

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

I	Introduction	1
II	Probability Theory – Discrete Case	2
1	Discrete Probability Spaces	2
2	Probability Mass Functions	4
3	Conditional Probability and Independence	6
4	Random Variables	8
5	Empirical Distribution	12
6	Special Discrete Distributions	14
7	Expected Value and Variance	21
III	Probability Theory – General Case	27
1	Absolutely Continuous Distributions	28
2	Basic Concepts	31
3	Cumulative Distribution Function and Quantiles	32
4	Limit Theorems	38
IV	Statistics – Parameter Estimation	47
1	Point Estimation	48
2	Interval Estimation	52
V	Statistics – Hypothesis Tests	58
1	Tests for Normal Distribution	61
2	Goodness-of-fit Tests	64
A	Combinatorics	65
B	Tables	67

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 girls (1), $N - k$ boys (0).

The relative frequencies are given by

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

and serve as a metric of the data (descriptive statistics).

Naive answers to the above questions read:

- (i) The sought probability is given by $\hat{p}(1)$.
- (ii) Yes if and only if $\hat{p}(0) > \hat{p}(1)$.

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 \emptyset . 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

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

respectively.

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

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

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

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

The events $A_1, A_2, A_3, \dots \subseteq \Omega$ are *pairwise disjoint* if $A_i \cap A_j = \emptyset$ 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 $A_1, \dots, A_n \subseteq \Omega$ we have

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

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

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

Definition 7. The set

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

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

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

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

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

Definition 9. A function $P : \mathcal{P}(\Omega) \rightarrow \mathbb{R}$ 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 $A_1, A_2, \dots \in \mathcal{P}(\Omega)$ we have

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad (\sigma\text{-additivity})$$

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) $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$ (*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) $P(A^c) = 1 - P(A)$,
- (v) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

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

$$A \cup B = A \cup B \cup \emptyset \cup \emptyset \cup \dots$$

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

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

Since $P(A \cup B) \in [0, 1]$, we have $P(\emptyset) = 0$.

ad (ii)-(v): See **Exercise 1.1**. □

2 Probability Mass Functions

Definition 1. Let P be a probability measure on Ω . The function $p: \Omega \rightarrow \mathbb{R}$ 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 $p: \Omega \rightarrow \mathbb{R}$ satisfies

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

- (iii) Every function $p: \Omega \rightarrow \mathbb{R}$ 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 occurred.

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 $B \neq \emptyset$ 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

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

where

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

Then $Q: \mathcal{P}(\Omega) \rightarrow \mathbb{R}$ given by

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

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

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

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 $\{B_1, \dots, B_n\} \subseteq \mathcal{P}(\Omega)$ be a partition¹ of Ω with $P(B_i) > 0$ for all $i = 1, \dots, n$.

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

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

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

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

Proof. ad (i): Since $\{B_1, \dots, B_n\}$ 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 $A \cap B_1, \dots, A \cap B_n \subseteq \Omega$. Hence we get

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

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

Since $P(B_i) > 0$ for all $i = 1, \dots, n$, the conditional probabilities $P(A | B_i)$ are well-defined such that $P(A \cap B_i) = P(A | B_i) \cdot P(B_i)$.

ad (ii): If $P(A) > 0$, the conditional probability $P(B_k | A)$ is well-defined and we have

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

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 $X: \Omega \rightarrow \mathfrak{X}$ 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 $X_1, X_2: \Omega \rightarrow \mathfrak{X}$ 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 $S: \Omega \rightarrow \mathfrak{S}$ given by

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

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

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

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) $p_X: \mathfrak{X} \rightarrow \mathbb{R}$ 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.

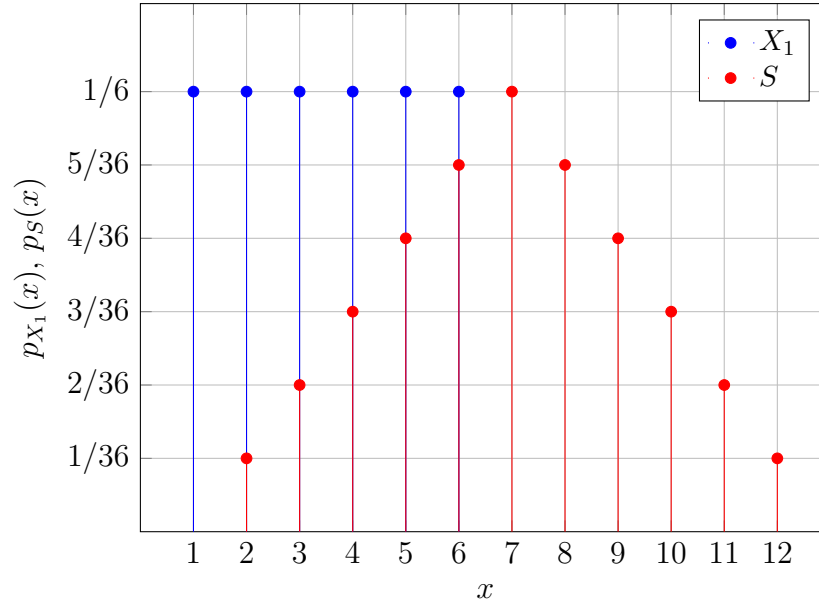


Figure 4.1: Probability mass functions p_{X_1} and p_S from Example 4.

Definition 8. Two random variables $X_1, X_2: \Omega \rightarrow \mathfrak{X}$ 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 $\{1, \dots, 6\}$. Hence X_1 and X_2 are identically distributed.

Proposition 10 (Criterion for identical distributions). $X_1, X_2: \Omega \rightarrow \mathfrak{X}$ 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)

$$X = (X_1, \dots, X_n): \Omega \rightarrow \mathfrak{X}$$

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

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

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

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

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

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

In the sequel we use

$$\begin{aligned} \{X = x\} &= \{\omega \in \Omega: X(\omega) = x\}, \\ \{X \in A\} &= \{\omega \in \Omega: X(\omega) \in A\}, \end{aligned}$$

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

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

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

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

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

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

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

Proof. “ \Rightarrow ”: Take $A_i = \{x_i\}$.

“ \Leftarrow ”: 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). [...] \square

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

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

$X_1 \backslash X_2$	0	\dots	k	Σ
0	$p_{0,0}$	\dots	$p_{0,k}$	$p_{0,\bullet}$
\vdots	\vdots	\ddots	\vdots	\vdots
k	$p_{k,0}$	\dots	$p_{k,k}$	$p_{k,\bullet}$
Σ	$p_{\bullet,0}$	\dots	$p_{\bullet,k}$	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 $X: \Omega \rightarrow \mathfrak{X}$ 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)

$$x_1, \dots, x_N \in \mathfrak{X} \quad (3)$$

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

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

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

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

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

Definition 3. \hat{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)).

$X_1 \backslash X_2$	0	\dots	k	Σ
0	$\hat{p}_{0,0}$	\dots	$\hat{p}_{0,k}$	$\hat{p}_{0,\bullet}$
\vdots	\vdots	\ddots	\vdots	\vdots
k	$\hat{p}_{k,0}$	\dots	$\hat{p}_{k,k}$	$\hat{p}_{k,\bullet}$
Σ	$\hat{p}_{\bullet,0}$	\dots	$\hat{p}_{\bullet,k}$	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)

$$X = (X_1, \dots, X_n): \Omega \rightarrow \mathfrak{X}$$

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 $(X_1, X_2): \Omega \rightarrow \mathfrak{X}$ and $\mathfrak{X} = \{0, \dots, k\}^2$ with $k \in \mathbb{N}$. Put

$$\hat{p}_{i,j} = \frac{|\{\ell \in \{1, \dots, N\}: x_\ell = (i, j)\}|}{N}$$

and

$$\hat{p}_{i,\bullet} = \frac{|\{\ell \in \{1, \dots, N\}: x_{\ell,1} = i\}|}{N}, \quad \hat{p}_{\bullet,j} = \frac{|\{\ell \in \{1, \dots, N\}: x_{\ell,2} = j\}|}{N}$$

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

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

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

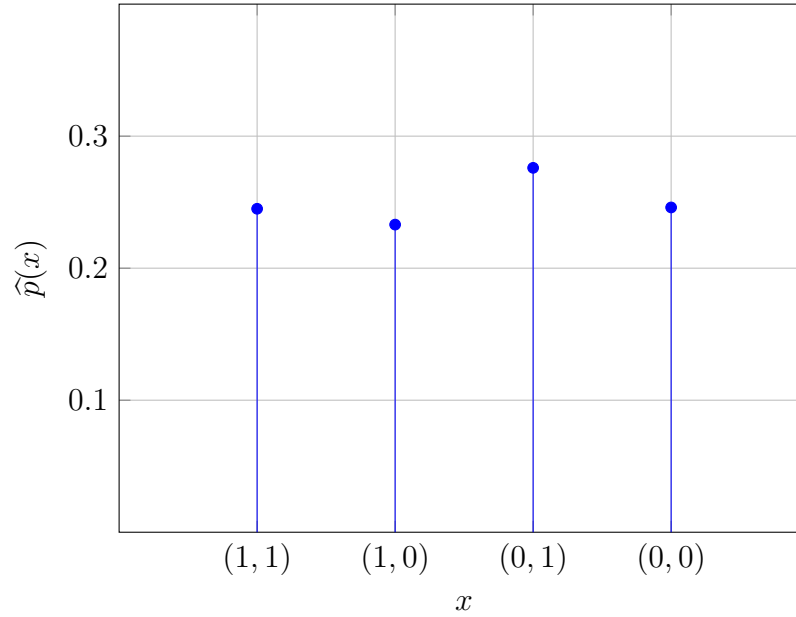
$$\hat{p}_{i,j} \approx \hat{p}_{i,\bullet} \cdot \hat{p}_{\bullet,j}$$

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

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.

$X_1 \backslash X_2$	0	1	Σ
0	0.246	0.276	0.522
1	0.233	0.245	0.478
Σ	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.

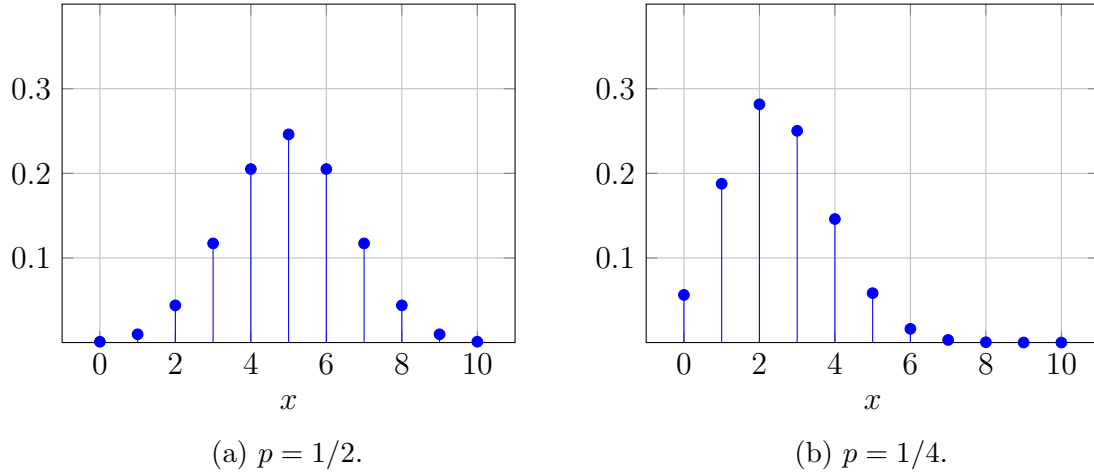
Binomial Distribution

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, \dots, X_n satisfying

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

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

Figure 6.1: Probability mass functions of $X \sim B(10, p)$ 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

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

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: $X \sim B(n, p)$.

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

Example 3. The probability mass functions of $X \sim B(n, p)$ 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 $X \sim B(n, p)$ and $Y \sim B(m, p)$ for $m, n \in \mathbb{N}$ and $p \in [0, 1]$. Then we have

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

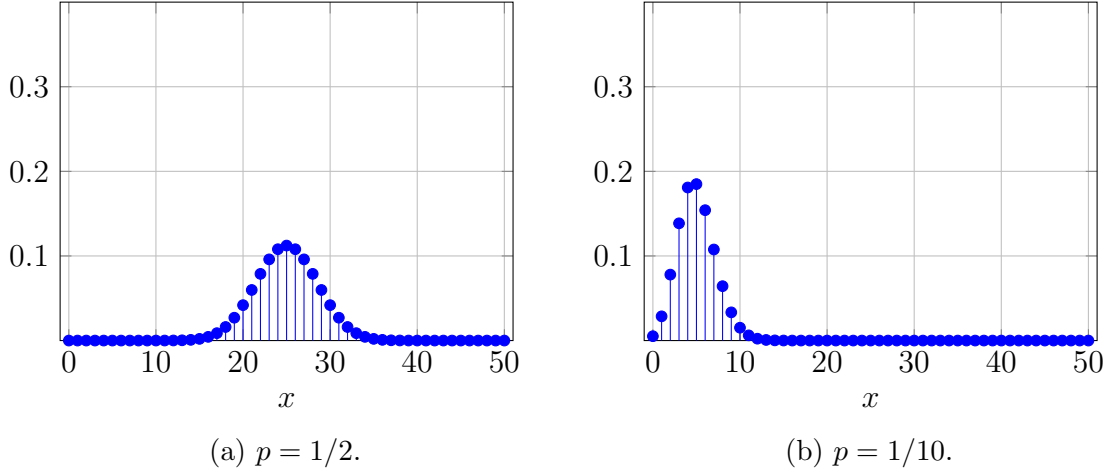


Figure 6.2: Probability mass functions of $X \sim B(50, p)$ with $p = 1/2$ and $p = 1/10$.

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

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

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 $S_j \sim B(n, p_j)$.

(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: $X \sim M(n, p_0, \dots, p_{m-1})$.

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 $n \leq N$ and $K \leq N$,
- (ii) the uniform distribution P on

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

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.5 shows

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

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

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

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 $n \leq N$ and $K \leq N$ if (5) holds for all $k \in \mathbb{N}_0$ with (4).

Notation: $X \sim H(N, K, n)$.

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 $X_N \sim H(N, K_N, n)$ for $N \in \mathbb{N}$ such that

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

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

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

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

Proof. By assumption, we have $\lim_{N \rightarrow \infty} K_N = \infty$. 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

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

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

Example 12. The probability mass functions of $X \sim H(100, 20, 10)$ and $Y \sim B(10, 1/5)$ 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 $X_n \sim B(n, p_n)$ for $n \in \mathbb{N}$ such that

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

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

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

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

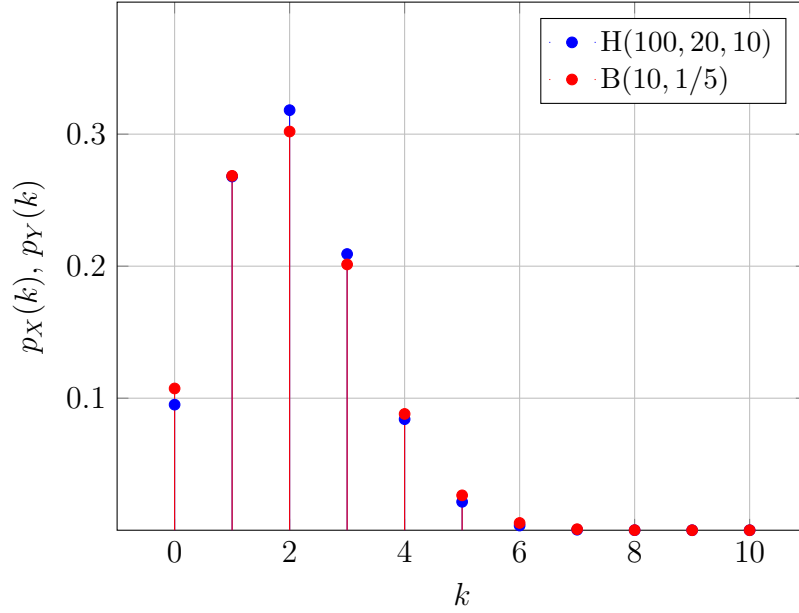


Figure 6.3: Probability mass functions of $X \sim H(100, 20, 10)$ and $Y \sim B(10, 1/5)$.

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

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

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

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

Hence we get

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

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

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

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.

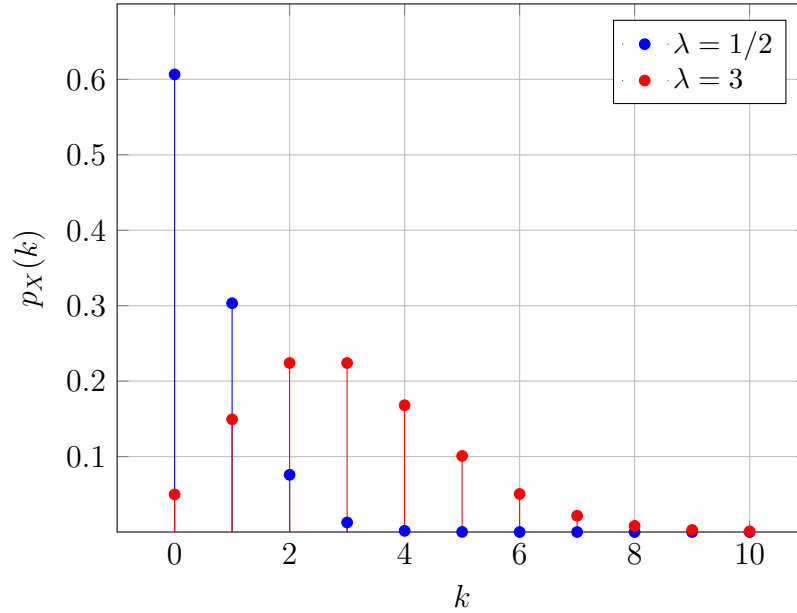


Figure 6.4: Probability mass functions of $X \sim \text{Poi}(\lambda)$ with $\lambda = 1/2$ and $\lambda = 3$.

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

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

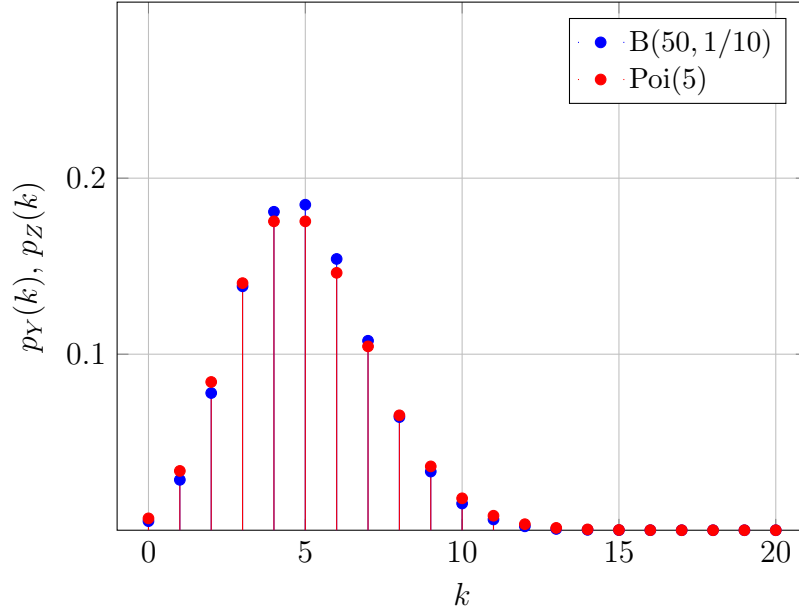


Figure 6.5: Probability mass functions of $Y \sim B(50, 1/10)$ and $Z \sim \text{Poi}(5)$.

and we put

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

otherwise, i.e., if $\{k \in \{1, \dots, n\} : X_k(\omega) = 1\} \neq \emptyset$. 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 $\lim_{n \rightarrow \infty} P(\{T_n = 0\}) = 0$.

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

7 Expected Value and Variance

In the sequel let $(\Omega, \mathcal{P}(\Omega), P)$ be a discrete probability space and let $X: \Omega \rightarrow \mathfrak{X}$ 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

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

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

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

In particular, for $\Omega = \{0, \dots, 36\}$ 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”)

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

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

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

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

$$E(c \cdot X + Y) = c \cdot E(X) + E(Y). \quad (\text{linearity})$$

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

$$E(X) \leq E(Y). \quad (\text{monotonicity})$$

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

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

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

$$E(X) = \sum_{\omega \in \Omega} X(\omega) \cdot P(\{\omega\}) \leq \sum_{\omega \in \Omega} Y(\omega) \cdot P(\{\omega\}) = E(Y). \quad \square$$

Proposition 6. For every function $g: \mathbb{R} \rightarrow \mathbb{R}$ we have

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

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

Proposition 7 (Expected value of special discrete distributions).

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

$$E(X) = n \cdot p.$$

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

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

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

$$E(X) = \lambda.$$

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

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

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

$$E(X \cdot Y) = E(X) \cdot E(Y).$$

Variance

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

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

In this case $\sigma_X = \sqrt{\text{Var}(X)}$ 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 $g(x) = (x - E(X))^2$ shows

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

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

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

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

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

Proposition 12.

(i) We have

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

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

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

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

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

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

$$\begin{aligned}\text{Var}(a \cdot X + b) &= E\left((aX + b - E(aX + b))^2\right) \\ &= E\left((aX + b - aE(X) - b)^2\right) \\ &= E\left(a^2 \cdot (X - E(X))^2\right) = a^2 \cdot \text{Var}(X).\end{aligned}\quad \square$$

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

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

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

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

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

Proposition 14 (Variance of special discrete distributions).

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

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

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

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

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

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

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

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

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

$$P(\{|X - E(X)| \geq \varepsilon\}) \leq \frac{\text{Var}(X)}{\varepsilon^2}.$$

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

$$A = \{|X - E(X)| \geq \varepsilon\} = \{\omega \in \Omega : (X(\omega) - E(X))^2 \geq \varepsilon^2\}$$

we obtain

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

Sample Mean and Sample Variance

In the sequel let

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

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

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

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

Remark 17. We have

$$\bar{x}_N = \frac{1}{N} \cdot \sum_{x \in \mathfrak{X}} x \cdot |\{\ell \in \{1, \dots, N\} : x_\ell = x\}| = \sum_{x \in \mathfrak{X}} x \cdot \hat{p}(x).$$

Hence \bar{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

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

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

Remark 19. We have

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

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 $X: \Omega \rightarrow \mathfrak{X}$ 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) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$,
- (iii) $A_1, A_2, \dots \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

A probability space is then a triple (Ω, \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 $P: \mathcal{A} \rightarrow \mathbb{R}$.

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 $X: \Omega \rightarrow \mathbb{R}$ is a so-called *measurable* function, i.e., we require

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

This property ensures

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

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 $f: \mathbb{R} \rightarrow \mathbb{R}$ is a (*probability*) *density* (*PDF*) if

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

Example 2. The following functions $f_1, f_2, f_3: \mathbb{R} \rightarrow \mathbb{R}$ 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 (Ω, \mathcal{A}, P) be a probability space and let $X: \Omega \rightarrow \mathbb{R}$ 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

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

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

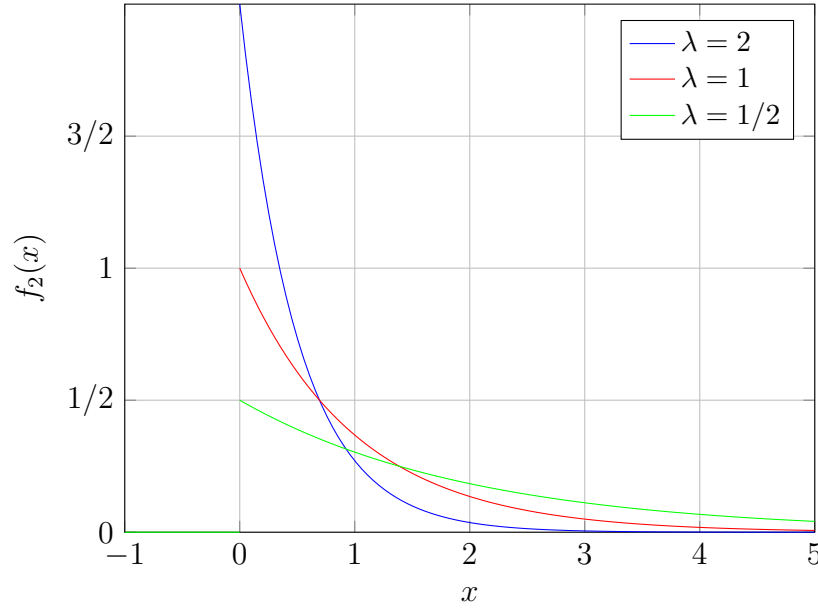


Figure 1.1: Probability density f_2 from Example 2.(ii) with $\lambda \in \{1/2, 1, 2\}$.

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

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

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: $X \sim U(a, b)$.
- (ii) X is *exponentially distributed* with parameter $\lambda \in]0, \infty[$ if $f_X = f_2$ according to Example 2.(ii).
Notation: $X \sim \text{Exp}(\lambda)$.
- (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: $X \sim N(\mu, \sigma^2)$.

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

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

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 .

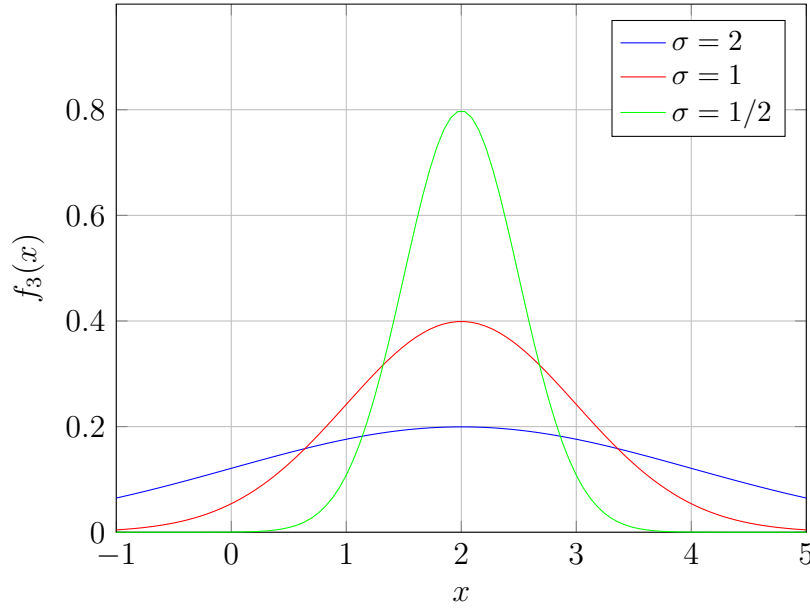


Figure 1.2: Probability density f_3 from Example 2.(iii) with $\mu = 2$ and $\sigma \in \{1/2, 1, 2\}$.

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

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

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

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

In this case $\sigma_X = \sqrt{\text{Var}(X)}$ 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 $X \sim U(a, b)$ with $a < b$ we have

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

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

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

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

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

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

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

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 (Ω, \mathcal{A}, P) be a probability space and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable.

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

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

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 $X_1, X_2: \Omega \rightarrow \mathbb{R}$ 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 $X_1, X_2: \Omega \rightarrow \mathbb{R}$ 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 $X_1, \dots, X_n: \Omega \rightarrow \mathbb{R}$ are *independent* if for all $A_1, \dots, A_n \in \mathfrak{B}(\mathbb{R})$ 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 (Ω, \mathcal{A}, P) be a probability space and let $X, Y: \Omega \rightarrow \mathbb{R}$ be random variables.

Definition 1. The function $F_X: \mathbb{R} \rightarrow [0, 1]$ 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

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

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).

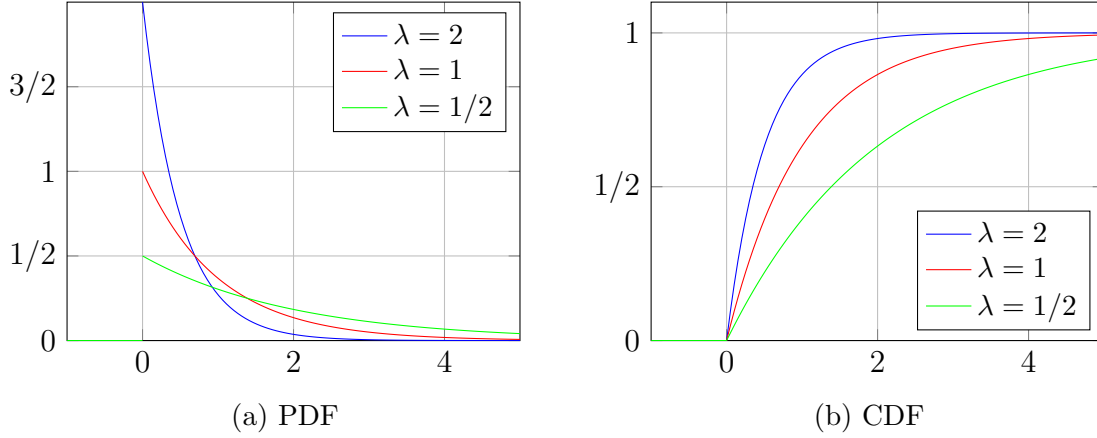
(i) For $Y \sim U(a, b)$ with $a, b \in \mathbb{R}$ and $a < b$ we have

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

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

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

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

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

(ii) For $Y \sim \text{Exp}(\lambda)$ 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 $X \sim \text{Exp}(1)$ 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 $Z \sim \text{Exp}(\lambda)$ due to Proposition 3.

The PDF and the CDF of $Y \sim \text{Exp}(\lambda)$ 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

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

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

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

for $A \in \mathfrak{B}(\mathbb{R})$, and hence $Z \sim N(\mu, \sigma^2)$. 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

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

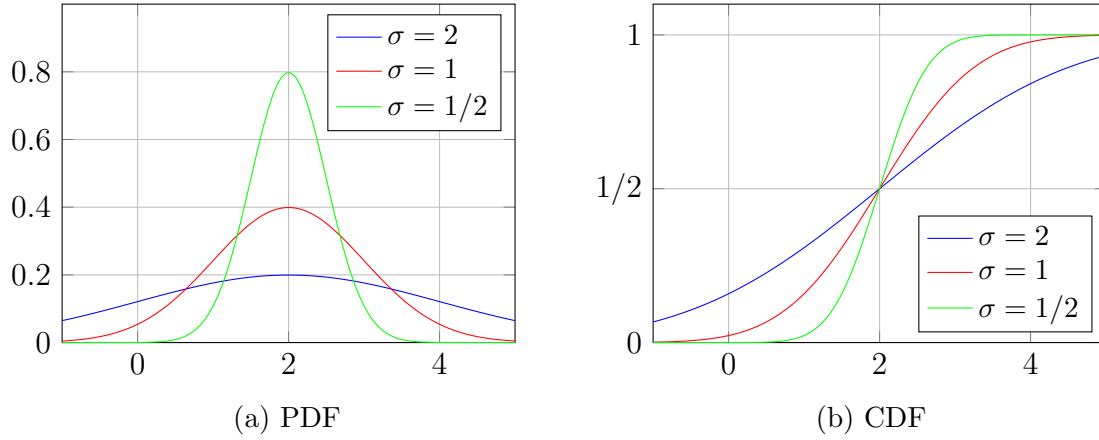


Figure 3.2: PDF and CDF of $Z \sim N(\mu, \sigma^2)$ with $\mu = 2$ and $\sigma \in \{1/2, 1, 2\}$ from Example 6.(iii).

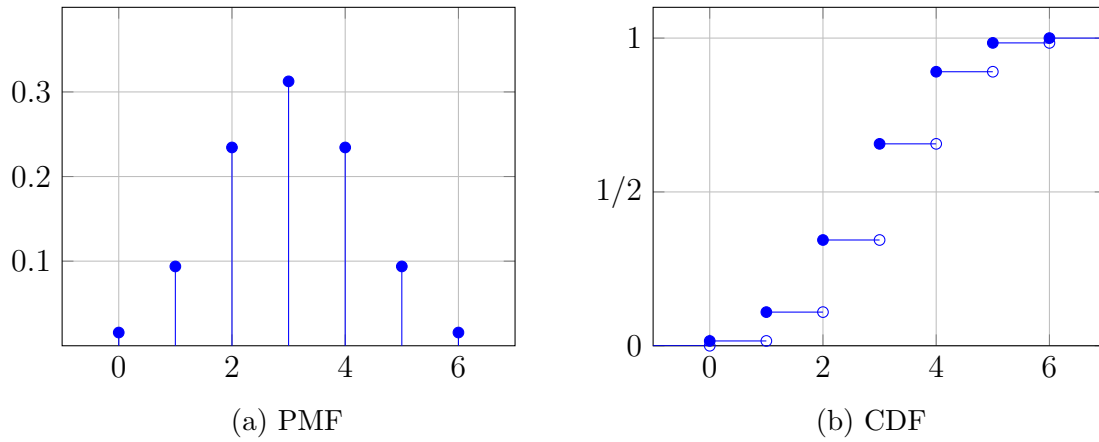


Figure 3.3: PMF and CDF of $X \sim B(6, 1/2)$.

The PDF and the CDF of $Z \sim N(\mu, \sigma^2)$ 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

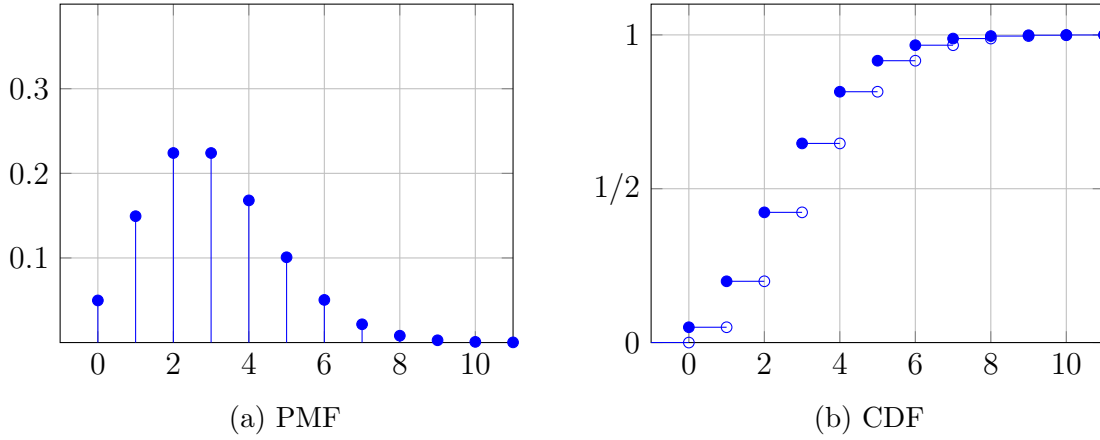
$$F_X(x) = \sum_{i=0}^k p_X(i)$$

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

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

Proposition 9 (Properties of CDFs, general case). Every cumulative distribution function $F_X: \mathbb{R} \rightarrow [0, 1]$ 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 $x \in \mathbb{R}$,

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

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

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

$$F_X(q) = p. \quad (1)$$

Example 10. Let $X \sim \text{Exp}(\lambda)$ 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 $X \sim \text{B}(2, 1/2)$. 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

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

The p -quantile for $p = 1/2$ is called *median* of F_X (of P_X , of X).

Notation: median $\mathbf{m}(X)$.

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

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

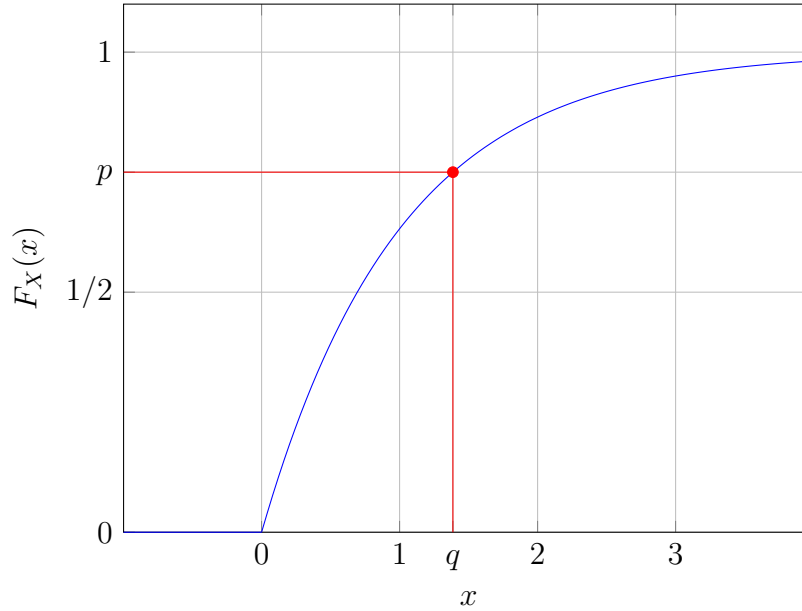


Figure 3.5: CDF of $X \sim \text{Exp}(1)$ with $p \in]0, 1[$ and $q \in \mathbb{R}$ such that $F_X(q) = p$, see Example 10.

Example 14 (Continuation of Example 11). Let $X \sim \text{B}(2, 1/2)$. 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

$$|E(X) - m(X)| \leq \sigma_X.$$

Empirical Distribution Function

As in Section II.5 we consider a sample

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

Definition 16. The *empirical distribution function* $\hat{F}: \mathbb{R} \rightarrow [0, 1]$ is defined by

$$\hat{F}(x) = \frac{|\{\ell \in \{1, \dots, N\} : x_\ell \leq x\}|}{N}.$$

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

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

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

Remark 18. Consider the sample mean \bar{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 \bar{x}_N or \tilde{x}_N attain an arbitrary value?

Answer:

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

(ii) Note that

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

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 \bar{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.

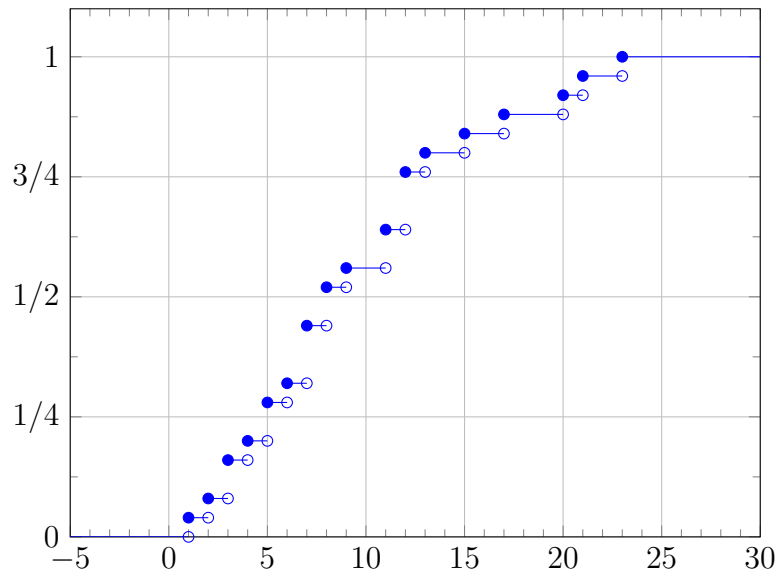


Figure 3.6: Empirical CDF of the sample from Example 19.

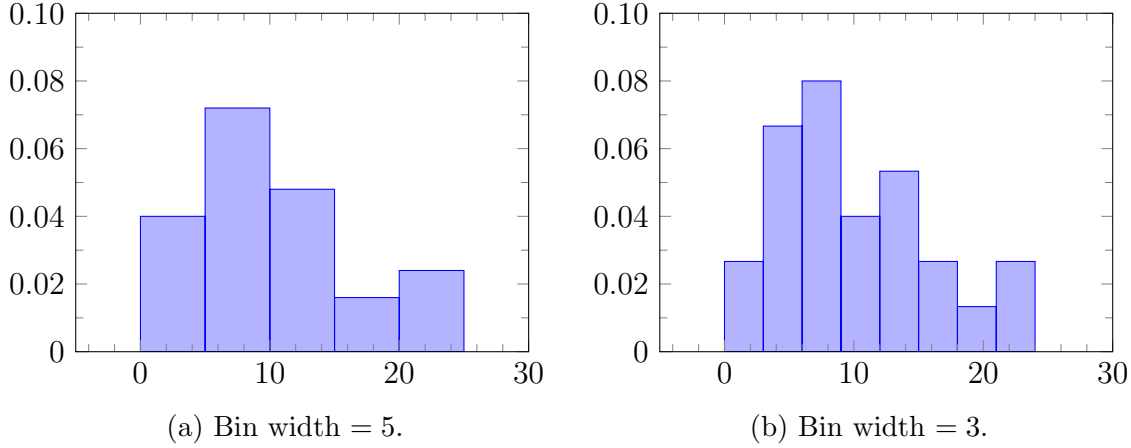


Figure 3.7: Histograms of the sample from Example 19.

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

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

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

4 Limit Theorems

In the sequel let (Ω, \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, \dots, 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 $E(X_1)$ and variance $\text{Var}(X_1)$. Moreover, let

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

be the arithmetic mean of the first n random variables.

- Example 2.** (i) \bar{X}_n can be a model for an average waiting time if $X_1 \sim \text{Exp}(\lambda)$.
(ii) \bar{X}_n is the relative frequency of successes if $X_1 \sim \text{B}(1, p)$.

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

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

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

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

$$\mathbb{E}(\bar{X}_n) = \mathbb{E}(X_1) \quad \text{and} \quad \text{Var}(\bar{X}_n) = \frac{\text{Var}(X_1)}{n}.$$

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

$$\mathbb{E}(\bar{X}_n) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \cdot \mathbb{E}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \cdot \sum_{i=1}^n \mathbb{E}(X_i) = \mathbb{E}(X_1),$$

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

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \cdot \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \cdot \sum_{i=1}^n \text{Var}(X_i) = \frac{\text{Var}(X_1)}{n}$$

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

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

$$\forall \varepsilon > 0: \lim_{n \rightarrow \infty} P(\{|\bar{X}_n - \mathbb{E}(X_1)| \geq \varepsilon\}) = 0.$$

Notation: $\bar{X}_n \xrightarrow{P} \mathbb{E}(X_1)$. (“ \bar{X}_n converges in probability to $\mathbb{E}(X_1)$.”)

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

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

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

Example 5.

- (i) Let $X_1 \sim \text{N}(\mu, \sigma^2)$. Then we have¹ $\bar{X}_n \sim \text{N}(\mu, \sigma^2/n)$, see Lemma 3. The corresponding PDF φ_n , given by

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

are illustrated in Figure 4.1.

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

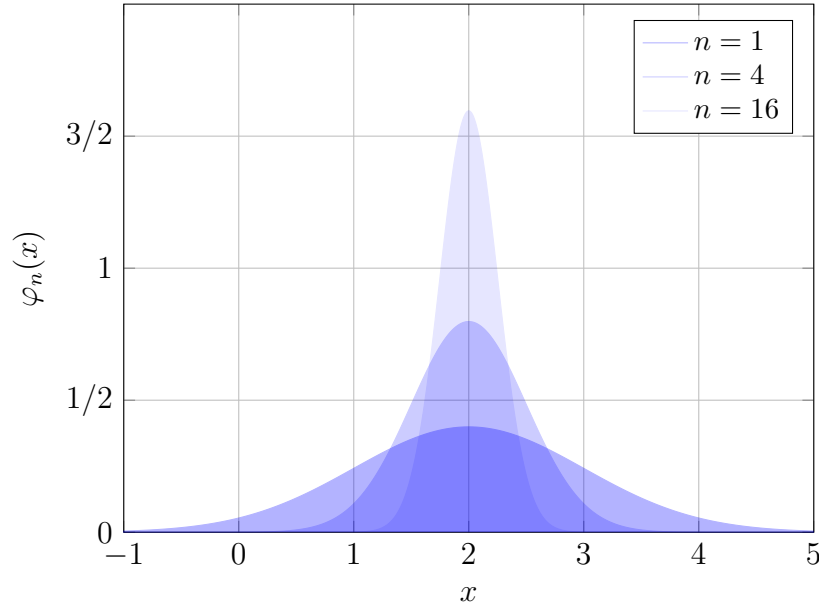


Figure 4.1: PDF of $\bar{X}_n \sim N(\mu, \sigma^2/n)$ with $\mu = 2$ and $\sigma = 1$ for $n = 1, 4, 16$.

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

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

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

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

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

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

We consider three realizations of \bar{X}_n :

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

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

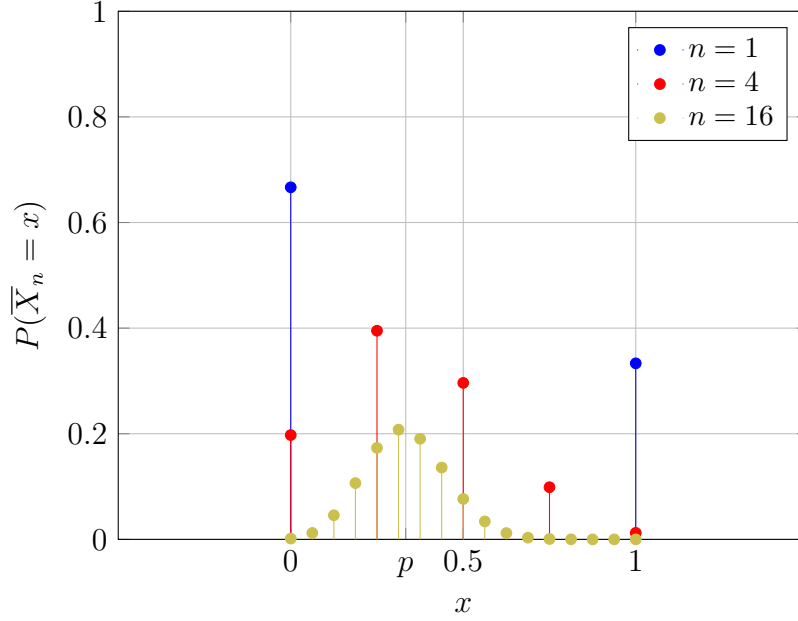


Figure 4.2: PMF of \bar{X}_n with $n \cdot \bar{X}_n \sim B(n, p)$ and $p = 1/3$ for $n = 1, 4, 16$.

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

$$P\left(\left\{\omega \in \Omega: \lim_{n \rightarrow \infty} |\bar{X}_n(\omega) - E(X_1)| = 0\right\}\right) = 1.$$

Notation: $\bar{X}_n \xrightarrow{\text{a.s.}} E(X_1)$. (“ \bar{X}_n converges almost surely to $E(X_1)$.”)

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

Remark 8.

- (i) For random variables Y and Y_1, Y_2, \dots 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 $\lim_{n \rightarrow \infty} \bar{X}_n(\omega) = E(X_1)$ for all $\omega \in \Omega$. In Example 6 there is the outcome $X_1(\omega) = X_2(\omega) = \dots = 6$ such that $\bar{X}_n(\omega) = 6$ for all $n \in \mathbb{N}$.

For $B \in \mathfrak{B}(\mathbb{R})$ let $\mathbf{1}_B: \mathbb{R} \rightarrow \mathbb{R}$ 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

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

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

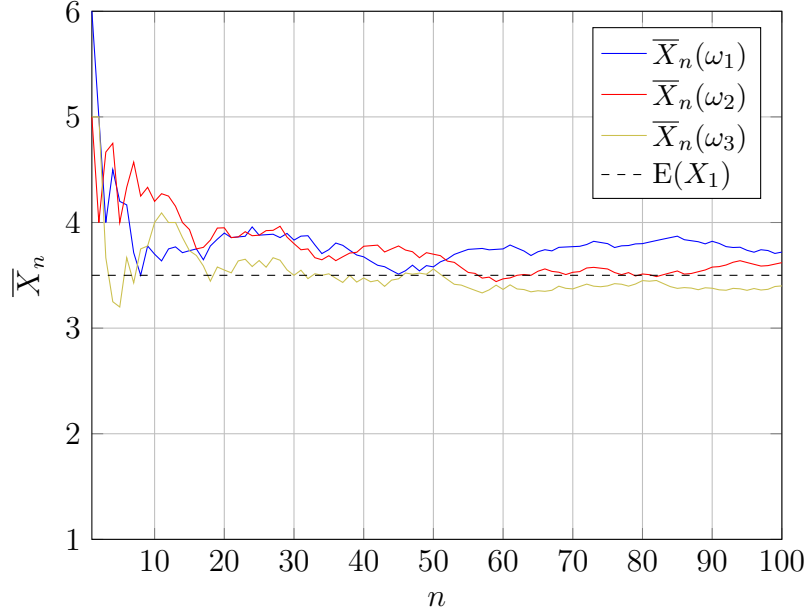


Figure 4.3: Three realizations of \bar{X}_n from Example 6 (fair dice).

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

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

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

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

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

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

$$\hat{F}_n(\cdot, \omega): \mathbb{R} \rightarrow [0, 1]$$

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

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

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

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

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

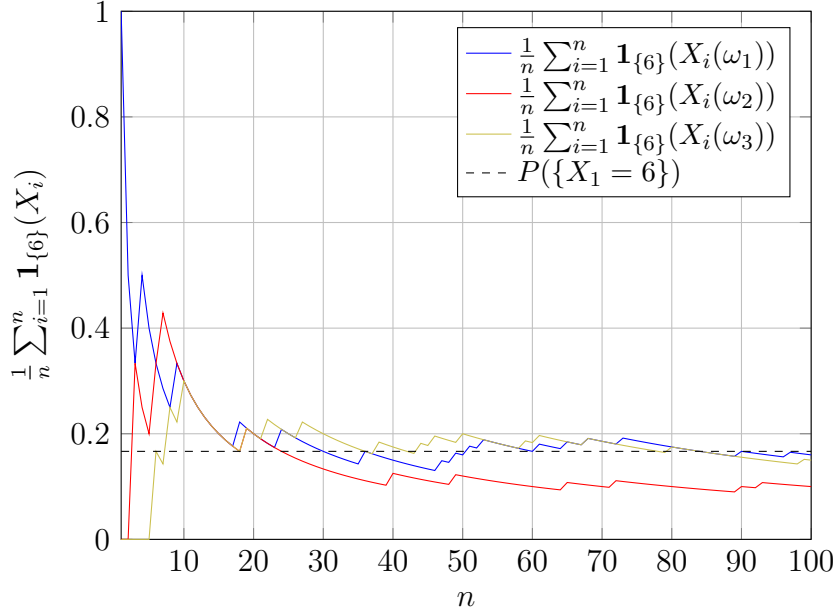


Figure 4.4: Three realizations of $\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{6\}}(X_i)$ from Example 10 (fair dice).

Note that

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

and apply Corollary 9 (SLLN for probabilities). \square

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

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

Remark 13 (Monte Carlo integration). For the computation of $E(X_1)$ 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 $E(X_1)$ by the arithmetic mean $\frac{1}{n} \cdot \sum_{i=1}^n x_i$.

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

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

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

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

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 $\mu = E(X_1)$ and variance $\sigma^2 = \text{Var}(X_1) > 0$. Moreover, let

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

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

Remark 14. Note that

$$E(\bar{X}_n^*) = 0 \quad \text{and} \quad \text{Var}(\bar{X}_n^*) = 1.$$

Moreover, we have

$$\bar{X}_n - \mu \xrightarrow{\text{a.s.}} 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{\sqrt{n}}{\sigma} = \infty.$$

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

$$\lim_{n \rightarrow \infty} P(\{\bar{X}_n^* \in I\}) = \frac{1}{\sqrt{2\pi}} \cdot \int_I \exp(-x^2/2) dx$$

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

$$\lim_{n \rightarrow \infty} P(\{\bar{X}_n^* \leq x\}) = \Phi(x).$$

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

Example 16. Let $X_1 \sim \text{Exp}(1)$. Then $\sum_{i=1}^n X_i$ is absolutely continuous with PDF²

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

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

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

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

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

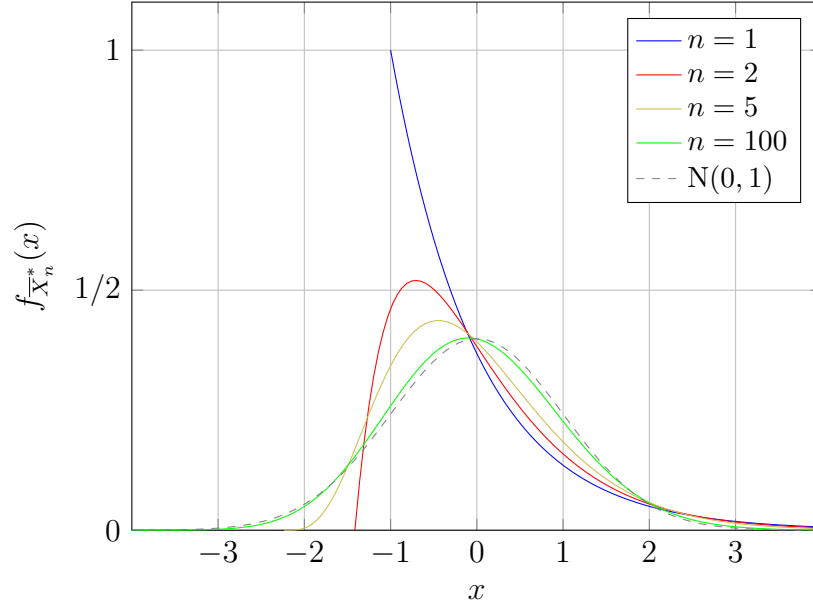


Figure 4.5: PDF $f_{\bar{X}_n^*}$ from Example 16 for $n \in \{1, 2, 5, 100\}$.

Example 17 (cf. Exercise 3.3 a)). Let $K \in \mathbb{N}$, $p, \alpha \in]0, 1[$, and $X_1 \sim B(1, p)$.

Aim: Find $n > K$ such that

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

We have (“standardizing”)

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

where

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

Proposition 15 (CLT) yields

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

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 (Ω, \mathcal{A}, P)	random experiment
X_1, \dots, X_n i.i.d.	n independent repetitions of a random experiment
realization $X_1(\omega), \dots, X_n(\omega)$	data x_1, \dots, x_n
probability $P(\{X \in A\})$	relative frequency $1/n \cdot \{i \in \{1, \dots, n\} : x_i \in A\} $
expected value $E(X)$	empirical mean $1/n \cdot \sum_{i=1}^n x_i$
CDF $F_X(x)$	empirical CDF $1/n \cdot \{i \in \{1, \dots, n\} : x_i \leq x\} $

Table III.1: Stochastic modelling.

Chapter IV

Statistics – Parameter Estimation

We consider a random experiment modelled by a random variable $X: \Omega \rightarrow \mathbb{R}$ on a probability space (Ω, \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

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

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

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

Example 1.

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

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

Note that

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

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

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

Note that

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

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

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

Note that

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

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

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

Note that

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

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

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

Note that

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

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

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

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

Note that

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

1 Point Estimation

Notation: We write \mathbb{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 $\mathbb{E}^\vartheta(X)$ and the variance $\text{Var}^\vartheta(X)$ of a random variable X . The general case consists of the estimation of $\gamma(\vartheta)$ where

$$\gamma: \Theta \rightarrow \mathbb{R}$$

is a known function.

Definition 1. A function

$$g_n: \mathbb{R}^n \rightarrow \mathbb{R}$$

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

$$\text{bias}^\vartheta(g_n) = \mathbb{E}^\vartheta(g_n(X_1, \dots, X_n)) - \gamma(\vartheta).$$

Moreover, an estimator g_n is *unbiased* if

$$\forall \vartheta \in \Theta: \text{bias}^\vartheta(g_n) = 0.$$

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

$$\forall \vartheta \in \Theta: P\left(\left\{\lim_{n \rightarrow \infty} g_n(X_1, \dots, X_n) = \gamma(\vartheta)\right\}\right) = 1.$$

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

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

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

$$\text{mse}^\vartheta(g_n) = \text{Var}^\vartheta(g_n(X_1, \dots, X_n)) + (\text{bias}^\vartheta(g_n))^2.$$

Proof. We have

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

Estimation of $E^\vartheta(X)$

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

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

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

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

Proof. By Lemma III.4.3 we have

$$E^\vartheta(g_n(X_1, \dots, X_n)) = E^\vartheta(\bar{X}_n) = E^\vartheta(X) = \gamma(\vartheta)$$

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

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

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

$$g_n(X_1, \dots, X_n) = \bar{X}_n \xrightarrow{\text{a.s.}} E^\vartheta(X) = \gamma(\vartheta). \quad \square$$

Example 5.

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

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

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

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

Example 6. Let $X \sim N(\mu, 1)$ 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

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

The corresponding empirical CDF, the CDF of $N(\bar{x}_{10}, 1)$, and the CDF of $N(2, 1)$ are illustrated in Figure 1.1.

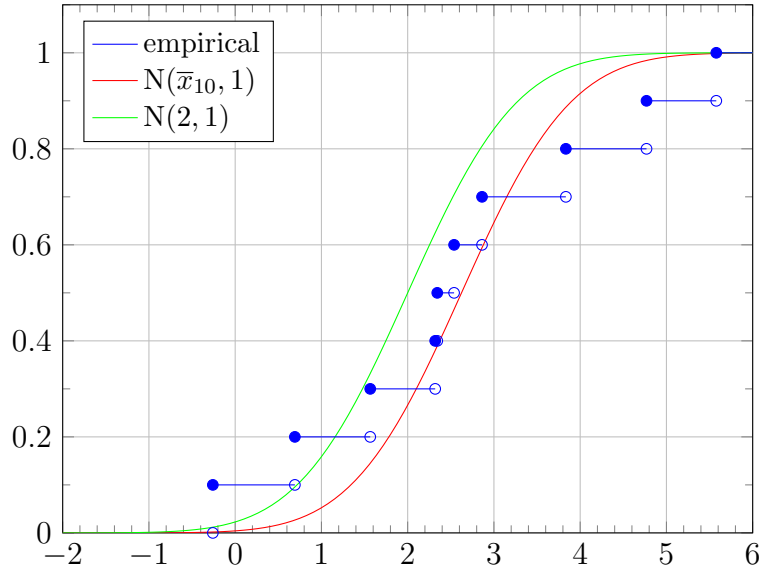


Figure 1.1: CDF (empirical, $N(\bar{x}_{10}, 1)$, $N(2, 1)$) from Example 6.

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

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

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

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

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

and

$$\mathbb{E}^\vartheta \left(\overline{X}_n^2 \right) = \text{Var}^\vartheta(\overline{X}_n) + (\mathbb{E}^\vartheta(\overline{X}_n))^2 = \frac{1}{n} \cdot \text{Var}^\vartheta(X) + (\mathbb{E}^\vartheta(X))^2.$$

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 $L(\cdot; x_1, \dots, x_n): \Theta \rightarrow \mathbb{R}$ given by

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

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

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

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

$$k = |\{i \in \{1, \dots, n\} : x_i = 1\}|.$$

Then we have

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

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

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

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 $L(\cdot; x_1, \dots, x_n): \Theta \rightarrow \mathbb{R}$ given by

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

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

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

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

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

see Exercise 7.5.

2 Interval Estimation

Put $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{X} = (X_1, \dots, X_n)$. 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.

Definition 1. Let $\alpha \in]0, 1[$. Two functions

$$a_n, b_n: \mathbb{R}^n \rightarrow \mathbb{R}$$

form a *confidence interval* with confidence level $1 - \alpha$ if

$$\forall \vartheta \in \Theta: P^\vartheta(\{\gamma(\vartheta) \in [a_n(\mathbf{X}), b_n(\mathbf{X})]\}) \geq 1 - \alpha.$$

In practice, one is interested in confidence intervals of “small” length $b_n(\mathbf{X}) - a_n(\mathbf{X})$.

Confidence Intervals for Normal Distribution

In the sequel let $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma \in]0, \infty[$, and let $\alpha \in]0, 1[$. We consider the following three cases:

(i) $\mu = \vartheta \in \Theta = \mathbb{R}$ and

$$\gamma(\vartheta) = \vartheta = E^\vartheta(X),$$

i.e., confidence interval for μ with known σ ,

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

$$\gamma(\vartheta) = \mu = E^\vartheta(X),$$

i.e., confidence interval for μ with unknown σ ,

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

$$\gamma(\vartheta) = \sigma^2 = \text{Var}^\vartheta(X),$$

i.e., confidence interval for σ^2 with unknown μ .

Proposition 2 (CI for μ, σ known). Let X_1, \dots, X_n be i.i.d. with $X_1 \sim N(\mu, \sigma^2)$.

(i) We have

$$\frac{\sqrt{n}}{\sigma} \cdot (\bar{X}_n - \mu) \sim N(0, 1).$$

(ii) A confidence interval for μ with confidence level $1 - \alpha$ is given by

$$a_n(\mathbf{x}) = \bar{x}_n - q_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \quad b_n(\mathbf{x}) = \bar{x}_n + q_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

where

$$q_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2).$$

Proof. ad (i): The sum of independent normally distributed random variables is normally distributed, cf. Example III.4.5. Moreover, we have

$$E(\bar{X}_n - \mu) = 0 \quad \text{and} \quad \text{Var}\left(\frac{\sqrt{n}}{\sigma} \cdot (\bar{X}_n - \mu)\right) = \frac{n}{\sigma^2} \cdot \text{Var}(\bar{X}_n) = 1,$$

cf. Lemma III.4.3

ad (ii): For $q_{\alpha/2} = \Phi^{-1}(\alpha/2)$ and $q_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ we obtain

$$\begin{aligned} 1 - \alpha &= P\left(\left\{\frac{\sqrt{n}}{\sigma} \cdot (\bar{X}_n - \mu) \in [q_{\alpha/2}, q_{1-\alpha/2}]\right\}\right) \\ &= P\left(\left\{q_{\alpha/2} \leq \frac{\sqrt{n}}{\sigma} \cdot (\bar{X}_n - \mu)\right\} \cap \left\{\frac{\sqrt{n}}{\sigma} \cdot (\bar{X}_n - \mu) \leq q_{1-\alpha/2}\right\}\right) \\ &= P\left(\left\{\mu \leq \bar{X}_n - q_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right\} \cap \left\{\bar{X}_n - q_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu\right\}\right) \\ &= P\left(\left\{\mu \in \left[\bar{X}_n - q_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X}_n - q_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]\right\}\right) \end{aligned}$$

and hence

$$P^\mu\left(\left\{\mu \in \left[\bar{X}_n - q_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X}_n - q_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]\right\}\right) \geq 1 - \alpha.$$

Finally, by symmetry of the density of the standard normal distribution we have

$$-q_{\alpha/2} = q_{1-\alpha/2}$$

see Figure 2.1. □

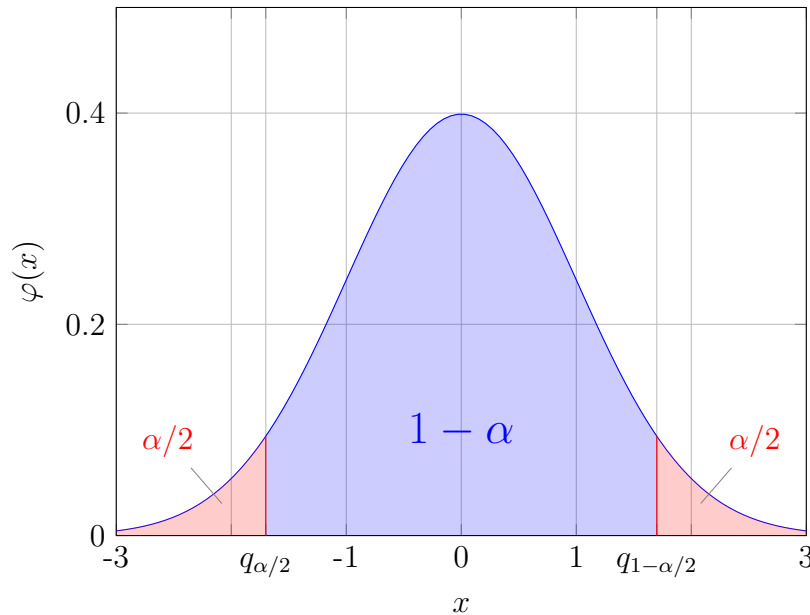


Figure 2.1: PDF of $N(0, 1)$ illustrating the situation from the proof of Proposition 2.

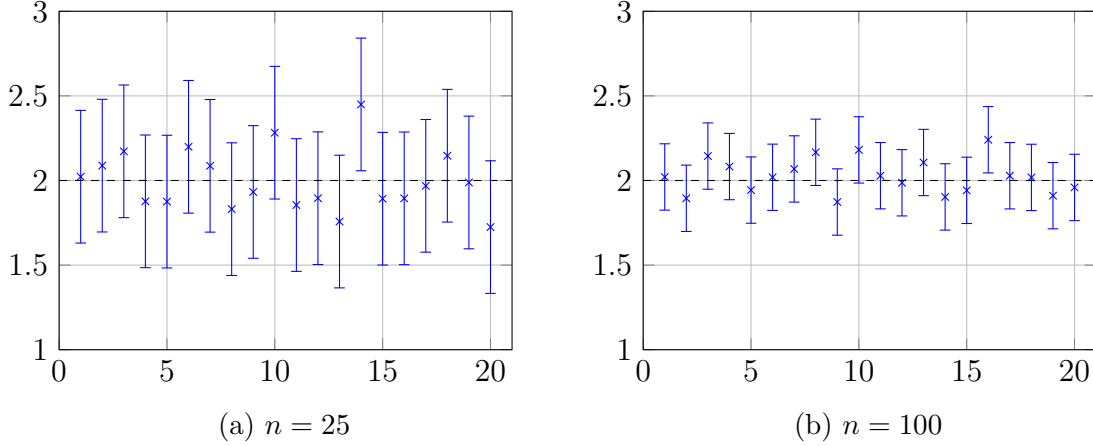


Figure 2.2: Computer simulation of 20 independently generated confidence intervals for $X \sim N(\mu, 1)$ with $\alpha = 0.05$, $\mu = 2$, and $n \in \{25, 100\}$, see Example 3.

Example 3. Figure 2.2 shows a computer simulation of 20 independently generated confidence intervals for $X \sim N(\mu, 1)$ with $\alpha = 0.05$, $\mu = 2$, and $n \in \{25, 100\}$ according to Proposition 2. The relative share of confidence intervals not containing the true value $\mu = 2$ is given by $1/20$ for both values of n .

In the sequel let $n \geq 2$ and

$$S_n = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X}_n)^2}.$$

Definition 4. Let $n \in \mathbb{N}$. A random variable X has a *t-distribution* (or *Student's t-distribution*) with n degrees of freedom if it is absolutely continuous with density $f_n: \mathbb{R} \rightarrow \mathbb{R}$ given by

$$f_n(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi} \cdot \Gamma(n/2)} \cdot \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}, \quad x \in \mathbb{R},$$

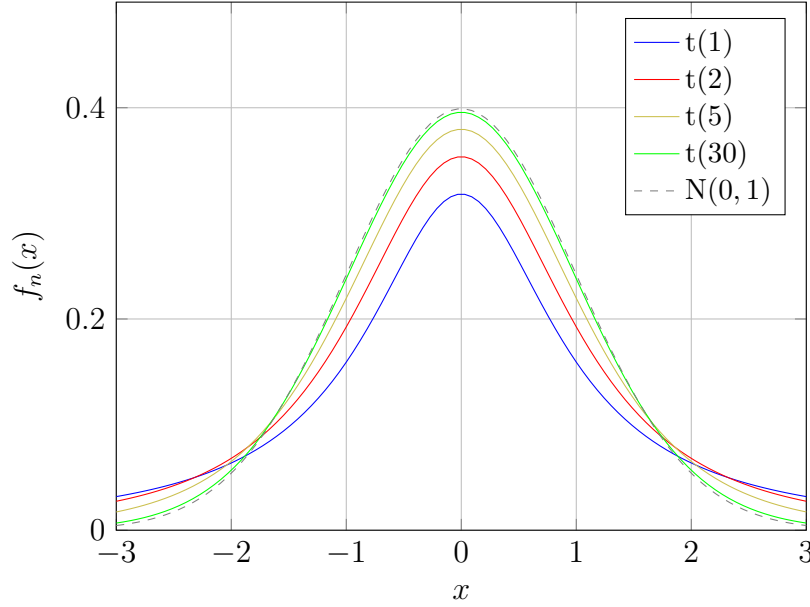
where Γ denotes the gamma function.

Notation: $X \sim t(n)$.

Remark 5 (Properties of t-distribution). For all $n \in \mathbb{N}$ the PDF of $t(n)$ is an even function and strictly positive. Hence the CDF of $t(n)$, denoted by F_n , satisfies

- (i) $F_n(-x) = 1 - F_n(x)$ for all $x \in \mathbb{R}$,
- (ii) for all $p \in]0, 1[$ there exists a unique $x \in \mathbb{R}$ such that $F_n(x) = p$,
- (iii) $F_n^{-1}(1 - p) = -F_n^{-1}(p)$ for all $p \in]0, 1[$.

The PDF of $t(n)$ is illustrated in Figure 2.3. Quantiles $F_n^{-1}(p)$ can be computed numerically and are listed in Table B.2 for typical values of p .

Figure 2.3: PDF of $t(n)$ for $n \in \{1, 2, 5, 30\}$.

Proposition 6 (CI for μ , σ unknown). Let X_1, \dots, X_n be i.i.d. with $X_1 \sim N(\mu, \sigma^2)$.

(i) We have

$$\frac{\sqrt{n}}{S_n} \cdot (\bar{X}_n - \mu) \sim t(n-1).$$

In particular, the distribution $t(n-1)$ does not depend on the unknown parameters μ and σ .

(ii) A confidence interval for μ with confidence level $1 - \alpha$ is given by

$$a_n(\mathbf{x}) = \bar{x}_n - t_{n-1, 1-\alpha/2} \cdot \frac{s_n}{\sqrt{n}}, \quad b_n(\mathbf{x}) = \bar{x}_n + t_{n-1, 1-\alpha/2} \cdot \frac{s_n}{\sqrt{n}},$$

where $t_{n-1, 1-\alpha/2}$ denotes the $(1 - \alpha/2)$ -quantile of $t(n-1)$.

Example 7. Figure 2.4 shows 20 confidence intervals according to Proposition 6. The underlying data is identical to Example 3. The relative share of confidence intervals not containing the true value $\mu = 2$ is given by $1/20$ for both values of n .

Definition 8. Let $n \in \mathbb{N}$. A random variable X has a χ^2 -distribution with n degrees of freedom if it is absolutely continuous with density $f_n: \mathbb{R} \rightarrow \mathbb{R}$ given by

$$f_n(x) = \begin{cases} (2^{n/2} \cdot \Gamma(n/2))^{-1} \cdot x^{n/2-1} \cdot \exp(-x/2), & \text{if } x > 0, \\ 0, & \text{if } x \leq 0, \end{cases}$$

where Γ denotes the gamma function.

Notation: $X \sim \chi^2(n)$.

Remark 9. The PDF of $\chi^2(n)$ is illustrated in Figure 2.5. Quantiles of $\chi^2(n)$ can be computed numerically and are listed in Table B.3 for typical values of p .

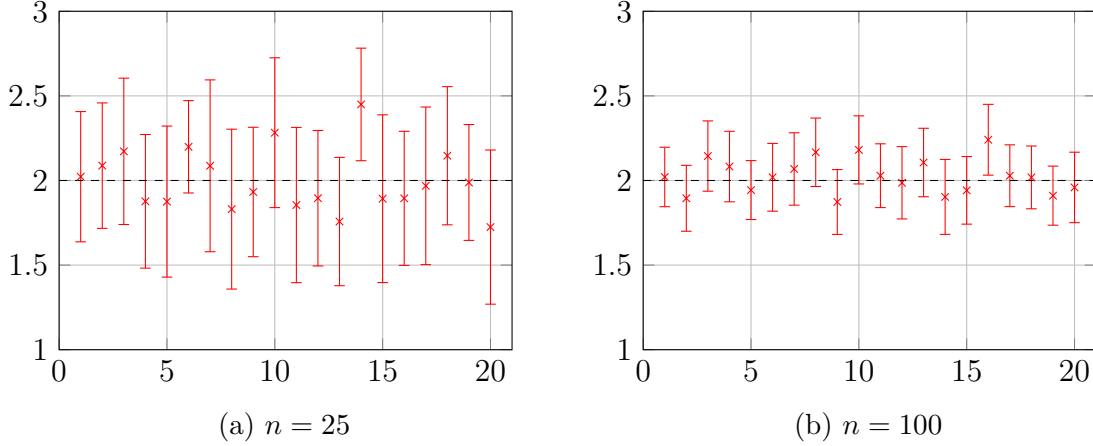


Figure 2.4: Computer simulation of 20 independently generated confidence intervals for $X \sim N(\mu, 1)$ with $\alpha = 0.05$, $\mu = 2$, and $n \in \{25, 100\}$, see Example 7.

Proposition 10 (CI for σ^2 , μ unknown). Let X_1, \dots, X_n be i.i.d. with $X_1 \sim N(\mu, \sigma^2)$.

(i) We have

$$\frac{n-1}{\sigma^2} \cdot S_n^2 \sim \chi^2(n-1).$$

In particular, the distribution $\chi^2(n-1)$ does not depend on the unknown parameters μ and σ .

(ii) A confidence interval for σ^2 with confidence level $1 - \alpha$ is given by

$$a_n(\mathbf{x}) = \frac{n-1}{\chi_{n-1, 1-\alpha/2}^2} \cdot s_n^2, \quad b_n(\mathbf{x}) = \frac{n-1}{\chi_{n-1, \alpha/2}^2} \cdot s_n^2,$$

where $\chi_{n-1, p}^2$ denotes the p -quantile of $\chi^2(n-1)$.

Asymptotic Confidence Intervals

In the sequel let $P_X = P_X^\vartheta$ with $\vartheta \in \Theta$ and let

$$\gamma(\vartheta) = E^\vartheta(X).$$

Moreover, let $\text{Var}^\vartheta(X) > 0$ for all $\vartheta \in \Theta$. Let $\alpha \in]0, 1[$.

Proposition 11 (Asymptotic confidence intervals). Put

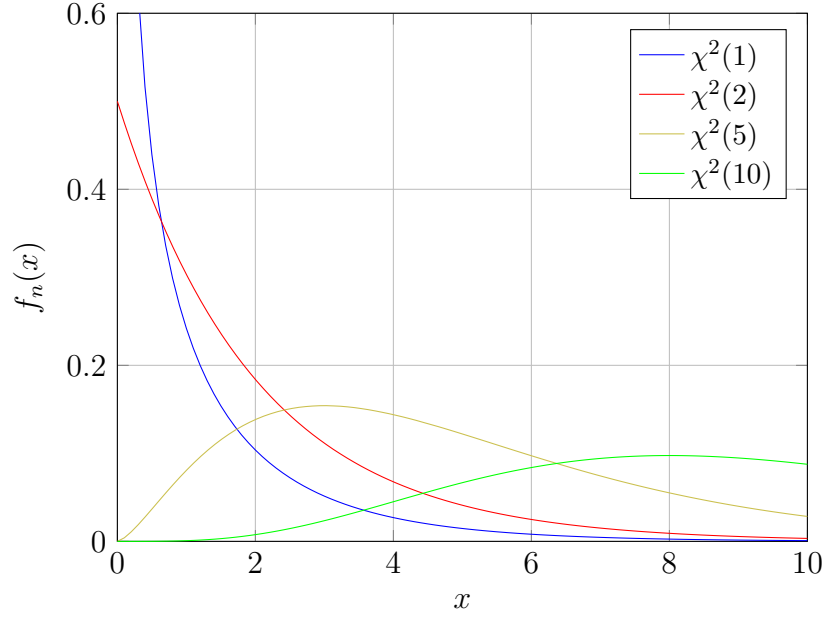
$$q_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$$

and

$$a_n(\mathbf{x}) = \bar{x}_n - q_{1-\alpha/2} \cdot \frac{s_n}{\sqrt{n}}, \quad b_n(\mathbf{x}) = \bar{x}_n + q_{1-\alpha/2} \cdot \frac{s_n}{\sqrt{n}}.$$

Then we have

$$\forall \vartheta \in \Theta: \lim_{n \rightarrow \infty} P^\vartheta(\{E^\vartheta(X) \in [a_n(\mathbf{X}), b_n(\mathbf{X})]\}) = 1 - \alpha.$$

Figure 2.5: PDF of $\chi^2(n)$ for $n \in \{1, 2, 5, 10\}$.

Example 12 (Continuation of Chapter I). Let $X \sim B(1, p)$ with $p = \vartheta \in \Theta = [0, 1]$ to model the gender of a newborn where p is the probability of having a girl. Empirical data from a sample of size

$$n = 25\,171\,123$$

yields

$$\bar{x}_n = \frac{12\,241\,392}{25\,171\,123} = 0.486326 \dots$$

and

$$s_n^2 = 0.2498 \dots$$

An asymptotic confidence interval for $\alpha = 0.001$ is given by

$$\begin{aligned} [a_n(\mathbf{x}), b_n(\mathbf{x})] &\approx [\bar{x}_n - 3.27 \cdot 10^{-4}, \bar{x}_n + 3.27 \cdot 10^{-4}] \\ &= [0.4859 \dots, 0.4866 \dots]. \end{aligned}$$

Chapter V

Statistics – Hypothesis Tests

We consider a random experiment modelled by a random variable $X: \Omega \rightarrow \mathbb{R}$ on a probability space (Ω, \mathcal{A}, P) where the distribution P_X is unknown. We aim at deciding whether

- a certain conjecture H_0 about P_X is “true” (H_0 is called *null hypothesis*) or
- an *alternative hypothesis* H_1 is “true”.

For this, a sample

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

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

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

Moreover, as in Chapter IV we consider the case where $P_X = P_X^\vartheta$ is known up to some parameter $\vartheta \in \Theta \subseteq \mathbb{R}^d$, and we consider

- a null hypothesis $H_0: \vartheta \in \Theta_0$ with $\Theta_0 \subseteq \Theta$,
- an alternative hypothesis $H_1: \vartheta \in \Theta_1$ with $\Theta_1 = \Theta \setminus \Theta_0$.

In the context of hypothesis tests there are two types of error, cf. Table V.1:

- (i) *Type I error*: H_0 is true and H_0 is rejected.
- (ii) *Type II error*: H_1 is true (i.e., H_0 is false) and H_0 is not rejected.

Example 1 (Continuation of Example IV.2.12). For testing the gender of a newborn (girl (1), boy (0)) we put

- $H_0: P_X^\vartheta \sim \text{B}(1, p)$ with $p = \vartheta \in \Theta_0 = [0, 1/2[$, i.e., it is more likely to have a boy than a girl.
- $H_1: P_X^\vartheta \sim \text{B}(1, p)$ with $p = \vartheta \in \Theta_1 = [1/2, 1]$.

	H_0 is true	H_0 is false
H_0 is rejected	type I error	right decision
H_0 is not rejected	right decision	type II error

Table V.1: Type I/II error for hypothesis tests.

A reasonable decision rule whether to reject the null hypothesis H_0 is given by:

- (i) Determine a critical value $c_{\text{crit}} \in [0, n[$.
- (ii) Reject H_0 if and only if

$$\sum_{i=1}^n x_i > c_{\text{crit}}.$$

Recall that for $X \sim B(1, p)$ we have

$$Y_n = \sum_{i=1}^n X_i \sim B(n, p).$$

Aim: For a given $\varepsilon > 0$ we try to choose c_{crit} such that the probabilities of both types of error are bounded by ε , i.e.,

- (i) $P(\{Y_n > c_{\text{crit}}\}) \leq \varepsilon$ for every $p < 1/2$ (type I error),
- (ii) $P(\{Y_n \leq c_{\text{crit}}\}) \leq \varepsilon$ for every $p \geq 1/2$ (type II error).

The probability $P(\{Y_n > c_{\text{crit}}\})$ is illustrated in Figure 0.1 for $n = 11$ and $c_{\text{crit}} = 6$ as a function of p , i.e.,

$$P(\{Y_n > c_{\text{crit}}\}) = \sum_{k=\lfloor c_{\text{crit}}+1 \rfloor}^n \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \sum_{k=7}^{11} \binom{11}{k} \cdot p^k \cdot (1-p)^{11-k}.$$

Let $Z \sim B(n, 1/2)$. Note that

- (i) under H_0 the probability $P(\{Y_n > c_{\text{crit}}\})$ can be arbitrarily close to $P(\{Z > c_{\text{crit}}\})$,
- (ii) under H_1 the probability $P(\{Y_n \leq c_{\text{crit}}\})$ can be equal to $P(\{Z \leq c_{\text{crit}}\})$,
- (iii) $P(\{Z > c_{\text{crit}}\}) + P(\{Z \leq c_{\text{crit}}\}) = 1$.

Hence we cannot choose c_{crit} such that the probabilities of both types of error are small. Therefore, we choose the critical value c_{crit} such that

- (i) $P(\{Z > c_{\text{crit}}\}) \leq \alpha$ for a specified threshold $\alpha \in]0, 1[$,
- (ii) $P(\{Z \leq c_{\text{crit}}\})$ as small as possible.

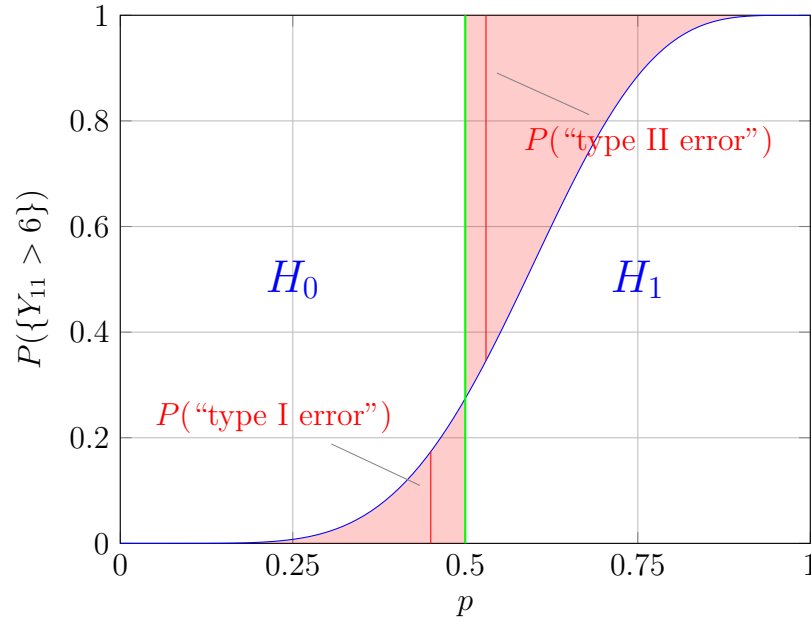


Figure 0.1: Probability $P(\{Y_n > c_{\text{crit}}\})$ of rejecting H_0 with $n = 11$ and $c_{\text{crit}} = 6$ as a function of p , see Example 1.

Definition 2.

- (i) A *statistical test* (or *hypothesis test*) consists of a *test statistic*

$$T_n = g_n(X_1, \dots, X_n)$$

with $g_n: \mathbb{R}^n \rightarrow \mathbb{R}$ and a *rejection region* (or *critical region*) $R_n \subseteq \mathbb{R}$. The null hypothesis is rejected if and only if $g_n(x_1, \dots, x_n) \in R_n$.

- (ii) A statistical test is of (*significance*) *level* α with $\alpha \in]0, 1[$ if

$$\forall \vartheta \in \Theta_0: P^\vartheta(\{T_n \in R_n\}) \leq \alpha.$$

Example 3 (Continuation of Example 1). We have

$$T_n = g_n(X_1, \dots, X_n) = \sum_{i=1}^n X_i$$

and $R_n =]c_{\text{crit}}, \infty[$. Moreover, we have

$$\sup_{\vartheta \in \Theta_0} P^\vartheta(\{T_n \in R_n\}) = \sup_{\vartheta \in [0, 1/2[} P^\vartheta(\{Y_n > c_{\text{crit}}\}) = P(\{Z > c_{\text{crit}}\}),$$

see Figure 0.2, and

$$\begin{aligned} P(\{Z > c_{\text{crit}}\}) \leq \alpha &\Leftrightarrow 1 - \alpha \leq P(\{Z \leq c_{\text{crit}}\}) \\ &\Leftrightarrow c_{\text{crit}} \geq b_{n, 1/2, 1-\alpha}, \end{aligned}$$

where $b_{n, 1/2, 1-\alpha}$ denotes the $(1 - \alpha)$ -quantile of $B(n, 1/2)$. Hence we get a level α test if and only if $c_{\text{crit}} \geq b_{n, 1/2, 1-\alpha}$. Since the probability of a type II error is monotonically increasing in c_{crit} , we choose $c_{\text{crit}} = b_{n, 1/2, 1-\alpha}$.

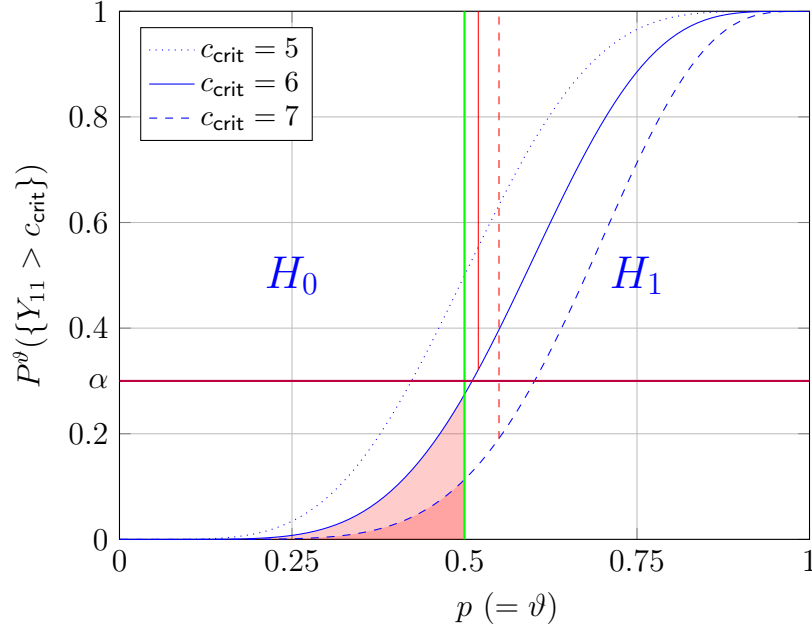


Figure 0.2: Probability $P^\vartheta(\{Y_n > c_{\text{crit}}\})$ of rejecting H_0 with $n = 11$ and $c_{\text{crit}} \in \{5, 6, 7\}$ as a function of $p (= \vartheta)$, see Example 3. The probability of a type I error is bounded by $\alpha = 0.3$ for $c_{\text{crit}} \in \{6, 7\}$. The probability of a type II error for $c_{\text{crit}} = 6$ is less than the corresponding probability for $c_{\text{crit}} = 7$.

Remark 4. (i) Notice the asymmetry regarding the choice of H_0 and H_1 . The probability of a type I error is bounded by α whereas the probability of a type II error does not satisfy a specified error bound. Consequently, the null hypothesis is chosen such that a type I error is more embarrassing (or more serious) than a type II error.

(ii) For a given significance level one is interested in constructing statistical tests with a “large” rejection region to obtain “small” probabilities for type II errors.

1 Tests for Normal Distribution

In the sequel let $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma \in]0, \infty[$, and let $\alpha \in]0, 1[$.

Proposition 1 (*Z-test (two-tailed)*). Let $\sigma > 0$ be known and let $\mu = \vartheta \in \Theta = \mathbb{R}$ and $\Theta_0 = \{\mu_0\}$. Put

- $H_0: X \sim N(\mu_0, \sigma^2)$,
- $H_1: X \sim N(\mu, \sigma^2)$ with $\mu \neq \mu_0$.

Let $q_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ and let

$$T_n = g_n(X_1, \dots, X_n) = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

The decision rule

$$H_0 \text{ is rejected} \Leftrightarrow |g_n(x_1, \dots, x_n)| > q_{1-\alpha/2}$$

defines a statistical test of significance level α .

Proposition 2 (Z-test (one-tailed)). Let $\sigma > 0$ be known and let $\mu = \vartheta \in \Theta = \mathbb{R}$ and $\Theta_0 =]-\infty, \mu_0]$. Put

- $H_0: X \sim N(\mu, \sigma^2)$ with $\mu \leq \mu_0$ (or $\mu < \mu_0$),
- $H_1: X \sim N(\mu, \sigma^2)$ with $\mu > \mu_0$ (or $\mu \geq \mu_0$).

Let $q_{1-\alpha} = \Phi^{-1}(1 - \alpha)$ and let

$$T_n = g_n(X_1, \dots, X_n) = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

The decision rule

$$H_0 \text{ is rejected} \Leftrightarrow g_n(x_1, \dots, x_n) > q_{1-\alpha}$$

defines a statistical test of significance level α .

Remark 3. Proposition 2 also holds for

- $H_0: X \sim N(\mu, \sigma^2)$ with $\mu \geq \mu_0$ (or $\mu > \mu_0$)

together with the decision rule

$$H_0 \text{ is rejected} \Leftrightarrow g_n(x_1, \dots, x_n) < -q_{1-\alpha}.$$

Example 4. Consider a fill level control with a minimum fill level of $\mu_0 = 5$ and known standard deviation $\sigma = 0.1$. The producer wants to verify the minimum fill level based on a statistical test of significance level $\alpha = 0.01$.

(i) Model: $X \sim N(\mu, \sigma^2)$ with $\mu = \vartheta \in \Theta = \mathbb{R}$; X_1, \dots, X_n i.i.d. with $P_{X_1} = P_X$.

(ii) Hypotheses:

- $H_0: \mu < \mu_0$,
- $H_1: \mu \geq \mu_0$.

(iii) Test statistic:

$$g_n(X_1, \dots, X_n) = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

Empirical data from a sample of size $n = 20$ yields $\bar{x}_{20} = 5.07$. Hence we get (one-tailed Z-test)

$$g_{20}(x_1, \dots, x_{20}) = \sqrt{20} \cdot \frac{5.07 - 5}{0.1} \approx 3.130 > 2.326 \approx q_{0.99}$$

such that H_0 is rejected. Note that H_0 is rejected if and only if

$$\bar{x}_n > 5 + \frac{0.01}{\sqrt{20}} \cdot q_{0.99} \approx 5.005.$$

Example 5 (Continuation of Example 4). Consider the hypothesis

- $H_0: \mu \geq \mu_0$.

Then H_0 is rejected if and only if

$$\bar{x}_n < 5 - \frac{0.01}{\sqrt{20}} \cdot q_{0.99} \approx 4.994,$$

see Remark 3.

In the sequel let $n \geq 2$ and recall that

$$S_n^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Proposition 6 (t-test (two-tailed)). Let

$$(\mu, \sigma) = \vartheta \in \Theta = \mathbb{R} \times]0, \infty[\quad \text{and} \quad \Theta_0 = \{\mu_0\} \times]0, \infty[.$$

Put

- $H_0: X \sim N(\mu_0, \sigma^2)$ with $\sigma > 0$,
- $H_1: X \sim N(\mu, \sigma^2)$ with $\mu \neq \mu_0$ and $\sigma > 0$.

Let $t_{n-1, 1-\alpha/2}$ be the $(1 - \alpha/2)$ -quantile of $t(n-1)$ and let

$$T_n = g_n(X_1, \dots, X_n) = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}.$$

The decision rule

$$H_0 \text{ is rejected} \Leftrightarrow |g_n(x_1, \dots, x_n)| > t_{n-1, 1-\alpha/2}$$

defines a statistical test of significance level α .

Proposition 7 (t-test (one-tailed)). Let

$$(\mu, \sigma) = \vartheta \in \Theta = \mathbb{R} \times]0, \infty[\quad \text{and} \quad \Theta_0 =]-\infty, \mu_0] \times]0, \infty[.$$

Put

- $H_0: X \sim N(\mu, \sigma^2)$ with $\mu \leq \mu_0$ (or $\mu < \mu_0$) and $\sigma > 0$,
- $H_1: X \sim N(\mu, \sigma^2)$ with $\mu > \mu_0$ (or $\mu \geq \mu_0$) and $\sigma > 0$.

Let $t_{n-1, 1-\alpha}$ be the $(1 - \alpha)$ -quantile of $t(n-1)$ and let

$$T_n = g_n(X_1, \dots, X_n) = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}.$$

The decision rule

$$H_0 \text{ is rejected} \Leftrightarrow g_n(x_1, \dots, x_n) > t_{n-1, 1-\alpha}$$

defines a statistical test of significance level α .

Remark 8. Proposition 7 also holds for

- $H_0: X \sim N(\mu, \sigma^2)$ with $\mu \geq \mu_0$ (or $\mu > \mu_0$) and $\sigma > 0$

together with the decision rule

$$H_0 \text{ is rejected} \Leftrightarrow g_n(x_1, \dots, x_n) < -t_{n-1, 1-\alpha}.$$

Example 9 (Continuation of Example 4). Assume that $\sigma > 0$ is not known anymore and consider the test statistic

$$g_n(X_1, \dots, X_n) = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}.$$

For $\bar{s}_{20} = 0.124$ we obtain (one-tailed t-test)

$$g_{20}(x_1, \dots, x_{20}) = \sqrt{20} \cdot \frac{5.07 - 5}{0.124} \approx 2.524 \not\approx 2.539 \approx t_{19, 0.99}$$

such that H_0 is not rejected.

2 Goodness-of-fit Tests

...

Appendix A

Combinatorics

In this chapter we consider four different urn models where k elements are drawn from the set $\Omega = \{1, \dots, n\}$ with $n \in \mathbb{N}$:

- (i) Model I: order of selection matters, with replacement.
- (ii) Model II: order of selection matters, without replacement.
- (iii) Model III: order of selection does not matter, without replacement.
- (iv) Model IV: order of selection does not matter, with replacement.

Note that the outcomes within Model I and Model II are k -tuples. The outcomes within Model III and Model IV are sets of cardinality k .

Proposition 1 (Model I, Cartesian products). For finite sets $\Omega_1, \dots, \Omega_k$ we have

$$|\Omega_1 \times \dots \times \Omega_k| = \prod_{i=1}^k |\Omega_i|.$$

In particular, we have

$$|\Omega^k| = n^k.$$

Example 2. Number of equipment variants of an e-bike:

$$\begin{array}{ll} \Omega_1 = \{51, 54, 56\} & \text{frame size} \\ \Omega_2 = \{\mathbf{f}, \mathbf{m}\} & \text{frame type} \\ \Omega_3 = \{\mathbf{d}, \mathbf{h}\} & \text{shifting system} \\ \Omega_4 = \{50, 80\} & \text{torque (motor)} \\ \vdots & \vdots \end{array}$$

Proposition 3 (Model II, k -permutations). For $1 \leq k \leq n$ we have

$$\left| \{(\omega_1, \dots, \omega_k)^\top \in \Omega^k : \omega_1, \dots, \omega_k \text{ pairwise distinct}\} \right| = n \cdot (n-1) \cdots (n-(k-1)).$$

In particular, we have

$$\left| \{(\omega_1, \dots, \omega_n)^\top \in \Omega^n : \omega_1, \dots, \omega_n \text{ pairwise distinct}\} \right| = n!.$$

Example 4. The number of possible final tables of the Bundesliga is given by

$$18! = 6\,402\,373\,705\,728\,000 \approx 6.4 \cdot 10^{15}.$$

Proposition 5 (Model III, k -combinations). For $0 \leq k \leq n$ we have

$$|\{A \subseteq \Omega: |A| = k\}| = \binom{n}{k}.$$

Example 6. The number of possible outcomes in a 6/49 game (lottery) is given by

$$\binom{49}{6} = 13\,983\,816 \approx 1.4 \cdot 10^7.$$

Remark 7 (Power set and binomial coefficients). Let $f: \mathcal{P}(\{1, \dots, n\}) \rightarrow \{0, 1\}^n$ be given by

$$f(A) = (x_1, \dots, x_n)$$

for $A \subseteq \{1, \dots, n\}$ with

$$x_i = \begin{cases} 1, & \text{if } i \in A, \\ 0, & \text{if } i \notin A. \end{cases}$$

Then f is bijective and hence

$$|\mathcal{P}(\Omega)| = |\mathcal{P}(\{1, \dots, n\})| = |\{0, 1\}^n| = 2^n = 2^{|\Omega|}$$

due to Proposition 1. Moreover, we obtain

$$|\{A \subseteq \Omega: |A| = k\}| = \left| \left\{ x \in \{0, 1\}^n: \sum_{i=1}^n x_i = k \right\} \right|,$$

and, by Proposition 5,

$$\sum_{k=0}^n \binom{n}{k} = \sum_{k=0}^n \left| \left\{ x \in \{0, 1\}^n: \sum_{i=1}^n x_i = k \right\} \right| = |\{0, 1\}^n| = 2^n.$$

Let $g: \mathcal{P}(\{1, \dots, n\}) \rightarrow \mathcal{P}(\{1, \dots, n\})$ be given by $g(A) = \{1, \dots, n\} \setminus A$. Then g is bijective and $|g(A)| = n - |A|$. Hence Proposition 5 shows for $0 \leq k \leq n$

$$\binom{n}{k} = |\{A \subseteq \Omega: |A| = k\}| = |\{A \subseteq \Omega: |A| = n - k\}| = \binom{n}{n-k}.$$

Proposition 8 (Model IV, k -combinations with repetition). We have

$$\left| \left\{ (h_1, \dots, h_n) \in \mathbb{N}_0^n: \sum_{i=1}^n h_i = k \right\} \right| = \binom{k+n-1}{k}.$$

Example 9. The number of allocations of $k = 2500$ students to $n = 4$ different meals in the university canteen is given by

$$\binom{2503}{2500} = \binom{2503}{3} = 2\,610\,421\,251 \approx 2.6 \cdot 10^9.$$

Appendix B

Tables

In this chapter the following tables are provided:

- (i) Approximate values of the cumulative distribution function Φ of the standard normal distribution, i.e.,

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

for $x \in \{0, 0.01, 0.02, \dots, 3.49\}$, are shown in Table B.1.

- (ii) The p -quantiles of $t(n)$ with $n \in \{1, 2, 3, \dots, 29\}$ and of $N(0, 1)$ are shown in Table B.2 for $p \in \{0.9, 0.95, 0.975, 0.99, 0.995\}$.
- (iii) The p -quantiles of $\chi^2(n)$ with $n \in \{1, 2, 3, \dots, 29, 30, 40, 50, \dots, 100\}$ are shown in Table B.3 for $p \in \{0.005, 0.01, 0.025, 0.05, 0.9, 0.95, 0.975, 0.99, 0.995\}$.

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 the standard normal distribution, i.e., $\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^x \exp(-y^2/2) dy$.

n	p				
	0.90	0.95	0.975	0.99	0.995
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576

Table B.2: p -quantiles of the t-distribution with n degrees of freedom. The row with $n = \infty$ shows the p -quantiles of $N(0, 1)$.

n	p								
	0.005	0.01	0.025	0.05	0.90	0.95	0.975	0.99	0.995
1	$4 \cdot 10^{-5}$	$1.6 \cdot 10^{-4}$	0.001	0.004	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	118.498	124.342	129.561	135.807	140.169

Table B.3: p -quantiles of the χ^2 -distribution with n degrees of freedom.