \dots, X_n$ satisfying

$$p = P(\{X_i = 1\}) = 1 - P(\{X_i = 0\})$$

for all $i = 1, \dots, n$.

For $\omega \in \Omega$ we put

$$T_n(\omega) = 0$$

if

$$X_1(\omega) = \dots = X_n(\omega) = 0,$$

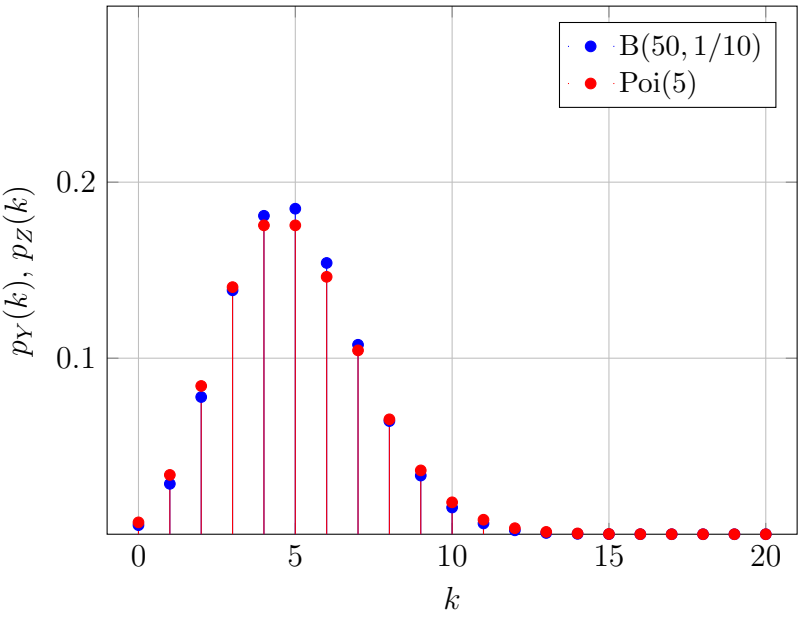


Figure 6.5: Probability mass functions of $Y \sim \mathrm{B}(50, 1/10)$ and $Z \sim \mathrm{Poi}(5)$.

and we put

$$T_n(\omega) = \min\{k \in \{1, \dots, n\} \colon X_k(\omega) = 1\}$$

otherwise, i.e., if $\{k \in \{1, \dots, n\} \colon X_k(\omega) = 1\} \neq \varnothing$. Note that T_n describes the (discrete) waiting time until the first success occurs within n trials.

Remark 17. (i) For $n \in \mathbb{N}$ and $k \in \{0, \dots, n\}$ we have

$$P(\{T_n = k\}) = \begin{cases} (1-p)^{k-1} \cdot p, & \text{if } k \in \{1, \dots, n\}, \\ (1-p)^n, & \text{if } k = 0. \end{cases}$$

In particular, we have $\lim_{n\to\infty} P(\{T_n = 0\}) = 0$.

(ii) For all $p \in]0, 1]$ we have $\sum_{k=1}^{\infty} (1-p)^{k-1} = 1/p$.

Definition 18. A random variable X is *geometrically distributed* with parameter $p \in]0, 1]$ if

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

for all $k \in \mathbb{N}$.

Notation: $X \sim \mathrm{Geo}(p)$.

7 Expected Value and Variance

In the sequel let $(\Omega, \mathcal{P}(\Omega), P)$ be a discrete probability space and let $X \colon \Omega \to \mathfrak{X}$ be a random variable with a countable set $\mathfrak{X} \subseteq \mathbb{R}$.

Expected Value

Definition 1. If $\sum_{\omega\in\Omega} |X(\omega)| \cdot P(\{\omega\}) < \infty$, the *expected value* (or *expectation*, *mean*) of X is defined by

$$\mathrm{E}(X) = \sum_{\omega\in\Omega} X(\omega) \cdot P(\{\omega\}).$$

Remark 2. If $|\Omega| < \infty$, the condition $\sum_{\omega\in\Omega} |X(\omega)| \cdot P(\{\omega\}) < \infty$ in Definition 1 is trivially satisfied for every random variable X.

Example 3. Let $|\Omega| < \infty$, and let P be the uniform distribution on Ω. Then we have

$$\mathrm{E}(X) = \frac{1}{|\Omega|} \cdot \sum_{\omega\in\Omega} X(\omega).$$

In particular, for $\Omega = \{0, \ldots, 36\}$ and

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \text{ is odd,} \\ -1, & \text{if } \omega \text{ is even,} \end{cases}$$

we have (Roulette, outside bet "odd")

$$\mathrm{E}(X) = -\frac{1}{37}.$$

Example 4. If X is constant, i.e., $X(\omega) = b \in \mathbb{R}$ for all $\omega \in \Omega$, we have

$$\mathrm{E}(X) = b \cdot \sum_{\omega\in\Omega} P(\{\omega\}) = b.$$

In the sequel let X and Y be random variables on $(\Omega, \mathcal{P}(\Omega), P)$. Moreover, we tacitly assume that the involved random variables satisfy the condition on absolute convergence in Definition 1.

Proposition 5.

(i) For all $c \in \mathbb{R}$ we have

$$\mathrm{E}(c \cdot X + Y) = c \cdot \mathrm{E}(X) + \mathrm{E}(Y). \qquad \textit{(linearity)}$$

(ii) If $X(\omega) \le Y(\omega)$ for all $\omega \in \Omega$, we have

$$\mathrm{E}(X) \le \mathrm{E}(Y). \qquad \textit{(monotonicity)}$$

Proof. ad (i): For $c \in \mathbb{R}$ we have

$$\begin{aligned} \mathrm{E}(c \cdot X + Y) &= \sum_{\omega\in\Omega} (c \cdot X(\omega) + Y(\omega)) \cdot P(\{\omega\}) \\ &= c \cdot \sum_{\omega\in\Omega} X(\omega) \cdot P(\{\omega\}) + \sum_{\omega\in\Omega} Y(\omega) \cdot P(\{\omega\}) = c \cdot \mathrm{E}(X) + \mathrm{E}(Y). \end{aligned}$$

ad (ii): Since $P(\{\omega\}) \ge 0$ for all $\omega \in \Omega$, we obtain

$$\mathrm{E}(X) = \sum_{\omega\in\Omega} X(\omega) \cdot P(\{\omega\}) \le \sum_{\omega\in\Omega} Y(\omega) \cdot P(\{\omega\}) = \mathrm{E}(Y). \qquad \square$$

Proposition 6. For every function $g\colon \mathbb{R} \to \mathbb{R}$ we have

$$\mathrm{E}(g(X)) = \sum_{x \in \mathfrak{X}} g(x) \cdot p_X(x).$$

In particular, the expected value of X only depends on the distribution of X.

Proposition 7 (Expected value of special discrete distributions)**.**

(i) For $X \sim \mathrm{B}(n,p)$ with $n \in \mathbb{N}$ and $p \in [0,1]$ we have

$$\mathrm{E}(X) = n \cdot p.$$

(ii) For $X \sim \mathrm{H}(N,K,n)$ with $N, K, n \in \mathbb{N}$, $n \leq N$, and $K \leq N$ we have

$$\mathrm{E}(X) = \frac{n \cdot K}{N}.$$

(iii) For $X \sim \mathrm{Poi}(\lambda)$ with $\lambda > 0$ we have

$$\mathrm{E}(X) = \lambda.$$

(iv) For $X \sim \mathrm{Geo}(p)$ with $p \in \,]0,1]$ we have

$$\mathrm{E}(X) = \frac{1}{p}.$$

Proposition 8 (Product of independent random variables)**.** If X and Y are independent, we have

$$\mathrm{E}(X \cdot Y) = \mathrm{E}(X) \cdot \mathrm{E}(Y).$$

Variance

Definition 9. If $\sum_{\omega \in \Omega} |X(\omega)|^2 \cdot P(\{\omega\}) < \infty$, the *variance* of X is defined by

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{E}\left(\big(X - \mathrm{E}(X)\big)^2\right).$$

In this case $\sigma_X = \sqrt{\mathrm{Var}(X)}$ is called *standard deviation* of X.

In the sequel we tacitly assume that the involved random variables satisfy the condition on absolute convergence in Definition 9. See also Remark 2.

Remark 10. Proposition 6 with $g(x) = (x - \mathrm{E}(X))^2$ shows

$$\mathrm{Var}(X) = \sum_{x \in \mathfrak{X}} (x - \mathrm{E}(X))^2 \cdot p_X(x).$$

In particular, the variance of X only depends on the distribution of X.

Example 11. For $X \sim \mathrm{B}(1,p)$ with $p \in [0,1]$ we have

$$\begin{aligned}\mathrm{Var}(X) &= (0-p)^2 \cdot P(\{X=0\}) + (1-p)^2 \cdot P(\{X=1\}) \\ &= p^2 \cdot (1-p) + (1-p)^2 \cdot p \\ &= p \cdot (1-p).\end{aligned}$$

As a function of p, $\mathrm{Var}(X)$ attains its maximum at $p = 1/2$, and $\mathrm{Var}(X)$ attains its minimum at $p = 0$ and $p = 1$.

Proposition 12.

(i) We have

$$\mathrm{Var}(X) = \mathrm{E}\big(X^2\big) - \big(\mathrm{E}(X)\big)^2.$$

(ii) For all $a, b \in \mathbb{R}$ we have

$$\mathrm{Var}(a \cdot X + b) = a^2 \cdot \mathrm{Var}(X).$$

Proof. ad (i): Using Proposition 5.(i) and Example 4 we obtain

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\left(\big(X - \mathrm{E}(X)\big)^2\right) \\ &= \mathrm{E}\left(X^2 - 2X \cdot \mathrm{E}(X) + \big(\mathrm{E}(X)\big)^2\right) \\ &= \mathrm{E}\big(X^2\big) - 2 \cdot \big(\mathrm{E}(X)\big)^2 + \big(\mathrm{E}(X)\big)^2 = \mathrm{E}\big(X^2\big) - \big(\mathrm{E}(X)\big)^2.\end{aligned}$$

ad (ii): By Proposition 5.(i) we have

$$\begin{aligned}\mathrm{Var}(a \cdot X + b) &= \mathrm{E}\left(\big(aX + b - \mathrm{E}(aX+b)\big)^2\right) \\ &= \mathrm{E}\left(\big(aX + b - a\,\mathrm{E}(X) - b\big)^2\right) \\ &= \mathrm{E}\left(a^2 \cdot \big(X - \mathrm{E}(X)\big)^2\right) = a^2 \cdot \mathrm{Var}(X).\end{aligned}$$

□

Proposition 13 (Bienaymé's identity)**.** If X and Y are independent, we have

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Proof. Using Proposition 12.(i) and Proposition 5.(i) we obtain

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{E}\big((X+Y)^2\big) - \big(\mathrm{E}(X+Y)\big)^2 \\ &= \mathrm{E}\big(X^2 + Y^2 + 2XY\big) - \big(\mathrm{E}(X) + \mathrm{E}(Y)\big)^2 \\ &= \mathrm{E}\big(X^2\big) + \mathrm{E}\big(Y^2\big) + 2\,\mathrm{E}(XY) - \big(\mathrm{E}(X)\big)^2 - \big(\mathrm{E}(Y)\big)^2 - 2\,\mathrm{E}(X)\,\mathrm{E}(Y) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \big(\mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)\big).\end{aligned}$$

Since X, Y are independent, we may apply Proposition 8. □

Proposition 14 (Variance of special discrete distributions)**.**

(i) For $X \sim \mathrm{B}(n,p)$ with $n \in \mathbb{N}$ and $p \in [0,1]$ we have

$$\mathrm{Var}(X) = n \cdot p \cdot (1-p).$$

(ii) For $X \sim \mathrm{H}(N,K,n)$ with $N,K,n \in \mathbb{N}$, $n \le N$, $K \le N$, and $N \ge 2$ we have

$$\mathrm{Var}(X) = \frac{n \cdot (N-n) \cdot (N-K) \cdot K}{N^2 \cdot (N-1)}.$$

(iii) For $X \sim \mathrm{Poi}(\lambda)$ with $\lambda > 0$ we have

$$\mathrm{Var}(X) = \lambda.$$

(iv) For $X \sim \mathrm{Geo}(p)$ with $p \in]0,1]$ we have

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

Proposition 15 (Chebyshev's inequality)**.** For all $\varepsilon > 0$ we have

$$P(\{|X - \mathrm{E}(X)| \ge \varepsilon\}) \le \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

Proof. For $\varepsilon > 0$ and $A \in \mathcal{P}(\Omega)$ given by

$$A = \{|X - \mathrm{E}(X)| \ge \varepsilon\} = \{\omega \in \Omega \colon (X(\omega) - \mathrm{E}(X))^2 \ge \varepsilon^2\}$$

we obtain

$$\begin{aligned} \varepsilon^2 \cdot P(A) = \sum_{\omega \in A} \varepsilon^2 \cdot P(\{\omega\}) &\le \sum_{\omega \in A} (X(\omega) - \mathrm{E}(X))^2 \cdot P(\{\omega\}) \\ &\le \sum_{\omega \in \Omega} (X(\omega) - \mathrm{E}(X))^2 \cdot P(\{\omega\}) = \mathrm{Var}(X). \end{aligned}$$

□

Sample Mean and Sample Variance

In the sequel let

$$x_1, \ldots, x_N \in \mathfrak{X}$$

be a sample with relative frequencies $\widehat{p}(x)$ for $x \in \mathfrak{X}$, see Section 5.

Definition 16. The *sample mean* (or *empirical mean*) $\overline{x}_N$ is defined by

$$\overline{x}_N = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i.$$

Remark 17. We have

$$\overline{x}_N = \frac{1}{N} \cdot \sum_{x \in \mathfrak{X}} x \cdot |\{\ell \in \{1, \ldots, N\} \colon x_\ell = x\}| = \sum_{x \in \mathfrak{X}} x \cdot \widehat{p}(x).$$

Hence $\overline{x}_N$ is the expected value of the associated empirical distribution.

Definition 18. Let $N \in \mathbb{N}$ with $N \geq 2$. The *sample variance* s_N^2 is defined by

$$s_N^2 = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \overline{x}_N)^2.$$

Moreover, $s_N = \sqrt{s_N^2}$ is called *sample standard deviation*.

Remark 19. We have

$$s_N^2 = \frac{1}{N-1} \cdot \left(\sum_{i=1}^{N} x_i^2 - N \cdot \overline{x}_N^2 \right).$$

Hence $(N-1)/N \cdot s_N^2$ is the variance of the associated empirical distribution.

Chapter III

Probability Theory – General Case

In Chapter II we restricted ourselves to probability spaces $(\Omega, \mathcal{P}(\Omega), P)$ and random variables $X\colon \Omega \to \mathfrak{X}$ where Ω and $\mathfrak{X} \subseteq \mathbb{R}$ were countable. In this chapter we consider sets Ω and $\mathfrak{X} \subseteq \mathbb{R}$ with possibly uncountably many elements. In this case it is in general not possible anymore

(i) to assign probabilities to arbitrary subsets $A \in \mathcal{P}(\Omega)$ of Ω,

(ii) to assign probabilities $P(\{X \in A\})$ for arbitrary subsets $A \in \mathcal{P}(\mathbb{R})$ of $\mathbb{R}$,

see, e.g., [Georgii, pp. 8-9].
In order to overcome this problem, the event spaces $\mathcal{P}(\Omega)$ and $\mathcal{P}(\mathbb{R})$ are replaced by smaller collections of subsets that do not contain the "problematic" subsets. These smaller collections are of the following type.

Definition 1. A collection $\mathcal{A} \subset \mathcal{P}(\Omega)$ of subsets of Ω is a *σ-algebra* (over Ω) if

(i) $\Omega \in \mathcal{A}$,

(ii) $A \in \mathcal{A} \;\Rightarrow\; A^{\mathsf{c}} \in \mathcal{A}$,

(iii) $A_1, A_2, \ldots \in \mathcal{A} \;\Rightarrow\; \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

A probability space is then a triple $(\Omega, \mathcal{A}, P)$, where $\mathcal{P}(\Omega)$ is consistently replaced by a σ-algebra $\mathcal{A}$ in Definition II.1.9. In particular, we have $P\colon \mathcal{A} \to \mathbb{R}$.
For the real numbers we consider the so-called *Borel-σ-algebra* $\mathfrak{B}(\mathbb{R})$, which contains all open subsets of $\mathbb{R}$ and which is generated by sets of the form

$$]-\infty, b], \quad b \in \mathbb{R}.$$

We stress that every interval is contained in $\mathfrak{B}(\mathbb{R})$. In particular, we have

$$]-\infty, b], [a, b], \{a\} \in \mathfrak{B}(\mathbb{R})$$

for all $a, b \in \mathbb{R}$. Moreover, Definition 1 ensures that unions of intervals and complements thereof are also contained in $\mathfrak{B}(\mathbb{R})$.

A random variable $X\colon \Omega \to \mathbb{R}$ is a so-called *measurable* function, i.e., we require

$$\forall\, b \in \mathbb{R}\colon X^{-1}(]-\infty, b]) \in \mathcal{A}.$$

This property ensures

$$\{X \in A\} = \{\omega \in \Omega\colon X(\omega) \in A\} \in \mathcal{A}$$

for all $A \in \mathfrak{B}(\mathbb{R})$. Hence we can assign probabilities $P(\{X \in A\})$. In the sequel we disregard measurability issues.

1 Absolutely Continuous Distributions

Aim: Model for continuous quantities as, e.g., waiting time, lifetime, or deviations in a production process.
Basic idea: Replace sums by integrals.

Definition 1. A function $f\colon \mathbb{R} \to \mathbb{R}$ is a *(probability) density (PDF)* if

$$(\forall\, x \in \mathbb{R}\colon f(x) \geq 0) \quad \wedge \quad \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

Example 2. The following functions $f_1, f_2, f_3\colon \mathbb{R} \to \mathbb{R}$ are probability densities:

(i) For $a, b \in \mathbb{R}$ with $a < b$ define

$$f_1(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a, b], \\ 0, & \text{else.} \end{cases}$$

(ii) For $\lambda > 0$ define

$$f_2(x) = \begin{cases} \lambda \cdot \exp(-\lambda \cdot x), & \text{if } x \geq 0, \\ 0, & \text{else.} \end{cases}$$

(iii) For $\mu \in \mathbb{R}$ and $\sigma > 0$ define

$$f_3(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

The functions f_2 and f_3 are illustrated in Figure 1.1 and Figure 1.2, respectively.

In the sequel let $(\Omega, \mathcal{A}, P)$ be a probability space and let $X\colon \Omega \to \mathbb{R}$ be a random variable.

Definition 3 (cf. Prop. II.4.5)**.** A random variable X is *absolutely continuous* if there exists a probability density f_X such that

$$P(\{X \in A\}) = \int_A f_X(x)\,\mathrm{d}x$$

for all $A \in \mathfrak{B}(\mathbb{R})$. In this case f_X is called *(probability) density* of X.

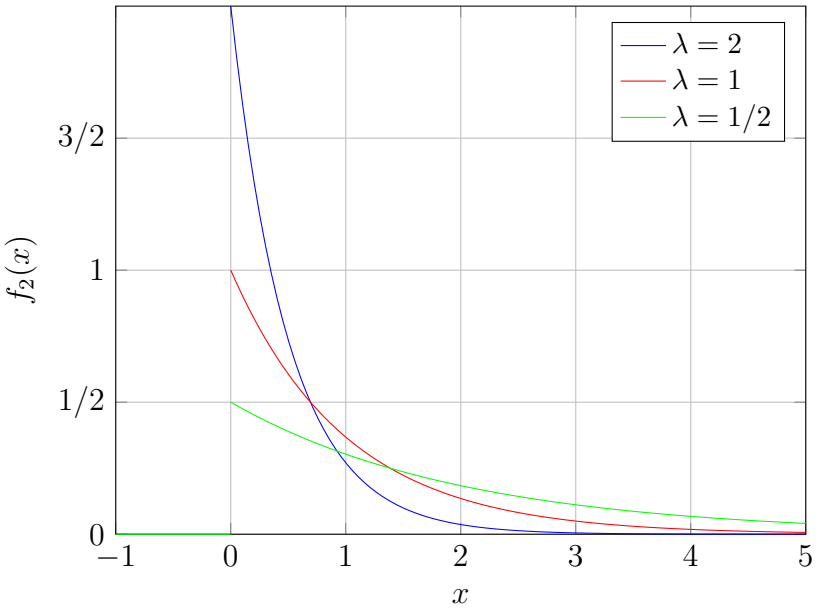


Figure 1.1: Probability density f_2 from Example 2.(ii) with $\lambda \in \{1/2, 1, 2\}$.

Remark 4. Let X be absolutely continuous with density f_X. Then we have

$$\forall\, x \in \mathbb{R}\colon\ P(\{X = x\}) = 0.$$

Definition 5. Let X be a random variable with density f_X.

(i) X is *uniformly distributed* on $[a, b]$ if $f_X = f_1$ according to Example 2.(i).
Notation: $X \sim \mathrm{U}(a, b)$.

(ii) X is *exponentially distributed* with parameter $\lambda \in \,]0, \infty[$ if $f_X = f_2$ according to Example 2.(ii).
Notation: $X \sim \mathrm{Exp}(\lambda)$.

(iii) X is *normally distributed* with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ if $f_X = f_3$ according to Example 2.(iii). X is *standard normally distributed* if $\mu = 0$ and $\sigma = 1$.
Notation: $X \sim \mathrm{N}(\mu, \sigma^2)$.

Remark 6. Let $X \sim \mathrm{Exp}(\lambda)$ with $\lambda > 0$, and let $s, t \geq 0$. Then we have

$$P(\{X \geq t\}) = \int_t^\infty \lambda \cdot \exp(-\lambda x)\,\mathrm{d}x = -\exp(-\lambda \cdot x)\big|_{x=t}^{x=\infty} = \exp(-\lambda \cdot t).$$

Hence the exponential distribution is memoryless, i.e.,

$$P(\{X \geq s + t\} \,|\, \{X \geq t\}) = \frac{P(\{X \geq s + t\})}{P(\{X \geq t\})} = P(\{X \geq s\}).$$

Application: Modelling of waiting times or lifetimes.

In the sequel let X be absolutely continuous with density f_X.

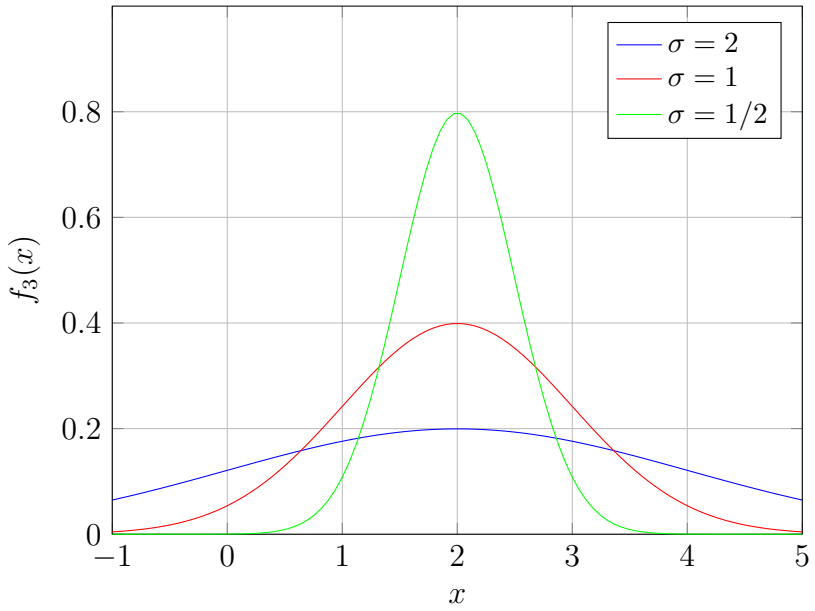


Figure 1.2: Probability density f_3 from Example 2.(iii) with $\mu = 2$ and $\sigma \in \{1/2, 1, 2\}$.

Definition 7 (cf. Prop. II.7.6, and Rem. II.7.10)**.** If $\int_{-\infty}^{\infty} |x| \cdot f_X(x)\,\mathrm{d}x < \infty$, the *expected value* (or *expectation*, *mean*) of X is defined by

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}x.$$

If $\int_{-\infty}^{\infty} x^2 \cdot f_X(x)\,\mathrm{d}x < \infty$, the *variance* of X is defined by

$$\sigma_X^2 = \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 \cdot f_X(x)\,\mathrm{d}x.$$

In this case $\sigma_X = \sqrt{\mathrm{Var}(X)}$ is called *standard deviation* of X.

In the sequel we tacitly assume that the involved random variables satisfy the required integrability conditions for the expected value and the variance.

Proposition 8 (Expected value and variance of special distributions)**.**

(i) For $X \sim \mathrm{U}(a, b)$ with $a < b$ we have

$$\mathrm{E}(X) = \frac{a+b}{2}, \qquad \mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

(ii) For $X \sim \mathrm{Exp}(\lambda)$ with $\lambda > 0$ we have

$$\mathrm{E}(X) = \frac{1}{\lambda}, \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

(iii) For $X \sim \mathrm{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$ we have

$$\mathrm{E}(X) = \mu, \qquad \mathrm{Var}(X) = \sigma^2.$$

Proposition 9 (cf. Prop. II.7.6)**.** For (piecewise) continuous functions $g\colon \mathbb{R} \to \mathbb{R}$ with $\int_{-\infty}^{\infty} |g(x)| \cdot f_X(x)\,\mathrm{d}x < \infty$, we have

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\,\mathrm{d}x.$$

Remark 10. Proposition II.7.5 (linearity and monotonicity) and Proposition II.7.8 (product of independent random variables) regarding the expected value as well as Proposition II.7.12 (affine linear transformation), Proposition II.7.13 (Bienaymé's identity), and Proposition II.7.15 (Chebyshev's inequality) regarding the variance are also valid for absolutely continuous random variables.

2 Basic Concepts

In the sequel let $(\Omega, \mathcal{A}, P)$ be a probability space and let $X\colon \Omega \to \mathbb{R}$ be a random variable.

Recall that $P_X\colon \mathfrak{B}(\mathbb{R}) \to \mathbb{R}$ is given by

$$P_X(A) = P(\{X \in A\}) = P(\{\omega \in \Omega\colon X(\omega) \in A\})$$

for $A \in \mathfrak{B}(\mathbb{R})$.

Proposition 1 (cf. Prop. II.4.5)**.** The function P_X is a probability measure (on $\mathbb{R}$).

Definition 2 (cf. Def. II.4.6)**.** The function P_X is called *distribution* of X.

Definition 3. X is *discrete* if there exists a countable set $\mathfrak{X} \subseteq \mathbb{R}$ such that

$$P_X(\mathfrak{X}) = 1.$$

Definition 4 (cf. Def. II.4.8)**.** Two random variables $X_1, X_2\colon \Omega \to \mathbb{R}$ are *identically distributed* if

$$P_{X_1}(A) = P_{X_2}(A)$$

for all $A \in \mathfrak{B}(\mathbb{R})$.

Remark 5. Proposition II.4.10 is not correct in general. In particular, for absolutely continuous random variables $X_1, X_2\colon \Omega \to \mathbb{R}$ we always have

$$P(\{X_1 = x\}) = 0 = P(\{X_2 = x\})$$

for all $x \in \mathbb{R}$.

Definition 6 (cf. Def. II.4.13)**.** Random variables $X_1, \ldots, X_n\colon \Omega \to \mathbb{R}$ are *independent* if for all $A_1, \ldots, A_n \in \mathfrak{B}(\mathbb{R})$ we have

$$P\left(\bigcap_{i=1}^{n} \{X_i \in A_i\}\right) = \prod_{i=1}^{n} P(\{X_i \in A_i\}).$$

Remark 7. Proposition II.4.15 is not correct in general.

3 Cumulative Distribution Function and Quantiles

In the sequel let $(\Omega, \mathcal{A}, P)$ be a probability space and let $X, Y \colon \Omega \to \mathbb{R}$ be random variables.

Definition 1. The function $F_X \colon \mathbb{R} \to [0,1]$ defined by

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

is called *cumulative distribution function (CDF)* of X.

Remark 2. For $x \in \mathbb{R}$ we have

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

For $u, v \in \mathbb{R}$ with $u < v$ we have

$$P(\{X \in]u, v]\}) = P(\{X \leq v\} \setminus \{X \leq u\}) = F_X(v) - F_X(u).$$

Proposition 3 (Uniqueness)**.** A cumulative distribution function uniquely determines the distribution, i.e., we have

$$F_X = F_Y \;\Leftrightarrow\; P_X = P_Y.$$

Remark 4. Let $Z = c \cdot X + d$ with $c > 0$ and $d \in \mathbb{R}$. Then we have

$$F_Z(x) = P(\{c \cdot X + d \leq x\}) = P(\{X \leq (x-d)/c\}) = F_X((x-d)/c))$$

for all $x \in \mathbb{R}$.

At first, we consider absolutely continuous random variables.

Proposition 5 (CDF, absolutely continuous random variables)**.** Let X be absolutely continuous with density f_X. Then we have

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y$$

for all $x \in \mathbb{R}$ and F_X is continuous. Moreover, if f_X is continuous at x, then F_X is differentiable at x with $F_X'(x) = f_X(x)$.

Proof. Fundamental theorem of calculus. □

Example 6 (CDF, continuation of Example 1.2)**.**

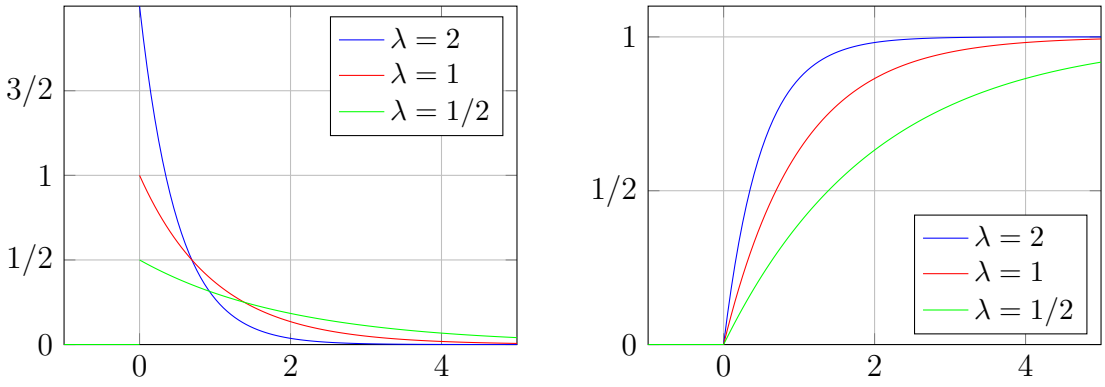
(i) For $Y \sim \mathrm{U}(a,b)$ with $a, b \in \mathbb{R}$ and $a < b$ we have

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

Moreover, let $X \sim \mathrm{U}(0,1)$ and $Z = (b-a) \cdot X + a$. By Remark 4, we have

$$F_Z(x) = F_X\big((x-a)/(b-a)\big) = F_Y(x)$$

for all $x \in \mathbb{R}$, and hence $Z \sim \mathrm{U}(a,b)$ due to Proposition 3.



(a) PDF (b) CDF

Figure 3.1: PDF and CDF of $Y \sim \mathrm{Exp}(\lambda)$ with $\lambda \in \{1/2, 1, 2\}$ from Example 6.(ii).

(ii) For $Y \sim \mathrm{Exp}(\lambda)$ with $\lambda > 0$ we have

$$F_Y(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1 - \exp(-\lambda \cdot x), & \text{if } x \geq 0. \end{cases}$$

Moreover, let $X \sim \mathrm{Exp}(1)$ and $Z = 1/\lambda \cdot X$. By Remark 4, we have

$$F_Z(x) = F_X(x \cdot \lambda) = F_Y(x)$$

for all $x \in \mathbb{R}$, and hence $Z \sim \mathrm{Exp}(\lambda)$ due to Proposition 3.

The PDF and the CDF of $Y \sim \mathrm{Exp}(\lambda)$ with $\lambda \in \{1/2, 1, 2\}$ are illustrated in Figure 3.1.

(iii) The CDF of a standard normal distribution is denoted by Φ, i.e., for $x \in \mathbb{R}$ we have

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x} \exp(-y^2/2)\,\mathrm{d}y.$$

Since the corresponding PDF is an even function, we have

$$\Phi(-x) = 1 - \Phi(x)$$

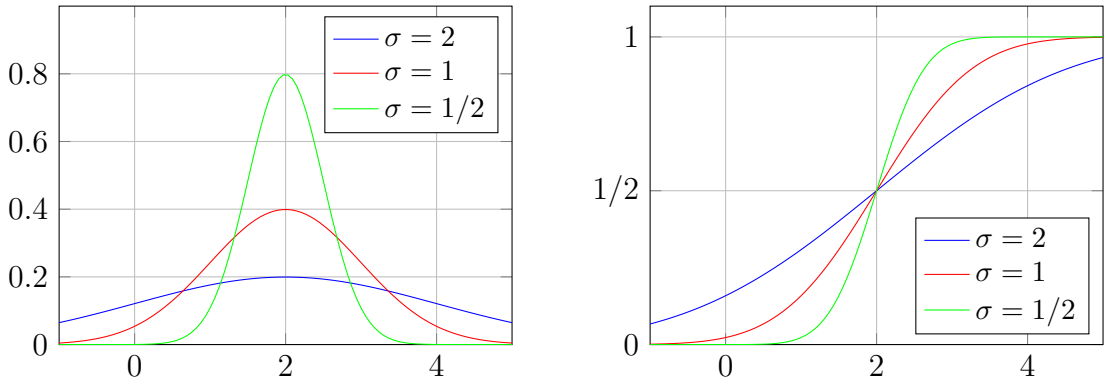
for all $x \in \mathbb{R}$. The function Φ cannot be expressed in terms of elementary functions. Function values $\Phi(x)$ can be computed numerically and are listed in Table B.1 for $x \geq 0$.

Let $X \sim N(0,1)$ and $Z = \sigma \cdot X + \mu$. The change of variables formula shows

$$P(\{Z \in A\}) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \int_A \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x$$

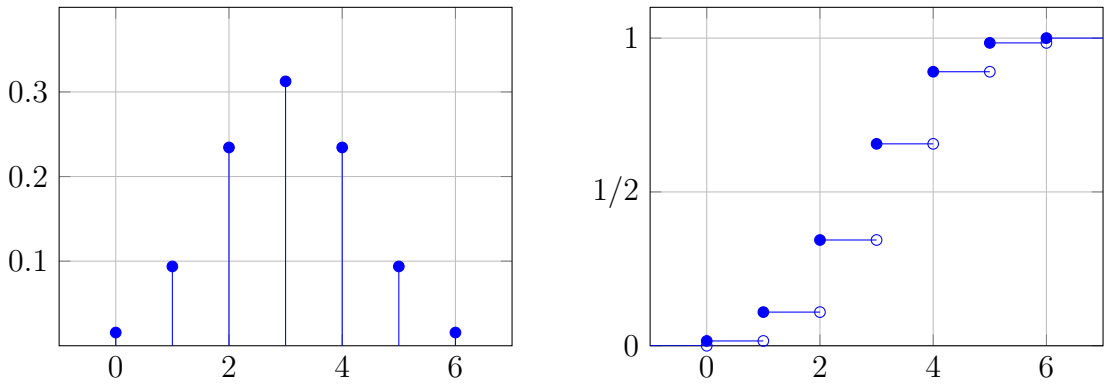
for $A \in \mathfrak{B}(\mathbb{R})$, and hence $Z \sim \mathrm{N}(\mu, \sigma^2)$. By Remark 4, the CDF of a normal distribution with arbitrary parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ can be reduced to the standard normal case, i.e., for all $x \in \mathbb{R}$ we have

$$F_Z(x) = F_X\big((x-\mu)/\sigma\big) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$



(a) PDF (b) CDF

Figure 3.2: PDF and CDF of $Z \sim \mathrm{N}(\mu, \sigma^2)$ with $\mu = 2$ and $\sigma \in \{1/2, 1, 2\}$ from Example 6.(iii).



(a) PMF (b) CDF

Figure 3.3: PMF and CDF of $X \sim \mathrm{B}(6, 1/2)$.

The PDF and the CDF of $Z \sim \mathrm{N}(\mu, \sigma^2)$ with $\mu = 2$ and $\sigma \in \{1/2, 1, 2\}$ are illustrated in Figure 3.2.

Proposition 7 (CDF, discrete random variables)**.** Let X be a discrete random variable with $P_X(\mathbb{N}_0) = 1$. Then we have

$$F_X(x) = \sum_{i=0}^{k} p_X(i)$$

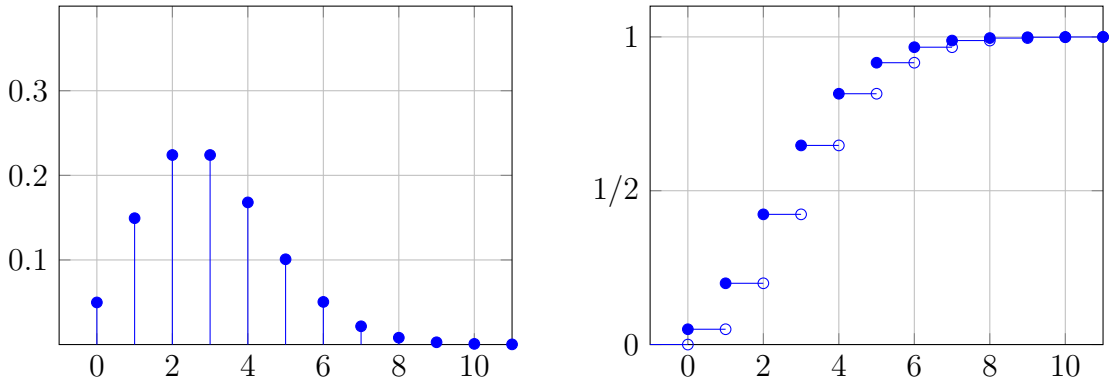
for all $k \in \mathbb{N}_0$ and $x \in [k, k+1[$.

Example 8. The PMF and CDF of $X \sim \mathrm{B}(6, 1/2)$ and $Y \sim \mathrm{Poi}(3)$ are illustrated in Figure 3.3 and Figure 3.4, respectively.

Proposition 9 (Properties of CDFs, general case)**.** Every cumulative distribution function $F_X \colon \mathbb{R} \to [0, 1]$ satisfies

(i) F_X is non-decreasing,

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



(a) PMF (b) CDF

Figure 3.4: PMF and CDF of $Y \sim \mathrm{Poi}(3)$.

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

Moreover, we have

$$P(\{X = x\}) = F_X(x) - \lim_{y\uparrow x} F_X(y)$$

for all $x \in \mathbb{R}$. In particular, $P(\{X = x\}) = 0$ iff F_X is continuous at x.

Quantiles

Given: Cumulative distribution function F_X and $p \in]0, 1[$.
Aim: Find $q \in \mathbb{R}$ with

$$F_X(q) = p. \tag{1}$$

Example 10. Let $X \sim \mathrm{Exp}(\lambda)$ with $\lambda > 0$. For every $p \in]0, 1[$ there exists a unique $q \in \mathbb{R}$ satisfying (1), namely

$$q = -\frac{1}{\lambda} \cdot \ln(1 - p).$$

E.g., for $\lambda = 1$ and $p = 3/4$ we have $q = \ln(4) \approx 1.3863$, see Figure 3.5.

Example 11. Let $X \sim \mathrm{B}(2, 1/2)$. For $p \in \{1/4, 3/4\}$ there exist infinitely many $q \in \mathbb{R}$ satisfying (1). For $p \in]0, 1[\setminus \{1/4, 3/4\}$ there is no $q \in \mathbb{R}$ satisfying (1).

Definition 12. The *p-quantile* q of F_X (of P_X, of X) is defined by

$$q = \min\{x \in \mathbb{R} \colon F_X(x) \geq p\}.$$

The p-quantile for $p = 1/2$ is called *median* of F_X (of P_X, of X).
Notation: median $\mathsf{m}(X)$.

Example 13 (Continuation of Example 10)**.** Let $X \sim \mathrm{Exp}(\lambda)$ with $\lambda > 0$. The p-quantile of F_X is given by

$$q = -\frac{1}{\lambda} \cdot \ln(1 - p).$$

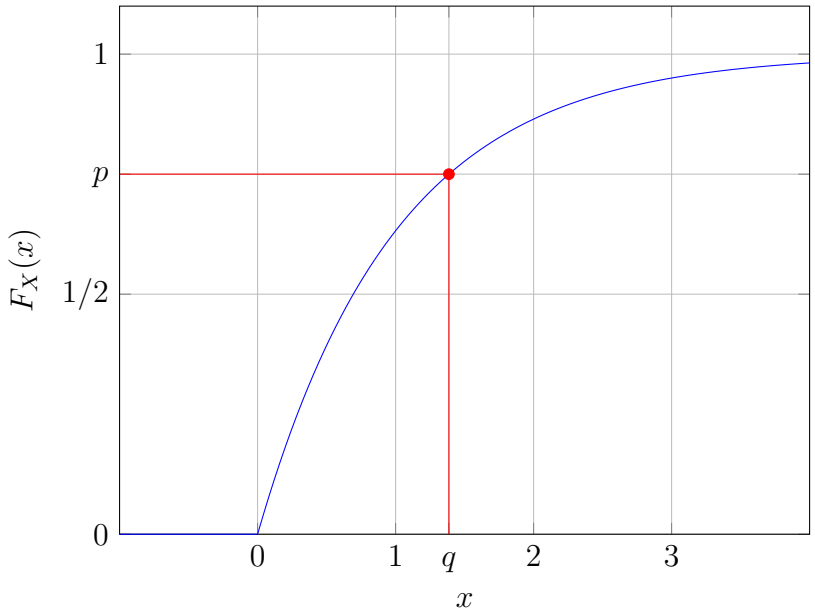


Figure 3.5: CDF of $X \sim \mathrm{Exp}(1)$ with $p \in]0,1[$ and $q \in \mathbb{R}$ such that $F_X(q) = p$, see Example 10.

Example 14 (Continuation of Example 11)**.** Let $X \sim \mathrm{B}(2, 1/2)$. The p-quantile of F_X is given by

$$q = \begin{cases} 0, & \text{if } 0 < p \leq 1/4, \\ 1, & \text{if } 1/4 < p \leq 3/4, \\ 2, & \text{if } 3/4 < p < 1. \end{cases}$$

Proposition 15. The p-quantile q of X satisfies

$$P(\{X \leq q\}) \geq p \quad \wedge \quad P(\{X \geq q\}) \geq 1 - p.$$

Moreover, we have

$$|\,\mathrm{E}(X) - \mathsf{m}(X)| \leq \sigma_X.$$

Empirical Distribution Function

As in Section II.5 we consider a sample

$$x_1, \ldots, x_N \in \mathbb{R}.$$

Definition 16. The *empirical distribution function* $\widehat{F}\colon \mathbb{R} \to [0,1]$ is defined by

$$\widehat{F}(x) = \frac{|\{\ell \in \{1, \ldots, N\}\colon x_\ell \leq x\}|}{N}.$$

Remark 17. The empirical distribution function is the CDF of the empirical distribution $\widehat{P}$ of the sample. The corresponding order statistics

$$x_{(1)} \leq \cdots \leq x_{(N)}$$

of the sample allows for a straightforward computation of $\widehat{p}(x)$, $\widehat{F}(x)$, and quantiles.

Remark 18. Consider the sample mean $\overline{x}_N$ and the *sample median* $\tilde{x}_N$ given by

$$\tilde{x}_N = \begin{cases} x_{((N+1)/2)}, & \text{if } N \text{ is odd,} \\ x_{(N/2)}, & \text{if } N \text{ is even.} \end{cases}$$

Question: How many values x_ℓ of the sample must be modified in order that $\overline{x}_N$ or $\tilde{x}_N$ attain an arbitrary value?

Answer:

(i) For $\overline{x}_N$ a single value x_ℓ suffices.

(ii) Note that

$$\frac{|\{\ell \in \{1, \dots, N\} \colon a \le x_\ell \le b\}|}{N} > 1/2$$

implies $\tilde{x}_N \in [a, b]$, see Proposition 15. For $\tilde{x}_N$ we need to modify at least $N/2$ values.

In this sense the sample median $\tilde{x}_N$ is more robust than the sample mean $\overline{x}_N$.

Example 19 (Empirical CDF and histogram)**.** Consider the following sample

$$3, 7, 9, 17, 23, 2, 7, 11, 21, 12, 5, 13, 15, 4, 1, 5, 12, 6, 12, 7, 8, 20, 3, 11, 8$$

of size $N = 25$. The corresponding order statistics read

$$1, 2, 3, 3, 4, 5, 5, 6, 7, 7, 7, 8, 8, 9, 11, 11, 12, 12, 12, 13, 15, 17, 20, 21, 23.$$

The sample median is given by $\tilde{x}_{25} = x_{(13)} = 8$. The empirical CDF of the sample is illustrated in Figure 3.6.

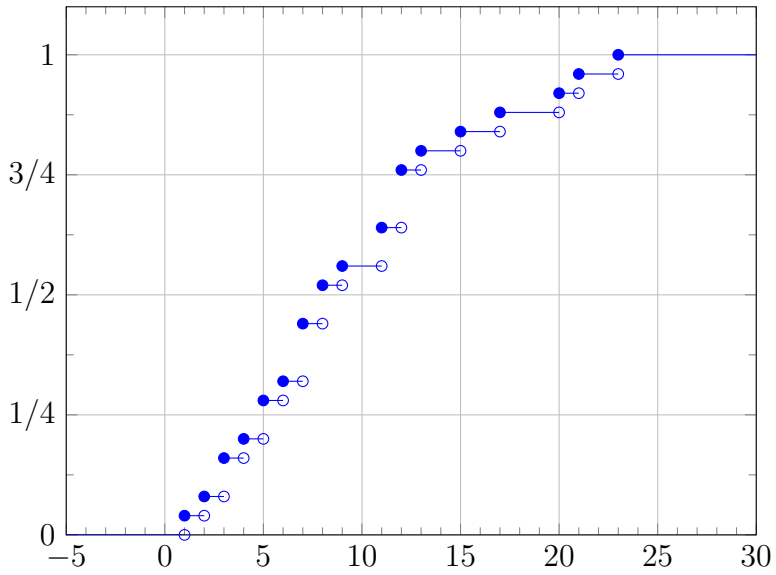
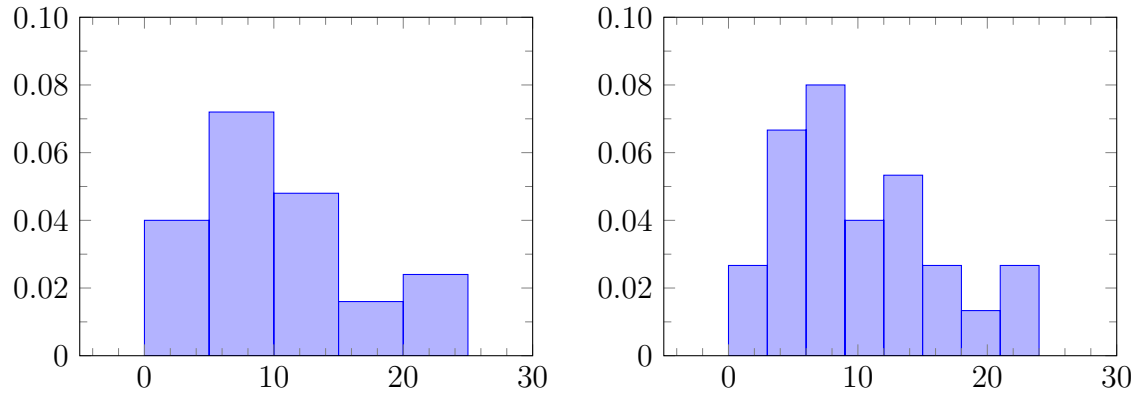


Figure 3.6: Empirical CDF of the sample from Example 19.



(a) Bin width = 5. (b) Bin width = 3.

Figure 3.7: Histograms of the sample from Example 19.

An approximate representation of the empirical distribution of the sample is given by a *histogram*: Consider a partition

$$[a_0, a_m[= \bigcup_{i=0}^{m-1} [a_i, a_{i+1}[$$

of $[a_0, a_m[$ with $m \in \mathbb{N}$ and $a_0 < a_1 < \ldots < a_m$ such that $x_1, \ldots, x_N \in [a_0, a_m[$. A subinterval $[a_i, a_{i+1}[$ is called *bin*. For each bin we define the height $h_i \in [0, \infty[$ such that

$$h_i \cdot (a_{i+1} - a_i) = \widehat{P}([a_i, a_{i+1}[) = \frac{|\{\ell \in \{1, \ldots, N\} \colon x_\ell \in [a_i, a_{i+1}[\}|}{N}.$$

A histogram is then a frequency distribution by means of m rectangles with width $a_{i+1} - a_i$ and height h_i for $i = 0, \ldots, m-1$, see Figure 3.7. By construction, the total area of the m rectangles is equal to one.

4 Limit Theorems

In the sequel let $(\Omega, \mathcal{A}, P)$ be a probability space and let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables.

Definition 1. $(X_n)_{n\in\mathbb{N}}$ is *independent* if $X_1, \ldots, X_n$ is independent for all $n \in \mathbb{N}$.

Law of Large Numbers

In the sequel let $(X_n)_{n\in\mathbb{N}}$ be i.i.d. random variables with expected value $\mathrm{E}(X_1)$ and variance $\mathrm{Var}(X_1)$. Moreover, let

$$\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

be the arithmetic mean of the first n random variables.

Example 2. (i) $\overline{X}_n$ can be a model for an average waiting time if $X_1 \sim \mathrm{Exp}(\lambda)$.

(ii) $\overline{X}_n$ is the relative frequency of successes if $X_1 \sim \mathrm{B}(1,p)$.

We aim at investigating the convergence of the sequence $(\overline{X}_n)_{n\in\mathbb{N}}$ of arithmetic means. Does, e.g., the sequence

$$\overline{X}_1(\omega), \overline{X}_2(\omega), \dots$$

of outcomes converge for all $\omega \in \Omega$?

Lemma 3. For $n \in \mathbb{N}$ we have

$$\mathrm{E}(\overline{X}_n) = \mathrm{E}(X_1) \qquad \text{and} \qquad \mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X_1)}{n}.$$

Proof. Let $n \in \mathbb{N}$. By linearity of the expected value, see Proposition II.7.5, we have

$$\mathrm{E}(\overline{X}_n) = \mathrm{E}\biggl(\frac{1}{n}\sum_{i=1}^n X_i\biggr) = \frac{1}{n}\cdot \mathrm{E}\biggl(\sum_{i=1}^n X_i\biggr) = \frac{1}{n}\cdot\sum_{i=1}^n \mathrm{E}(X_i) = \mathrm{E}(X_1),$$

where the last equality holds since $X_1, \dots, X_n$ are identically distributed. Similarly, we obtain

$$\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\biggl(\frac{1}{n}\sum_{i=1}^n X_i\biggr) = \frac{1}{n^2}\cdot \mathrm{Var}\biggl(\sum_{i=1}^n X_i\biggr) = \frac{1}{n^2}\cdot\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\mathrm{Var}(X_1)}{n}$$

due to Proposition II.7.12 and Proposition II.7.13 (Bienaymé's identity). □

Proposition 4 (Weak law of large numbers (WLLN))**.** We have

$$\forall\, \varepsilon > 0\colon \lim_{n\to\infty} P\bigl(\{|\overline{X}_n - \mathrm{E}(X_1)| \geq \varepsilon\}\bigr) = 0.$$

Notation: $\overline{X}_n \xrightarrow{P} \mathrm{E}(X_1)$. ("$\overline{X}_n$ converges in probability to $\mathrm{E}(X_1)$.")

Proof. Let $\varepsilon > 0$. Lemma 3 and Proposition II.7.15 (Chebyshev's inequality) show

$$\begin{aligned} 0 \leq P\bigl(\{|\overline{X}_n - \mathrm{E}(X_1)| \geq \varepsilon\}\bigr) &= P\bigl(\{|\overline{X}_n - \mathrm{E}(\overline{X}_n)| \geq \varepsilon\}\bigr) \\ &\leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} \\ &= \frac{\mathrm{Var}(X_1)}{\varepsilon^2}\cdot\frac{1}{n}. \end{aligned}$$

The claim follows from $\lim_{n\to\infty} 1/n = 0$. □

Example 5.

(i) Let $X_1 \sim \mathrm{N}(\mu,\sigma^2)$. Then we have[1] $\overline{X}_n \sim \mathrm{N}(\mu,\sigma^2/n)$, see Lemma 3. The corresponding PDF φ_n, given by

$$\varphi_n(x) = \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}}\cdot\exp\biggl(-\frac{n(x-\mu)^2}{2\sigma^2}\biggr), \qquad x \in \mathbb{R},$$

are illustrated in Figure 4.1.

[1] The sum of independent normally distributed random variables is normally distributed.

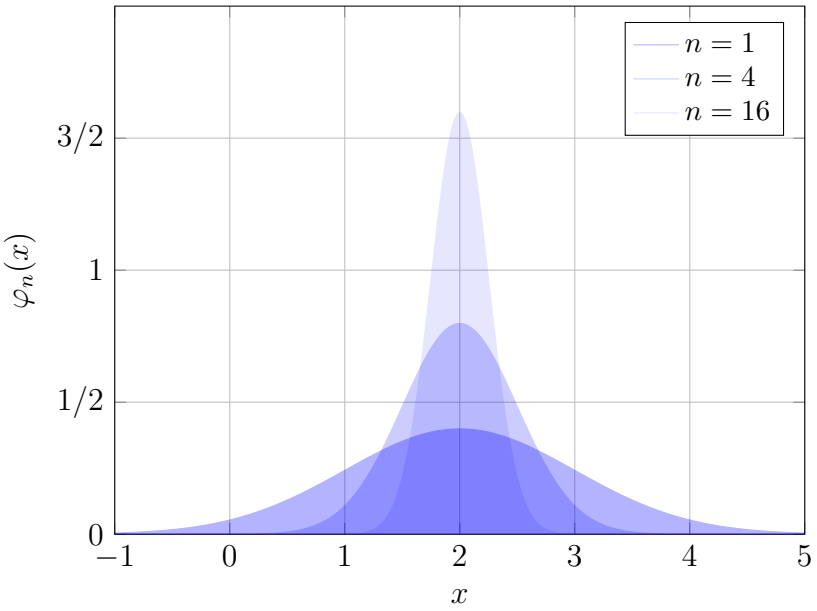


Figure 4.1: PDF of $\overline{X}_n \sim \mathrm{N}(\mu, \sigma^2/n)$ with $\mu = 2$ and $\sigma = 1$ for $n = 1, 4, 16$.

(ii) Let $X_1 \sim \mathrm{B}(1, p)$. Then we have $n \cdot \overline{X}_n = \sum_{i=1}^n X_i \sim \mathrm{B}(n, p)$ with

$$P(\{\overline{X}_n = k/n\}) = P(\{n \cdot \overline{X}_n = k\}) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

for $k \in \{0, \ldots, n\}$. The PMF of $\overline{X}_n$ are illustrated in Figure 4.2.

In both cases, the PDF (or PMF) of $\overline{X}_n$ become more concentrated around $\mathrm{E}(X_1)$ with increasing n.

Example 6. Let P_{X_1} be the uniform distribution on $\{1, \ldots, 6\}$, i.e., $\overline{X}_n$ is the arithmetic mean of the number of pips of n independently rolled fair dice. Clearly, we have

$$\mathrm{E}(\overline{X}_n) = \mathrm{E}(X_1) = 3.5.$$

We consider three realizations of $\overline{X}_n$:

n	1	2	3	4	5	...
$X_n(\omega_1)$	6	4	2	6	3	...
$\overline{X}_n(\omega_1)$	6	5	4	4.5	4.2	...
$X_n(\omega_2)$	5	3	6	5	1	...
$\overline{X}_n(\omega_2)$	5	4	4.667	4.75	4	...
$X_n(\omega_3)$	5	5	1	2	3	...
$\overline{X}_n(\omega_3)$	5	5	3.667	3.25	3.2	...

These realizations of $\overline{X}_n$ are illustrated in Figure 4.3 for $n = 1, \ldots, 100$. All depicted realizations seem to converge to $\mathrm{E}(X_1)$ for large n. Note that this phenomenon is not explained by Proposition 4 (WLLN).

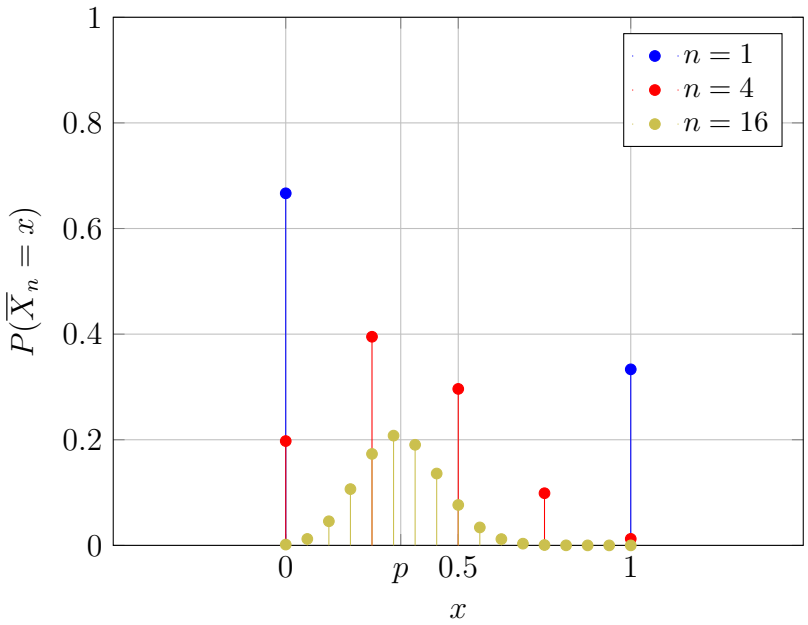


Figure 4.2: PMF of $\overline{X}_n$ with $n \cdot \overline{X}_n \sim \mathrm{B}(n, p)$ and $p = 1/3$ for $n = 1, 4, 16$.

Proposition 7 (Strong law of large numbers (SLLN))**.** We have

$$P\left(\left\{\omega \in \Omega\colon \lim_{n\to\infty} \left|\overline{X}_n(\omega) - \mathrm{E}(X_1)\right| = 0\right\}\right) = 1.$$

Notation: $\overline{X}_n \xrightarrow{\text{a.s.}} \mathrm{E}(X_1)$. ("$\overline{X}_n$ converges almost surely to $\mathrm{E}(X_1)$.")

Proof. See, e.g., [Georgii, Thm. 5.16, p. 129]. □

Remark 8.

(i) For random variables Y and $Y_1, Y_2, \ldots$ we have

$$Y_n \xrightarrow{\text{a.s.}} Y \;\Rightarrow\; Y_n \xrightarrow{P} Y.$$

The converse is not true in general.

(ii) The SLLN does not imply that $\lim_{n\to\infty} \overline{X}_n(\omega) = \mathrm{E}(X_1)$ for all $\omega \in \Omega$. In Example 6 there is the outcome $X_1(\omega) = X_2(\omega) = \ldots = 6$ such that $\overline{X}_n(\omega) = 6$ for all $n \in \mathbb{N}$.

For $B \in \mathfrak{B}(\mathbb{R})$ let $\mathbf{1}_B\colon \mathbb{R} \to \mathbb{R}$ be the *indicator function* defined by

$$\mathbf{1}_B(x) = \begin{cases} 1, & \text{if } x \in B, \\ 0, & \text{if } x \notin B. \end{cases}$$

Corollary 9 (SLLN for probabilities)**.** For all $B \in \mathfrak{B}(\mathbb{R})$ we have

$$P\left(\left\{\omega \in \Omega\colon \lim_{n\to\infty} \left|\frac{1}{n} \cdot \sum_{i=1}^{n} \mathbf{1}_B(X_i(\omega)) - p\right| = 0\right\}\right) = 1$$

where $p = P(\{X_1 \in B\})$.

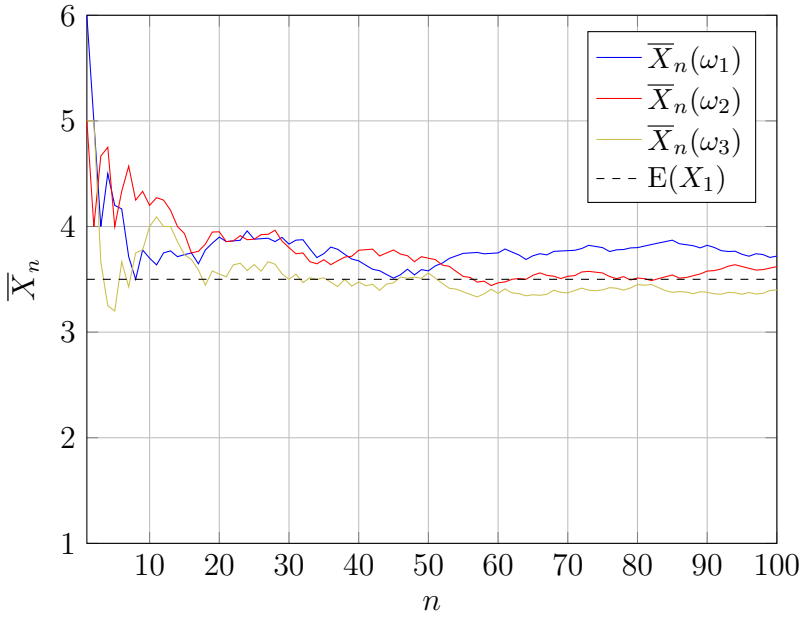


Figure 4.3: Three realizations of $\overline{X}_n$ from Example 6 (fair dice).

Proof. For all $B \in \mathfrak{B}(\mathbb{R})$ the sequence of random variables $\mathbf{1}_B(X_1), \mathbf{1}_B(X_2), \ldots$ is i.i.d. with $\mathbf{1}_B(X_1) \sim \mathrm{B}(1, p)$ where

$$p = P(\{X_1 \in B\}).$$

In particular, we have $\mathrm{E}(\mathbf{1}_B(X_1)) = p$. Apply Proposition 7 (SLLN). □

Example 10 (Continuation of Example 6)**.** Let P_{X_1} be the uniform distribution on $\{1, \ldots, 6\}$ and let $B = \{6\}$. Clearly, we have

$$\mathrm{E}(\mathbf{1}_B(X_1)) = P(\{X_1 = 6\}) = 1/6.$$

The same three realizations as in Example 6 are used to illustrate $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_B(X_i)$ in Figure 4.4.

Consider the empirical distribution function

$$\widehat{F}_n(\cdot, \omega) \colon \mathbb{R} \to [0, 1]$$

of the realization $X_1(\omega), \ldots, X_n(\omega)$.

Corollary 11 (SLLN for empirical CDF)**.** For all $x \in \mathbb{R}$ we have

$$P\left(\left\{\omega \in \Omega \colon \lim_{n\to\infty} \left|\widehat{F}_n(x, \omega) - F_{X_1}(x)\right| = 0\right\}\right) = 1.$$

Proof. Let $x \in \mathbb{R}$ and $B =]-\infty, x]$. Then we have

$$\widehat{F}_n(x, \omega) = \frac{|\{i \in \{1, \ldots, n\} \colon X_i(\omega) \leq x\}|}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbf{1}_B(X_i(\omega)).$$

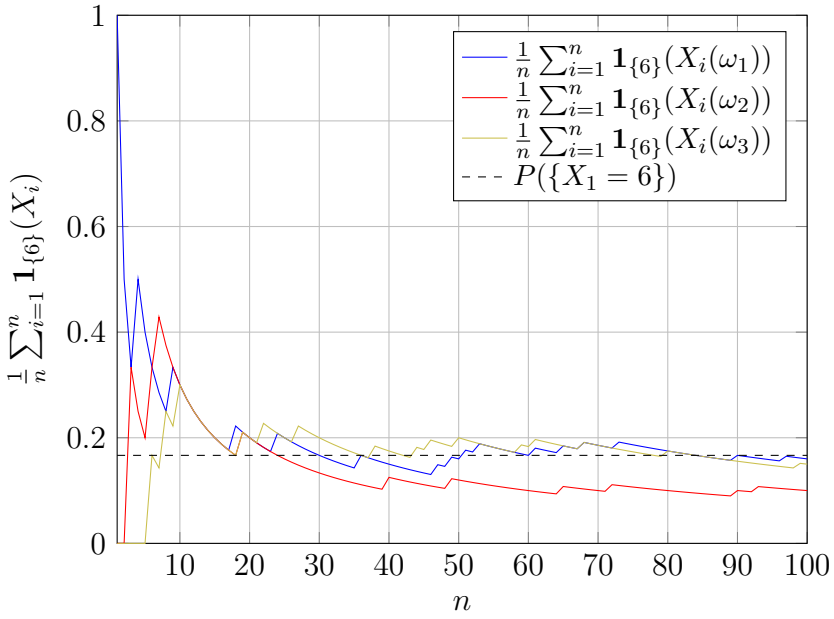


Figure 4.4: Three realizations of $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{6\}}(X_i)$ from Example 10 (fair dice).

Note that

$$\begin{aligned}\mathrm{E}(\mathbf{1}_B(X_1)) &= 0\cdot P(\{\mathbf{1}_B(X_1)=0\}) + 1\cdot P(\{\mathbf{1}_B(X_1)=1\})\\ &= P(\{X_1\in B\})\\ &= F_{X_1}(x)\end{aligned}$$

and apply Corollary 9 (SLLN for probabilities). □

Remark 12 (Fundamental theorem of statistics)**.** The pointwise convergence in Corollary 11 (SLLN for empirical CDF) can be strengthened to uniform convergence. In fact, the Glivenko-Cantelli theorem yields

$$P\left(\left\{\omega\in\Omega\colon \lim_{n\to\infty}\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x,\omega)-F_{X_1}(x)\right|=0\right\}\right)=1.$$

Remark 13 (Monte Carlo integration)**.** For the computation of $\mathrm{E}(X_1)$ we can use i.i.d. random variables $X_1,\dots,X_n$ as follows:

1. Generate a realization $(x_1,\dots,x_n)\in\mathbb{R}^n$ of $(X_1,\dots,X_n)$.
2. Approximate $\mathrm{E}(X_1)$ by the arithmetic mean $\frac{1}{n}\cdot\sum_{i=1}^{n}x_i$.

Numerical integration: Let $X_1\sim\mathrm{U}(0,1)$ with density f_{X_1} and let $g\colon[0,1]\to\mathbb{R}$ be integrable. Then we have

$$\mathrm{E}\big(g(X_1)\big)=\int_{-\infty}^{\infty}g(x)\cdot f_{X_1}(x)\,\mathrm{d}x=\int_0^1 g(x)\,\mathrm{d}x.$$

For an i.i.d. sequence $X_1,X_2,\dots$ the SLLN shows

$$\frac{1}{n}\cdot\sum_{i=1}^{n}g(X_i)\xrightarrow{\text{a.s.}}\int_0^1 g(x)\,\mathrm{d}x.$$

Matlab/Octave: `mc_int.m`.

Central Limit Theorem

The SLLN deals with the convergence of outcomes of random experiments. In the following we focus on the convergence of probabilities as in Proposition II.6.11 and Proposition II.6.13 (Poisson limit theorem).

In the sequel let $(X_n)_{n\in\mathbb{N}}$ be i.i.d. random variables with expected value $\mu = \mathrm{E}(X_1)$ and variance $\sigma^2 = \mathrm{Var}(X_1) > 0$. Moreover, let

$$\overline{X}_n^* = \sum_{i=1}^{n} \frac{X_i - \mu}{\sqrt{n}\cdot\sigma} = \frac{\left(\sum_{i=1}^n X_i\right) - n\mu}{\sqrt{n}\cdot\sigma} = \frac{\sqrt{n}}{\sigma}\cdot\left(\overline{X}_n - \mu\right)$$

be the *standardized sum* of the first n random variables.

Remark 14. Note that

$$\mathrm{E}(\overline{X}_n^*) = 0 \qquad \text{and} \qquad \mathrm{Var}(\overline{X}_n^*) = 1.$$

Moreover, we have

$$\overline{X}_n - \mu \xrightarrow{\text{a.s.}} 0 \qquad \text{and} \qquad \lim_{n\to\infty} \frac{\sqrt{n}}{\sigma} = \infty.$$

Proposition 15 (Central Limit Theorem (CLT))**.** We have

$$\lim_{n\to\infty} P(\{\overline{X}_n^* \in I\}) = \frac{1}{\sqrt{2\pi}}\cdot\int_I \exp(-x^2/2)\,\mathrm{d}x$$

for every interval $I \subseteq \mathbb{R}$. In particular, for all $x \in \mathbb{R}$ we have

$$\lim_{n\to\infty} P(\{\overline{X}_n^* \le x\}) = \Phi(x).$$

Terminology: $\overline{X}_n^*$ is asymptotically standard normally distributed.

Example 16. Let $X_1 \sim \mathrm{Exp}(1)$. Then $\sum_{i=1}^n X_i$ is absolutely continuous with PDF[2]

$$f(x) = \begin{cases} \frac{x^{n-1}}{(n-1)!}\cdot\exp(-x), & \text{if } x \ge 0,\\ 0, & \text{else.}\end{cases}$$

Note that $\mu = \sigma = 1$ in this case. By the change of variables formula the PDF of $\overline{X}_n^*$ is given by

$$f_{\overline{X}_n^*}(x) = \begin{cases} \sqrt{n}\cdot f(\sqrt{n}\cdot x + n), & \text{if } x \ge -\sqrt{n},\\ 0, & \text{else.}\end{cases}$$

The PDF of $\overline{X}_n^*$ are illustrated in Figure 4.5 for $n \in \{1, 2, 5, 100\}$.

[2]The sum of independent exponentially distributed random variables (with the same parameter $\lambda > 0$) admits an Erlang distribution.

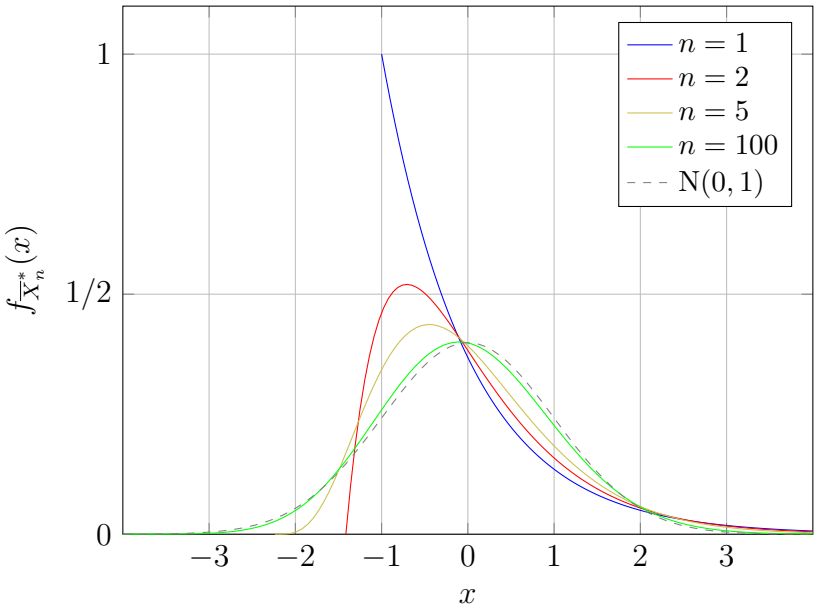


Figure 4.5: PDF $f_{\overline{X}_n^*}$ from Example 16 for $n \in \{1, 2, 5, 100\}$.

Example 17 (cf. Exercise 3.3 a)). Let $K \in \mathbb{N}$, $p, \alpha \in]0, 1[$, and $X_1 \sim \mathrm{B}(1, p)$.
Aim: Find $n > K$ such that

$$P\left(\left\{\sum_{i=1}^{n} X_i > K\right\}\right) \approx \alpha.$$

We have ("standardizing")

$$\left\{\sum_{i=1}^{n} X_i > K\right\} = \left\{\overline{X}_n^* > c_n\right\}$$

where

$$c_n = (K - n \cdot p)/\sqrt{n \cdot p \cdot (1 - p)}.$$

Proposition 15 (CLT) yields

$$P\left(\left\{\sum_{i=1}^{n} X_i > K\right\}\right) = P\left(\left\{\overline{X}_n^* > c_n\right\}\right) = 1 - P\left(\left\{\overline{X}_n^* \leq c_n\right\}\right) \approx 1 - \Phi(c_n).$$

Hence we choose $n > K$ such that $1 - \Phi(c_n) \approx \alpha$, i.e,

$$c_n \approx \Phi^{-1}(1 - \alpha).$$

For $K = 555$, $p = 0.98$, and $\alpha = 0.01$ we obtain $c_n \approx 2.33$ and $n \approx 558$.

Stochastic model	**Application**
random variable X on $(\Omega, \mathcal{A}, P)$	random experiment
$X_1, \ldots, X_n$ i.i.d.	n independent repetitions of a random experiment
realization $X_1(\omega), \ldots, X_n(\omega)$	data $x_1, \ldots, x_n$
probability $P(\{X \in A\})$	relative frequency $1/n \cdot \lvert\{i \in \{1, \ldots, n\} \colon x_i \in A\}\rvert$
expected value $\mathrm{E}(X)$	empirical mean $1/n \cdot \sum_{i=1}^{n} x_i$
CDF $F_X(x)$	empirical CDF $1/n \cdot \lvert\{i \in \{1, \ldots, n\} \colon x_i \leq x\}\rvert$

Table III.1: Stochastic modelling.

Chapter IV

Statistics – Parameter Estimation

We consider a random experiment modelled by a random variable $X\colon \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{A}, P)$ where the distribution

$$P_X = P_X^\vartheta$$

is known up to some parameter $\vartheta \in \Theta \subseteq \mathbb{R}^d$. We aim at determining the distribution P_X, i.e., estimating the unknown parameter ϑ. For this, a sample

$$x_1, \ldots, x_n \in \mathbb{R}$$

is available, which is assumed to be a realization of an i.i.d. sequence $X_1, \ldots, X_n$ having the same distribution as X, i.e.,

$$P_X = P_{X_1} = \ldots = P_{X_n}.$$

Example 1.

(i) Let $m \in \mathbb{N}$ and $X \sim \mathrm{B}(m, p)$ with unknown parameter

$$p = \vartheta \in \Theta = [0, 1].$$

Note that

$$\vartheta = \frac{1}{m} \cdot \mathrm{E}(X).$$

(ii) Let $N \in \mathbb{N}$, $m \in \{1, \ldots, N\}$, and $X \sim \mathrm{H}(N, K, m)$ with unknown parameter

$$K = \vartheta \in \Theta = \{0, \ldots, N\}.$$

Note that

$$\vartheta = \frac{N}{m} \cdot \mathrm{E}(X).$$

(iii) Let $X \sim \mathrm{Poi}(\lambda)$ with unknown parameter

$$\lambda = \vartheta \in \Theta =]0, \infty[\,.$$

Note that

$$\vartheta = \mathrm{E}(X).$$

(iv) Let $X \sim \mathrm{Exp}(\lambda)$ with unknown parameter

$$\lambda = \vartheta \in \Theta =]0, \infty[\,.$$

Note that

$$\vartheta = \frac{1}{\mathrm{E}(X)}.$$

(v) Let $\sigma > 0$ and $X \sim \mathrm{N}(\mu, \sigma^2)$ with unknown parameter

$$\mu = \vartheta \in \Theta = \mathbb{R}.$$

Note that

$$\vartheta = \mathrm{E}(X).$$

In all these examples the estimation of ϑ is basically an estimation of $\mathrm{E}(X)$.

Example 2. Let $X \sim \mathrm{N}(\mu, \sigma^2)$ with unknown parameter vector

$$(\mu, \sigma) = \vartheta \in \Theta = \mathbb{R} \times]0, \infty[\,.$$

Note that

$$\vartheta = \left(\mathrm{E}(X), \sqrt{\mathrm{Var}(X)}\right).$$

1 Point Estimation

Notation: We write E^ϑ and Var^ϑ for the expected value and the variance, respectively, if $P_X = P_X^\vartheta$.
In this section we focus on the estimation of the expected value $\mathrm{E}^\vartheta(X)$ and the variance $\mathrm{Var}^\vartheta(X)$ of a random variable X. The general case consists of the estimation of $\gamma(\vartheta)$ where

$$\gamma\colon \Theta \to \mathbb{R}$$

is a known function.

Definition 1. A function

$$g_n\colon \mathbb{R}^n \to \mathbb{R}$$

and the corresponding random variable $g_n(X_1, \ldots, X_n)$ are called *(point) estimator*. A function value $g_n(x_1, \ldots, x_n)$ is called *estimate*.

We consider different concepts to measure the quality of an estimator.

Definition 2.

(i) The *bias* of an estimator g_n is defined by

$$\mathrm{bias}^\vartheta(g_n) = \mathrm{E}^\vartheta\big(g_n(X_1, \ldots, X_n)\big) - \gamma(\vartheta).$$

Moreover, an estimator g_n is *unbiased* if

$$\forall\, \vartheta \in \Theta\colon\ \mathrm{bias}^\vartheta(g_n) = 0.$$

(ii) A sequence of estimators $g_1, g_2, \ldots$ is *strongly consistent* if

$$\forall\, \vartheta \in \Theta\colon P\Big(\Big\{\lim_{n\to\infty} g_n(X_1,\ldots,X_n) = \gamma(\vartheta)\Big\}\Big) = 1.$$

(iii) The *mean squared error* of an estimator g_n is defined by

$$\mathrm{mse}^\vartheta(g_n) = \mathrm{E}^\vartheta\left(\big(g_n(X_1,\ldots,X_n) - \gamma(\vartheta)\big)^2\right).$$

Lemma 3. For every estimator g_n and for all $\vartheta \in \Theta$ we have

$$\mathrm{mse}^\vartheta(g_n) = \mathrm{Var}^\vartheta\big(g_n(X_1,\ldots,X_n)\big) + \big(\mathrm{bias}^\vartheta(g_n)\big)^2.$$

Proof. We have

$$\begin{aligned}
\mathrm{mse}^\vartheta(g_n) &= \mathrm{E}^\vartheta\left(\big(g_n(X_1,\ldots,X_n) - \gamma(\vartheta)\big)^2\right)\\
&= \mathrm{E}^\vartheta\left(\big(g_n(X_1,\ldots,X_n)\big)^2 + \big(\gamma(\vartheta)\big)^2 - 2\cdot g_n(X_1,\ldots,X_n)\cdot\gamma(\vartheta)\right)\\
&= \mathrm{E}^\vartheta\left(\big(g_n(X_1,\ldots,X_n)\big)^2\right) + \big(\gamma(\vartheta)\big)^2 - 2\cdot\gamma(\vartheta)\cdot\mathrm{E}^\vartheta\big(g_n(X_1,\ldots,X_n)\big)\\
&= \mathrm{E}^\vartheta\left(\big(g_n(X_1,\ldots,X_n)\big)^2\right) - \Big(\mathrm{E}^\vartheta\big(g_n(X_1,\ldots,X_n)\big)\Big)^2\\
&\quad + \Big(\mathrm{E}^\vartheta\big(g_n(X_1,\ldots,X_n)\big)\Big)^2 + \big(\gamma(\vartheta)\big)^2 - 2\gamma(\vartheta)\cdot\mathrm{E}^\vartheta\big(g_n(X_1,\ldots,X_n)\big)\\
&= \mathrm{Var}^\vartheta\big(g_n(X_1,\ldots,X_n)\big) + \big(\mathrm{bias}^\vartheta(g_n)\big)^2.
\end{aligned}$$

□

Estimation of $\mathrm{E}^\vartheta(X)$

In the sequel let $\gamma(\vartheta) = \mathrm{E}^\vartheta(X)$ and let

$$g_n(x_1,\ldots,x_n) = \overline{x}_n = \frac{1}{n}\cdot\sum_{i=1}^n x_i$$

be the sample mean.

Proposition 4 (Expected value estimation by sample mean)**.** The sample mean g_n is an unbiased estimator and the sequence $(g_n)_{n\in\mathbb{N}}$ is strongly consistent. Moreover, we have

$$\mathrm{mse}^\vartheta(g_n) = \frac{1}{n}\cdot\mathrm{Var}^\vartheta(X).$$

Proof. By Lemma III.4.3 we have

$$\mathrm{E}^\vartheta\big(g_n(X_1,\ldots,X_n)\big) = \mathrm{E}^\vartheta(\overline{X}_n) = \mathrm{E}^\vartheta(X) = \gamma(\vartheta)$$

and hence $\mathrm{bias}^\vartheta(g_n) = 0$. This combined with Lemma 3 and Lemma III.4.3 shows

$$\mathrm{mse}^\vartheta(g_n) = \mathrm{Var}^\vartheta(\overline{X}_n) = \frac{1}{n}\cdot\mathrm{Var}^\vartheta(X).$$

Finally, Proposition III.4.7 (SLLN) yields for every $\vartheta \in \Theta$

$$g_n(X_1,\ldots,X_n) = \overline{X}_n \xrightarrow{\text{a.s.}} \mathrm{E}^\vartheta(X) = \gamma(\vartheta).$$

□

Example 5.

(i) Let $X \sim \mathrm{B}(1,p)$ with unknown $p = \vartheta \in \Theta = [0,1]$. Then we have

$$\mathrm{mse}^{\vartheta}(g_n) = \frac{1}{n} \cdot \vartheta \cdot (1-\vartheta).$$

(ii) Let $X \sim \mathrm{Exp}(\lambda)$ with unknown $\lambda = \vartheta \in \Theta =]0,\infty[$. Then we have

$$\mathrm{mse}^{\vartheta}(g_n) = \frac{1}{n} \cdot \frac{1}{\vartheta^2}.$$

Example 6. Let $X \sim \mathrm{N}(\mu,1)$ with unknown $\mu = \vartheta \in \Theta = \mathbb{R}$. A computer simulation with $\mu = 2$ and $n = 10$ generates the sample data

i	1	2	3	4	5	6	7	8	9	10
x_i	2.538	3.834	−0.259	2.862	2.319	0.692	1.566	2.343	5.578	4.769

with sample mean

$$\overline{x}_{10} = 2.624.$$

The corresponding empirical CDF, the CDF of $\mathrm{N}(\overline{x}_{10},1)$, and the CDF of $\mathrm{N}(2,1)$ are illustrated in Figure **??** TO DO.

Estimation of $\mathrm{Var}^{\vartheta}(X)$

In the sequel let $\gamma(\vartheta) = \mathrm{Var}^{\vartheta}(X)$, $n \geq 2$ and let

$$g_n(x_1,\ldots,x_n) = s_n^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \Big(x_i - \overline{x}_n\Big)^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} x_i^2 - n \cdot \big(\overline{x}_n\big)^2\right).$$

be the sample variance.

Proposition 7. The sample variance g_n is an unbiased estimator and the sequence $(g_n)_{n\in\mathbb{N}}$ is strongly consistent.

Proof. Use

$$\mathrm{E}^{\vartheta}\left(\sum_{i=1}^{n} X_i^2\right) = n \cdot \mathrm{E}^{\vartheta}(X^2) = n \cdot \Big(\mathrm{Var}^{\vartheta}(X) + \big(\mathrm{E}^{\vartheta}(X)\big)^2\Big)$$

and

$$\mathrm{E}^{\vartheta}\Big(\overline{X}_n^2\Big) = \mathrm{Var}^{\vartheta}\big(\overline{X}_n\big) + \big(\mathrm{E}^{\vartheta}(\overline{X}_n)\big)^2 = \frac{1}{n} \cdot \mathrm{Var}^{\vartheta}(X) + \big(\mathrm{E}^{\vartheta}(X)\big)^2.$$

Cf. Exercise 7.3. □

Maximum Likelihood Estimation

Definition 8 (MLE, discrete case)**.** Let P_X^ϑ be a discrete distribution for all $\vartheta \in \Theta$. The function $L(\,\cdot\,; x_1, \ldots, x_n)\colon \Theta \to \mathbb{R}$ given by

$$L(\vartheta; x_1, \ldots, x_n) = \prod_{i=1}^n P_X^\vartheta(\{x_i\})$$

is called *likelihood function* for the data $x_1, \ldots, x_n$.

An estimator g_n is a *maximum likelihood estimator (MLE)* of ϑ if $g_n(x_1, \ldots, x_n)$ is a global maximizer of $L(\,\cdot\,; x_1, \ldots, x_n)$.

Example 9. Let $P_X^\vartheta \sim \mathrm{B}(1, p)$ with $p = \vartheta \in \Theta = [0, 1]$. Put

$$k = |\{i \in \{1, \ldots, n\}\colon x_i = 1\}|.$$

Then we have

$$L(p; x_1, \ldots, x_n) = \prod_{i=1}^n \Big(p^{x_i} \cdot (1-p)^{1-x_i}\Big) = p^k \cdot (1-p)^{n-k}.$$

According to Exercise 3.1 b) the MLE of p is given by

$$g_n(x_1, \ldots, x_n) = \frac{k}{n} = \frac{1}{n} \cdot \sum_{i=1}^n x_i = \overline{x}_n.$$

Definition 10 (MLE, continuous case)**.** Let P_X^ϑ be an absolutely continuous distribution with density f_X^ϑ for all $\vartheta \in \Theta$. The function $L(\,\cdot\,; x_1, \ldots, x_n)\colon \Theta \to \mathbb{R}$ given by

$$L(\vartheta; x_1, \ldots, x_n) = \prod_{i=1}^n f_X^\vartheta(x_i)$$

is called *likelihood function* for the data $x_1, \ldots, x_n$.

An estimator g_n is a *maximum likelihood estimator (MLE)* of ϑ if $g_n(x_1, \ldots, x_n)$ is a global maximizer of $L(\,\cdot\,; x_1, \ldots, x_n)$.

Example 11. Let $P_X^\vartheta \sim \mathrm{Exp}(\lambda)$ with $\lambda = \vartheta \in \Theta =]0, \infty[$. The MLE of λ is given by

$$g_n(x_1, \ldots, x_n) = \frac{n}{\sum_{i=1}^n x_i} = (\overline{x}_n)^{-1},$$

see Exercise 7.5.

2 Interval Estimation

Put $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{X} = (X_1, \ldots, X_n)$. For the estimation of $\gamma(\vartheta)$ we now aim at constructing an interval $[a_n(\mathbf{x}), b_n(\mathbf{x})]$ such that

$$\gamma(\vartheta) \in [a_n(\mathbf{X}), b_n(\mathbf{X})] \quad \text{“with high probability”}$$

rather than constructing a point $g_n(\mathbf{x})$ as before.

Chapter V

Statistics – Hypothesis Tests

Appendix A

Combinatorics

Appendix B

Tables

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Table B.1: Values of the cumulative distribution function Φ of a standard normal distribution, i.e., $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, \mathrm{d}y$.