Utrecht University

ASYMPTOTIC STATISTICS: MOMENT ESTIMATORS AND M-ESTIMATORS IN PARAMETRIC MODELS

Bachelor Thesis
January 17, 2020

Author: Wessel Kroon
Studentnumber: 5989353

Supervisor: Dr. Karma Dajani

Contents

Chapter 1

Introduction

1.1 Asymptotic statistics

Asymptotic statistics explores the properties of statistical procedures in the case where the sample size n tends to infinity. So throughout this thesis we implicitly assume that n tends to infinity. There are two main reasons why the study of asymptotic statistics is worthwhile. Firstly, statistics becomes much simpler as the sample size n goes to infinity. In the limit we can obtain results that are very hard, or even impossible, to obtain for a finite sample size n. If we take a large but finite sample size, our statistical procedures approximate the results we obtained for n tending to infinity. So the study of asymptotic statistics allows us to draw conclusions about large samples which could otherwise not be drawn. This brings us to the second reason. Modern techniques allow us to actually gather large samples, which makes the developments in asymptotic statistics useful for actual statistical research. A term that has become increasingly popular for statistical research which utilizes large sample sizes is 'big data', the results of asymptotic statistics play a key role in this kind of research.

The results of asymptotic statistis can be used among other things to approximate statistical procedures or to provide asymptotic optimal solutions for statistical problems. In this thesis our focus will mainly lie on parameter estimation in asymptotic statistics. Specifically, we will focus on moment estimators and M-estimators in parametric models. Our goal is to develop the underlying theory of why such estimators are consistent and asymptotically normal. Furthermore, we want to demonstrate that the theoretical results we obtain when the sample size n tends to infinity are also applicable when having a large finite sample size.

In chapter 2 we will demarcate what knowledge we assume the reader to

possess. Additionally, we will expand probability theory to include random vectors and brush up on some well known results from statistics.

Chapter 3 concerns itself with the three modes of convergence required in order to do asymptotic statistics and how these modes relate to each other.

The next chapter concerns itself with the Delta Method and some of its applications. The Delta Method can be used to establish convergence of a sequence as a consequence of the convergence of another sequence. We introduce the concept of a parametric model and briefly discuss some of the limitations of asymptotic theory. Lastly, we will examine moment estimators and prove that these are asymptotically normal using the Delta Method.

Chapter 5 deals with the most important class of estimators: M-estimators. These estimators use a maximization procedure in order to find an estimate for a parameter θ. We study the consistency and asymptotic normality of this class of estimators. Particularly, we will look at the maximum likelihood estimator which is the most important kind of M-estimator. This chapter concludes the purely theoretical part of this thesis.

In Chapter 6 we will investigate to what extent the abstract theory is applicable in actual statistical research. We conduct several simulations in order to show how the theory we developed is related to statistical procedures with a finite sample size.

Subsequently, we give a conclusion in which we summarize our findings.

1.2 Consulted Literature

The first paragraph of the previous section and chapters 3-5 are based on Asymptotic Statistics by van der Vaart [6]. Most examples, theorems and proofs that we present are also given in this book. Van der Vaart's book is the most important source used in this thesis and therefore deserves explicit acknowledgement in this introduction. However, it is not the only source. Appendix A provides a precise overview describing where each definition, lemma, theorem and proof has been taken from or where our idea originated. Literal adoptions from other literature do not occur in this thesis, with the exception of Lemma 2.9. Also, if we have chosen to omit the proof of a theorem, then a reference is given in the text. Lastly, not every part of this thesis is based on existent literature, Section 3.2 and Chapter 6 provide examples of this among other things.

The reader is referred to appendix A if interested in the source of a specific part of this thesis.

1.3 Notation

$\mathbf{1}$	all-ones vector
$\xrightarrow{as}$	convergence almost surely
$\xrightarrow{d}$	convergence in distribution
$\xrightarrow{P}$	convergence in probability
$\mathrm{Cov}(X, Y)$	covariance of X and Y
$\mathrm{E}(X)$	expectation of X
$\mathbb{1}_A$	indicator function of the set A
$N_k(\mu, \Sigma)$	multivariate normal distribution with mean μ and covariance matrix Σ
$\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$	number fields and sets
$\mathbb{P}$	probability measure
$o_p(1)$, $O_p(1)$	stochastic o and O symbols
$\mathrm{Var}(X)$	variance of X
$\subseteq$	(not necessarily strict) subset
$\overline{\mathbb{R}}$	$\mathbb{R} \cup \{\infty, -\infty\}$

Chapter 2

Preliminaries

In order to study asymptotic statistics we need some mathematical knowledge. We assume the reader has had some introduction to probability theory and statistics. We also assume that the reader has some basic calculus skills and knowledge about analysis. This chapter provides the additional necessary knowledge. The chapter is divided into three sections: analysis, probability theory and statistics. The most important section is the second section on probability theory. Readers familiar with mathematical statistics may skip the last section, we merely included this section because the statistical results presented in it are of fundamental importance in (asymptotic) statistics. Lastly, this chapter does not focus on proving the theorems presented, but any reader that is interested in proofs will be referred to the appropriate literature.

2.1 Analysis

We assume that the reader is familiar with (uniformly) bounded functions and Lipschitz functions. Moreover, we expect the reader to be familiar with the Heine-Cantor theorem, the Heine-Borel theorem and the inverse function theorem. All of these things are treated in chapters 3, 4 and 9 of Mathematical Analysis by Apostol [1].

Some experience with topology is also required. Chapter 2 of Munkres [4] should suffice in order to understand everything related to topology in this thesis.

Furthermore, we expect the reader to have knowledge about partial derivatives and total differentiability. Chapter 12 of Mathematical Analysis [1] contains all the required information.

2.2 Probability Theory

This section is based on the lecture notes of the course "Stochastic Processes" given by Dirksen [3].

We use probability spaces to mathematically model real world processes or experiments.

2.1 Definition (Probability space)**.** A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where
(i) Ω is a non-empty set, called the sample space.
(ii) $\mathcal{F}$ is a σ-algebra of subsets of Ω, called the event space.
(iii) $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$.

The set Ω contains all possible outcomes of our process or experiment. However, we might not be interested in individual outcomes, but in groups containing these outcomes. For example, if we have a simple dice we have $\Omega = \{1, 2, \ldots, 6\}$. When we are interested whether we throw an even number of eyes or an odd number of eyes we want to know if the outcome of our throw is an element of $\{1, 3, 5\}$ or of $\{2, 4, 6\}$. We use the σ-algebra $\mathcal{F}$ to characterize such sets of outcomes, we can consider such sets as events or as representations of an event. So the subset $\{1, 3, 5\}$ represents the event that we throw an odd number of eyes. Generally speaking, when $\omega \in \Omega$ occurs we say that all events $A \in \mathcal{F}$ for which $\omega \in A$ have occurred. The definition underneath gives the formal constraints we must place on $\mathcal{F}$.

2.2 Definition (Sigma-algebra)**.** A collection $\mathcal{F}$ of subsets of a non-empty set Ω is called a σ-algebra on Ω if
(i) $\mathcal{F} \neq \emptyset$.
(ii) If $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$.
(iii)If $A_i \in \mathcal{F}$ for $i \geq 1$, then $\cup_{i \geq 1} A_i \in \mathcal{F}$.

After we have obtained $(\Omega, \mathcal{F})$ we only need a probability measure working on $\mathcal{F}$: a map that gives the probability of each event in $\mathcal{F}$.

2.3 Definition (Probability measure)**.** A map $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called a probability measure on $(\Omega, \mathcal{F})$ if
(i) $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$.
(ii) $\mathbb{P}(\Omega) = 1$.
(iii) If for all $i \geq 1$ the events A_i are disjoint then $\mathbb{P}(\cup_{i \geq 1} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Now that we have briefly treated each component of a probability space, we can start looking at random variables. Suppose that we have some probability space, then we can rigorously define random variables.

2.4 Definition (Random variable)**.** A map $X : \Omega \to \mathbb{R}$ is called a random variable on $(\Omega, \mathcal{F})$ if for all $x \in \mathbb{R}$ we have

$$\{X \leq x\} = X^{-1}(-\infty, x] = \{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}.$$

All random variables correspond to a distribution function. Moreover, if we have some distribution function F, then there exists a probability space and a random variable X defined on this space such that F is the distribution function of X.

2.5 Definition (Distribution function)**.** The distribution function of a random variable X is defined by $F_X : \mathbb{R} \to [0, 1]$, with $F_X(x) = \mathbb{P}(X \leq x)$ and having the following properties:
(i) F_X is non-decreasing.
(ii) $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
(iii) F_X is right-continuous.

Quite often we will not be considering one individual random variable, but a group of random variables. In such a case it is practical to collect these random variables in a vector.

2.6 Definition (Random vector)**.** A tuple $(X_1, \ldots, X_k)$ is called a random vector if X_i is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ for $i = 1, \ldots, k$.

The distribution function of a random vector X thus becomes a function $F_X : \mathbb{R}^k \to [0, 1]$. When we consider a random vector X and write $\mathbb{P}(X \leq x)$, we implicitly assume x to be a vector as well such that $\mathbb{P}(X \leq x)$ can be interpreted as $\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_k)$. Random vectors have an expectation similar to random variables.

2.7 Definition (Expectation of random vector)**.** The expectation of a random vector X is given by

$$\mathrm{E}(X) = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_n) \end{pmatrix}.$$

However, the concept of variance cannot be expanded to include random vectors as straightforward as the concept of expectation. This is due to possible dependency between the random variables in the vector. Therefore we need to have an understanding of the covariance between all the random variables in the vector. The following matrix gives a plain overview of the covariances between all the different random variables and can be considered as the "variance" of random vectors.

2.8 Definition (Covariance matrix)**.** The covariance matrix of a k-dimensional random vector X is given by

$$\mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_k) \\ \mathrm{Cov}(X_2, X_1) & \cdots & \mathrm{Cov}(X_2, X_k) \\ \vdots & & \vdots \\ \mathrm{Cov}(X_k, X_1) & \cdots & \mathrm{Cov}(X_k, X_k) \end{pmatrix}.$$

Note that the diagonal of the covariance matrix above contains the variances of all the random variables X_i. Furthermore, if all random variables in a random vector are independent of each other, then the covariance matrix is a diagonal matrix. The following lemma states some basic properties of the expectation and covariance matrix of a random vector.

2.9 Lemma. *For every matrix A, vector b and random vector X the following statements are true:*
(i) $\mathrm{E}(AX + b) = A\mathrm{E}(X) + b$
(ii) $\mathrm{Cov}(AX) = A\mathrm{Cov}(X)A^T$.
(iii) $\mathrm{Cov}(X)$ *is symmetric and non-negative definite.*

Proof. For a proof, see [7], Lemma 2.1, p.13-14. □

We conclude this section with the definition of a very important distribution: the multivariate normal distribution.

2.10 Definition. A random vector X is said to be multivariate-normally distributed with parameters μ and Σ, if it has the same distribution as the vector $\mu + LZ$, for a matrix L with $\Sigma = LL^T$ and $Z = (Z_1, \ldots, Z_k)^T$ a vector whose coordinates are independent $N(0, 1)$-distributed random variables. We will denote this distribution as $N_k(\mu, \Sigma)$.

If the context makes it sufficiently clear that we are dealing with a multivariate normal distribution then the subscript 'k' will often be omitted.

2.3 Statistics

We briefly look at some important theorems from statistics. All the results we look at concern limit theorems. We start off with the law of large numbers.

2.11 Theorem (Weak law of large numbers)**.** *Let $X_1, \ldots, X_n, \ldots$ be a sequence of i.i.d. random variables with $\mathrm{E}(|X_1|) < \infty$, $\mathrm{E}(X_1) = \mu$ and Var$(X_1) = \sigma^2$. For the sample average $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ we find that for any $\epsilon > 0$,*

$$\mathbb{P}(|\overline{X_n} - \mu| > \epsilon) \to 0$$

as $n \to \infty$.

Proof. For a proof, see [5], Theorem A, p.178. □

The law of large numbers thus states that the sample average $\overline{X_n}$ converges to μ in some way. As the name of the theorem suggests, there is also a strong version of this law. Under the same assumptions, it states that $\mathbb{P}(\lim_{n\to\infty} \overline{X_n} = \mu) = 1$. In the next chapter we will see that the weak law of large numbers corresponds to convergence in probability and the strong law to convergence almost surely.

2.12 Theorem (Central limit theorem)**.** *Let* $X_1, \ldots, X_n$ *be a sample of i.i.d. random variables with* $\mathrm{E}(|X_1|) < \infty$ *and* $Var(X_1) = \sigma^2 < \infty$. *Also, let* $\mu = \mathrm{E}(X_1) < \infty$ *and let* $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. *Then*

$$\sqrt{n}(\overline{X_n} - \mu)$$

tends to a normal distribution with mean zero and variance σ^2 *as* $n \to \infty$.

Proof. For a proof, see [5], theorem B, p.184.

So the law of large numbers tells us that the sample average converges to μ, while the central limit theorem tells us something about the distribution of the sample average around μ during this convergence. The following theorem generalizes the central limit theorem to random vectors.

2.13 Theorem. *Let* $||\cdot||$ *denote the Euclidean norm. Suppose* $X_1, X_2, \ldots$ *are i.i.d. random vectors in* $\mathbb{R}^k$ *with* $\mathrm{E}(||X_1||) < \infty$ *and finite covariance matrix* $\mathrm{Cov}(X) = \Sigma$. *Let* μ *be a* k*-dimensional vector such that* $\mu_i = \mathrm{E}(X_i)$ *for all* $i = 1, \ldots, k$. *Then the sequence*

$$\sqrt{n}(\overline{X_n} - \mu)$$

tends to $N_k(0, \Sigma)$ *as* $n \to \infty$.

Proof. For a proof, see [2], Theorem 29.5, p.409. □

Lastly, we take a look at the concept of Fisher information and the Cramér-Rao bound. The Fisher information $I(\theta)$ of a random variable is a measure of the amount of information that a random variable X possesses about a parameter θ that determines the distribution of X. We express the Fisher information in terms of the score.

2.14 Definition (Score)**.** Consider a random variable with a density function p_θ, i.e. the density of the random variable depends on the k-dimensional parameter θ. Then the score of this random variable is the k-dimensional vector of partial derivatives $\ell'_{\theta,i} = \frac{\partial}{\partial\theta_i}\log(p_\theta)$ for $i = 1,\dots,k$. Thus the score is the gradient of the log likelihood with respect to the k-dimensional parameter θ.

Now, we are able to define the Fisher information.

2.15 Definition. The Fisher information $I(\theta)$ is equal to the variance of the score.

In the case where θ is one-dimenional the variance of the score, and thus the Fisher information, is equal to

$$I(\theta) = \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log(p_\theta(x))\right)^2\right) - \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log(p_\theta(x))\right)^2$$

However, it turns out that the second term on the right is equal to zero under suitable regularity conditions, since

$$\begin{aligned}
\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log(p_\theta(x))\right) &= \int p_\theta(x)(\frac{\partial}{\partial\theta}\log(p_\theta(x)))dx \\
&= \int p_\theta(x)\frac{1}{p_\theta(x)}(\frac{\partial}{\partial\theta}p_\theta(x))dx \\
&= \frac{\partial}{\partial\theta}\int p_\theta(x)dx \\
&= \frac{\partial}{\partial\theta}1 \\
&= 0.
\end{aligned}$$

Notice that the required regularity conditions must enable us to interchange the integration and derivative symbols. Given that the first moment of the score is equal to zero, we find that

$$I(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log(p_\theta(x))\right)^2\right].$$

The treatment above is rather vague about what the Fisher information looks like in the case that θ is higher-dimensional. In the case that θ is k-dimensional the Fisher information $I(\theta)$ is a $k \times k$-matrix with

$$I(\theta)_{i,j} = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta_i}\log(p_\theta(x))\right)\left(\frac{\partial}{\partial\theta_j}\log(p_\theta(x))\right)\right],$$

or equivalently

$$I(\theta) = \mathrm{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log(p_\theta(x)) \right) \left(\frac{\partial}{\partial \theta} \log(p_\theta(x)) \right)^T \right].$$

We conclude the preliminaries with the statement of the Cramér-Rao inequality. This inequality provides a lower bound for the variance of unbiased estimators.

2.16 Theorem (Cramér-Rao inequality)**.** *Let $X_1, \ldots, X_n$ be an i.i.d. sample from a distribution with density function $p_\theta(x)$ to which suitable regularity conditions apply. If $\hat{\theta}_n$ is an unbiased estimate of θ, then*

$$Var(\hat{\theta}_n) \geq \frac{1}{nI(\theta)}.$$

Proof. For a proof, see [5], theorem A, p.300-301.

Thus when finding an unbiased estimator $\hat{\theta}_n$ having variance $\frac{1}{nI(\theta)}$, we know that this estimator is in some sense optimal. Towards the end of this thesis this inequality becomes interesting with regard to the maximum likelihood estimator.

Chapter 3

Stochastic Convergence

In this chapter we will introduce three modes of stochastic convergence: convergence in distribution, convergence in probability and convergence almost surely. The first two sections will be concerned with introducing these modes of convergence and studying some of their properties. In the last section we will study the relations between these three modes of convergence.

3.1 Convergence in distribution

We begin with a definition.

3.1 Definition (Convergence in distribution)**.** Let X be a random vector and let X_n be a sequence of random vectors. If

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$$

for all x at which the limit distribution function $x \mapsto \mathbb{P}(X \leq x)$ is continuous, then we say that X_n converges in distribution to X, denoted by $X_n \xrightarrow{d} X$.

Remember that if X_n and X are vectors, then x is a vector as well and $\mathbb{P}(X_n \leq x)$ and $\mathbb{P}(X \leq x)$ should be interpreted coordinate-wise. This means that $\mathbb{P}(X \leq x) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k)$ where $X = (X_1, \ldots, X_k)$ and $x = (x_1, \ldots, x_k)$. Throughout this thesis we will use $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ to denote $\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$. We will mostly use the name 'weak convergence' instead of 'convergence in distribution'. This alternative naming will turn out to be intuitively appealing since convergence in distribution will be the weakest form of convergence we will treat.

Another name that is sometimes used for this kind of convergence is 'convergence in law', since this kind of convergence is only dependent on

the induced laws of the vectors and not on the probability spaces on which they are defined. This means that the notion of weak convergence is still meaningful if X and X_n are defined on different probability spaces.

We could wonder why we only require $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ for x for which the limit distribution function is continuous. Since we are dealing with a distribution function, it is right-continuous and increasing. This in turn implies that there are countably many "jumps", which means that there are uncountably many $x \in \mathbb{R}^k$ where $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ has to hold. The exclusion of the jump points therefore does not seem unreasonable.

We look at a simple example of a sequence of random variables that converges weakly.

3.2 Example. Let X be an exponentially distributed random variable with parameter $\lambda > 0$, so that

$$\mathbb{P}(X \leq x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

and let X_n be a sequence of random variables with distribution function

$$\mathbb{P}(X_n \leq x) = \begin{cases} 1 - (1 - \frac{1}{n+1})^{(n+1)\lambda x} & \text{if } x > 0 \\ 0 & \text{elsewhere} \end{cases}.$$

For $x > 0$ we find that

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(X_n < x) &= \lim_{n\to\infty} 1 - (1 - \frac{1}{n+1})^{(n+1)\lambda x} \\ &= 1 - e^{-\lambda x} \\ &= \mathbb{P}(X < x) \end{aligned}.$$

For $x \leq 0$ the result is trivial. We conclude that $X_n \xrightarrow{d} X$. □

Another example of convergence in distribution that is of fundamental importance in statistics is given below.

3.3 Example (Central limit theorem)**.** Let $X_1, \ldots, X_n$ be i.i.d. random vectors with $\mathrm{E}(||X_i||^2) < \infty$ and $\mathrm{E}(X_i) = \mu$. If the average of these random vectors is given by $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$, then by the central limit theorem we find that $\sqrt{n}(\overline{X_n} - \mu) \xrightarrow{d} N(0, \mathrm{Cov}(X_i))$. □

The central limit theorem will reappear often in this thesis and in statistics in general.

The following lemma gives a few equivalent definitions of weak convergence, which will be useful in subsequent proofs and helps us in developing a broader understanding of weak convergence.

3.4 Lemma (Portmanteau)**.** *Let X be a random vector and X_n be a sequence of random vectors. The following statements are equivalent:*
(i) $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ *for all x at which the limit distribution function $x \mapsto \mathbb{P}(X \leq x)$ is continuous, i.e.* $X_n \xrightarrow{d} X$.
(ii) $\mathrm{E}(f(X_n)) \to \mathrm{E}(f(X))$ *for all bounded, continuous functions f.*
(iii) $\mathrm{E}(f(X_n)) \to \mathrm{E}(f(X))$ *for all bounded, Lipschitz functions f.*
(iv) $\liminf \mathrm{E}(f(X_n)) \geq \mathrm{E}(f(X))$ *for all nonnegative, continuous functions f.*
(v) $\liminf \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G)$ *for every open set G.*
(vi) $\limsup \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$ *for every closed set F.*
(vii) $\mathbb{P}(X_n \in B) \to \mathbb{P}(X \in B)$ *for all Borel sets B with $\mathbb{P}(X \in \delta B) = 0$, where $\delta B = \overline{B} - \mathring{B}$ is the boundary of B.*

Proof. For a proof, see [6], Lemma 2.2, p.6-7. □

Before studying some properties of weakly converging sequences of random vectors we look at another example.

3.5 Example. Let X be a Poisson distributed random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with parameter λ such that there exists some sequence $\{P_n\}_{n\in\mathbb{N}}$ consisting of real numbers in $[0,1]$ with $nP_n \to \lambda \in \mathbb{R}^+$.

Additionally, let Ω_n be the set of n-tuples consisting of ones and zeros for all $n \in \mathbb{N}$. We define $\mathcal{F}_n$ to be the power set of Ω_n and $\mathbb{P}_n$ to be a probability measure function that assigns the value $(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k}$ to each $\omega \in \Omega$ having k ones. Then $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ is a probability space on which a random variable X_n can be interpreted as the outcome of n Bernoulli trials with parameters $\frac{\lambda}{n}$. Equivalently, a random variable X_n in $(\Omega_n, \mathcal{F}_n, P_n)$ can be interpreted as a binomial distribution with parameter (n, P_n).

A basic result of probability theory is the Poisson limit theorem which states that

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}$$

if P_n is a sequence of real numbers in $[0,1]$ such that the sequence nP_n converges to a $\lambda \in \mathbb{R}^+$. Hence we can conclude that $X_n \xrightarrow{d} X$. □

The above is an example of a sequence X_n converging weakly to X even though all X_n and X are defined on different probability spaces. The rest of this section will be devoted to proving Prokhorov's theorem, which is a generalization of the Heine-Borel theorem to sequences of random vectors. Before stating the theorem we need a definition for uniform tightness.

3.6 Definition (Uniform tightness)**.** A set of random vectors $\{X_a \mid a \in A\}$ is called uniformly tight or bounded in probability if for every $\epsilon > 0$ there exists a $M > 0$ such that $\sup_{a \in A} \mathbb{P}(||X_a|| > M) < \epsilon$.

In other words, a sequence of k-dimensional random vectors X_n is uniformly tight if for every $\epsilon > 0$ there is a constant M such that $\mathbb{P}(||X_n|| > M) < \epsilon$ for all $n \in N$, where $||.||$ is the euclidean norm. Sometimes uniformly tight is called 'bounded in probability'. Such a name makes sense because a set of random vectors X_n is uniformly tight if there exists a compact set C such that $\mathbb{P}(X_n \in C) > 1 - \epsilon$ for each $n \in \mathbb{N}$. The following proposition relates uniform tightness of a sequence of random vectors to the distribution function it converges to. Note that $\mathbf{1}$ denotes the all-ones vector, so that $F(M\mathbf{1}) = \mathbb{P}(X_1 \leq M, \ldots, X_n \leq M)$ where $X = (X_1, \ldots, X_n)$.

3.7 Proposition. *Let $\epsilon > 0$ and let F be the distribution function corresponding to the random vector X. If $\mathbb{P}(||X|| > M) < \epsilon$, then $F(M\mathbf{1}) > 1 - \epsilon$ and $F(-M\mathbf{1}) < \epsilon$.*

Proof. Let $\epsilon > 0$ and assume that $\mathbb{P}(||X|| > M) < \epsilon$. If we have an X for which $\mathbb{P}(X > M\mathbf{1}) + \mathbb{P}(X \leq -M\mathbf{1}) < \epsilon$ then surely $\mathbb{P}(||X|| > M) < \epsilon$ as well, because $|x_1|, \ldots, |x_n| > M$ implies $||X|| > M$. Thus we have

$$\{x \in \Omega \mid \mathbb{P}(X > M\mathbf{1}) + \mathbb{P}(X \leq -M\mathbf{1}) < \epsilon\} \subseteq \{x \in \Omega \mid \mathbb{P}(||X|| > M) < \epsilon\}.$$

Using subaddivity of probability measures we can write

$$\begin{aligned} \mathbb{P}(||X|| > M) &\geq \mathbb{P}(X > M\mathbf{1}) + \mathbb{P}(X \leq -M\mathbf{1}) \\ &\geq \mathbb{P}(X > M\mathbf{1}) \\ &= 1 - \mathbb{P}(X \leq M\mathbf{1}) \\ &= 1 - F(M\mathbf{1}) \end{aligned}$$

which implies that $F(M\mathbf{1}) \geq 1 - \mathbb{P}(||X|| > M) > 1 - \epsilon$. For the second inequality, we take

$$\begin{aligned} \mathbb{P}(||X|| > M) &\geq \mathbb{P}(X > M\mathbf{1}) + \mathbb{P}(X \leq -M\mathbf{1}) \\ &\geq \mathbb{P}(X \leq -M\mathbf{1}) \\ &= F(-M\mathbf{1}) \end{aligned}$$

which implies $F(-M\mathbf{1}) < \epsilon$. □

Uniform tightness and weak convergence turn out to be closely related to each other. The following theorem explicates this.

3.8 Theorem (Prokhorov's theorem)**.** *If X_n is a sequence of random vectors in $\mathbb{R}^k$, then the following statements are true:*
(i) If $X_n \xrightarrow{d} X$ for some random vector X, then the sequence X_n is uniformly tight.
(ii) If X_n is uniformly tight, then there exists a subsequence X_{n_j} with $X_{n_j} \xrightarrow{d} X$ as $j \to \infty$ for some random vector X.

In order to prove this theorem we need another result which is known as Helly's lemma. We will prove this lemma before proving Prokhorov's theorem.

3.9 Lemma (Helly's lemma)**.** *Any sequence F_n consisting of cumulative distribution functions on $\mathbb{R}^k$ has a subsequence F_{n_j} with the property that $F_{n_j}(x) \to F(x)$ at all continuity points of some possibly defective distribution function F. Where with 'defective distribution function' we refer to a function having all the properties of a distribution function except that its limit at $-\infty$ may be greater than 0 and its limit at ∞ may be less than 1.*

Proof. Let $\mathbb{Q}^k = \{q_1, q_2, \dots\}$ be the set of vectors with rational coordinates, ordered in anyway we like. We consider the sequence $F_n(q_1)$. This sequence is contained in the closed and bounded set $[0, 1]$ since F_n is a distribution function. It follows that there exists a converging subsequence $F_{n_j}(q_1)$. We define the index of this subsequence as $\{n_j^1\}_{j=1}^{\infty}$ and the corresponding limit as $G(q_1)$. Similarly, we can take $F_n(q_2)$ and extract a subsequence with index $\{n_j^2\} \subseteq \{n_j^1\}$ that converges to some limit $G(q_2)$. We can repeat this process indefinitely for every q_i. Now, we define $n_j := n_j^j$, which has the property that $n_j \in \{n_j^i\}_{j=1}^{\infty}$ for all $i \in \mathbb{N}$. Using the sequence $\{n_j\}_{j=1}^{\infty}$ as an index we find that $F_{n_j}(q_i) \to G(q_i)$ for every $i \in \mathbb{N}$.

For the remainder of this proof, we take $q, q' \in \mathbb{Q}^k$. If $q \leq q'$, then $G(q) \leq G(q')$ because F_n is nondecreasing for all $n \in \mathbb{N}$. We define the function $F(x) = \inf_{q>x} G(q)$ i.e. $F(x)$ is equal to the infimum of all limits $G(q)$ with $q > x$, this ensures that $F(x)$ is nondecreasing.

We can also show that the function F is right-continuous. As a consequence of the definition of F there exists $q > x$ with $|G(q) - F(x)| = G(q) - F(x) < \epsilon$. It follows that for all $x \leq y \leq q$ we have $|F(x) - F(y)| = F(y) - F(x) < \epsilon$. In turn, this form of uniform continuity implies that F is right-continuous.

Consequently, for every $\epsilon > 0$ there exists $q < x < q'$ such that $|G(q) - G(q')| = G(q') - G(q) < \epsilon$. Since F is nondecreasing, we also have $G(q) \leq F(x) \leq G(q')$ for all $q \leq x \leq q'$. Combining these results gives us the inequality

$$G(q) = \lim_{j\to\infty} F_{n_j}(q) \leq \liminf_{j\to\infty} F_{n_j}(x) \leq \lim_{j\to\infty} F_{n_j}(q') = G(q')$$

and we conclude that $|\liminf_{j\to\infty}(x) - F(x)| < \epsilon$. By an analogous argument we can obtain

$$G(q) = \lim_{j\to\infty} F_{n_j}(q) \leq \limsup_{j\to\infty} F_{n_j}(x) \leq \lim_{j\to\infty} F_{n_j}(q') = G(q')$$

and thus $|\limsup_{j\to\infty} F_{n_j}(x) - F(x)| < \epsilon$. We conclude that for every continuity point of x of F we have $\lim_{j\to\infty} F_{n_j}(x) = F(x)$.

Now, we have proven the lemma for the one-dimensional case. For higher-dimensional cases we would still have to prove that the expressions defining masses of cells are nonnegative. When all corners of a cell are continuity points this property follows from the convergence of F_{n_j} to F and from F being a distribution function. After that the other cases follow by right continuity of distribution functions. For a slightly more comprehensive treatment of the higher-dimensional case we refer the reader to [6], Theorem 2.5, p.9. □

Proof (of Theorem 3.8)**.** (i) Let $\epsilon > 0$ and $X_n \xrightarrow{d} X$ for some sequence of random vectors X_n and random vector X. As mentioned before, there is an $M > 0$ such that $\mathbb{P}(||X|| \geq M) < \epsilon$. We assume that M is a continuity point of the distribution function of X, otherwise we replace M by M' with M' a continuity point of the distribution function. Because norms are continuous functions it follows from $\mathbb{P}(X_n \leq M\mathbf{1}) \to \mathbb{P}(X \leq M\mathbf{1})$ that $\mathbb{P}(||X_n|| \leq M) \to \mathbb{P}(||X|| \leq M)$. Similarly, $\mathbb{P}(||X_n|| \geq M) \to \mathbb{P}(||X|| \geq M)$ as well. So there must exist an N such that for all $n \geq N$ we have $\mathbb{P}(||X_n|| \geq M) < 2\epsilon$. We are only left with finitely many $n < N$. Each X_n corresponding to such an n is uniformly tight, so there is an M for each $n < N$ such that $\mathbb{P}(||X_n|| \geq M) < 2\epsilon$. Thus we can increase our fixed value of M in a way that ensures $\mathbb{P}(||X_n|| \geq M) < 2\epsilon$ for all $n \in \mathbb{N}$, from which it immediately follows that X_n is uniformly tight.

(ii) Given Helly's lemma and a sequence of random vectors X_n, we know that the sequence of distribution functions $\mathbb{P}(X_n \leq x) = F_n(x)$ contains a subsequence that converges weakly to a possibly defective distribution function $F(x)$. We only have to show that F is not defective, that is, we have to show that $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.

Let $\epsilon > 0$. We assume that X_n is uniformly tight, therefore, by Proposition 3.7, there exists an M such that $F_n(M) = \mathbb{P}(X_n \leq M\mathbf{1}) > 1 - \epsilon$ for all $n \in \mathbb{N}$. Since F is nondecreasing we can always find an M large enough such that M is also a continuity point of F. We find that

$$1 \geq F(M) = \lim_{j\to\infty} F_{n_j}(M) \geq 1 - \epsilon$$

which means that $\lim_{x\to\infty} F(x) = 1$.

We can treat the other limit in a similar way. By uniform tightness and Proposition 3.7 there exists an $M > 0$ such that $F(-M) < \epsilon$, if necessary, we increase M until it is a continuity point of F. Now, we find

$$0 \leq F(-M) = \lim_{j\to\infty} F_{n_j}(-M) \leq \epsilon$$

which implies that $\lim_{x\to-\infty} F(x) = 0$, so F is not defective. □

3.2 Convergence in probability and convergence almost surely.

We begin by giving the definitions of the two remaining modes of convergence.

3.10 Definition (Convergence in probability)**.** Let X be a random vector and let X_n be a sequence of random vectors. Consider the topology induced by some metric $d(x, y)$ on $\mathbb{R}^k$. If

$$\mathbb{P}(d(X_n, X) > \epsilon) \to 0$$

for all $\epsilon > 0$, then we say that X_n converges in probability to X, denoted by $X_n \xrightarrow{P} X$.

The notation $d(X_m, X) \xrightarrow{P} 0$ is equivalent to $X_n \xrightarrow{P} X$ and will sometimes be used. The following definition is quite similar to the one above, but it will turn out to be an even stronger form of convergence.

3.11 Definition (Convergence almost surely)**.** Let X be a random vector and let X_n be a sequence of random vectors. Consider the topology induced by some metric $d(x, y)$ on $\mathbb{R}^k$. If

$$\mathbb{P}(\lim_{n\to\infty} d(X_n, X) = 0) = 1,$$

then we say that X_n converges almost surely to X, denoted by $X_n \xrightarrow{as} X$.

Contrary to convergence in distribution, for both convergence in probability and convergence almost surely it is necessary that all random vectors of the sequence X_n and X are defined on the same probability space. Otherwise the distance $d(X_n, X)$ would not make any sense.

3.12 Example (Weak/strong law of large numbers)**.** Let $X_1, X_2, \ldots$ be an infinite sequence of i.i.d. Lebesgue integrable random vectors with $\mathrm{E}(X_i) = \mu$, where μ is a vector. We denote the sample average as $\overline{X_n}$. Then the weak law of large numbers states that

$$\mathbb{P}(d(\overline{X_n}, \mu) > \epsilon) \to 0$$

and the strong law of large numbers states that

$$\mathbb{P}(\lim_{n\to\infty} d(\overline{X_n}, \mu) = 0) = 1.$$

These laws correspond to convergence in probability and convergence almost surely respectively.

It turns out that continuous functions preserve all of the three modes of convergence. The following theorem formalizes this and will turn out to be very useful.

3.13 Theorem (Continuous mapping theorem)**.** *If $g : \mathbb{R}^k \to \mathbb{R}^m$ is a function that is continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$, then the following statements are true.*
(i) If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.
(ii) If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.
(iii) If $X_n \xrightarrow{as} X$, then $g(X_n) \xrightarrow{as} g(X)$.

Proof. (i): Let f be any bounded continuous function and g be any continuous function. We define $h = f \circ g$. Since f and g are both continuous and f is bounded the function h is also bounded and continuous. Now, if $X_n \xrightarrow{d} X$, then by (ii) of Lemma 3.4 we also have $\mathrm{E}(h(X_n)) \to \mathrm{E}(h(X))$, which is equivalent to $\mathrm{E}(f(g(X_n))) \to \mathrm{E}(f(g(X)))$. If we apply (ii) of Lemma 3.4 again we can conclude that $g(X_n) \xrightarrow{d} g(X)$.

(ii): Suppose $X_n \xrightarrow{P} X$ and $\epsilon > 0$. For every $\delta > 0$ we define

$$B_\delta = \{x \in \mathbb{R}^k \mid \exists y \in \mathbb{R}^k : d(x, y) < \delta \text{ and } d(g(x), g(y)) > \epsilon\}.$$

The set B_δ consists of all $x \in \mathbb{R}^k$ within the δ-neighborhood that map outside the ϵ-neighborhood in $\mathbb{R}^m$. From the continuity of g we obtain $\lim_{\delta\to 0} B_\delta = \emptyset$. Now, if we assume that $d(g(X_n), g(X)) > \epsilon$, then either $d(X, X_n) \geq \delta$ or $X \in B_\delta$. The probabilities of these events can be described by

$$\mathbb{P}(d(g(X_n), g(X)) > \epsilon) \leq \mathbb{P}(d(X_n, X) \geq \delta) + \mathbb{P}(X \in B_\delta),$$

notice that this is an inequality since we do not subtract the intersection of both events on the right-hand side. Since $X_n \xrightarrow{P} X$, the probability $\mathbb{P}(d(X_n, X) \geq \delta) \to 0$ for any $\delta > 0$. Now, as mentioned before, if δ becomes arbitrarily small the set B_δ converges to $\emptyset$ and therefore $\lim_{\delta \to 0} \mathbb{P}(X \in B_\delta) = 0$. Since both terms on the right-hand side of the inequality converge to 0, the term $\mathbb{P}(d(g(X_n), g(X)) > \epsilon) \to 0$ as well, which means that $g(X_n) \xrightarrow{P} g(X)$.

(iii): Given that a function g is continuous on a set C such that $\mathbb{P}(X \in C) = 0$. Using the continuity of g we find that

$$\begin{aligned} 1 = \mathbb{P}(\lim_{n\to\infty} d(X_n, X) = 0) &= \mathbb{P}(\lim_{n\to\infty} X_n = X) \\ &= \mathbb{P}(\lim_{n\to\infty} g(X_n) = g(X)) \\ &= \mathbb{P}(\lim_{n\to\infty} d(g(X_n), g(X)) = 0) \end{aligned}$$

which means that $g(X_n) \xrightarrow{as} g(X)$. □

An important implication of the continuous mapping theorem is that if we are interested in $\phi(\theta)$ and we have a sequence of estimators T_n that converges to some parameter θ, then $\phi(T_n)$ converges to $\phi(\theta)$, provided that ϕ is continuous. We further demonstrate the utility of the continuous mapping theory by giving a simple example.

3.14 Example. Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathrm{E}(X_i) = \mu < \infty$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. The central limit theorem states that $\frac{\sqrt{n}}{\sigma}(\overline{X_n} - \mu) \xrightarrow{d} N(0, 1)$. Now, it follows from the continuous mapping theorem that

$$\left(\frac{\sqrt{n}}{\sigma}(\overline{X_n} - \mu)\right)^2 = \frac{n}{\sigma^2}(\overline{X_n} - \mu)^2 \xrightarrow{d} \chi^2_1.$$

□

3.3 Relationships between modes of convergence

Now that we have defined the three relevant modes of convergence we can investigate the relationships between them. As mentioned in previous sections, almost sure convergence is the strongest mode of convergence, while convergence in distribution is the weakest form of convergence. In the following theorem we will formally state and prove this among other things.

3.15 Theorem. *Let X_n and Y_n be sequences of random vectors, X and Y be random vectors and $c \in \mathbb{R}$ be a constant. The following statements are true:*
(i) If $X_n \xrightarrow{as} X$, then $X_n \xrightarrow{P} X$.
(ii) If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{d} X$.
(iii) If $X_n \xrightarrow{as} X$, then $X_n \xrightarrow{d} X$.
(iv) The sequence $X_n \xrightarrow{P} c$ if and only if $X_n \xrightarrow{d} c$.
(v) If $X_n \xrightarrow{d} X$ and $d(X_n, Y_n) \xrightarrow{P} 0$, then $Y_n \xrightarrow{d} X$.
(vi) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$, then $(X_n, Y_n) \xrightarrow{d} (X, c)$.
(vii) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $(X_n, Y_n) \xrightarrow{P} (X, Y)$.

Proof. (i) Let $\epsilon > 0$. We define the decreasing sequence of sets

$$A_n = \cup_{m \geq n} \{d(X_n, X) > \epsilon\}.$$

If $X_n(\omega) \to X(\omega)$, for almost every $\omega \in \Omega$, as is the case when $X_n \xrightarrow{as} X$, then $\mathbb{P}(A_n) \to 0$. Therefore, if $X_n \xrightarrow{as} X$, then $\mathbb{P}(d(X_n, X) > \epsilon) \leq \mathbb{P}(A_n) \to 0$. We conclude that $X_n \xrightarrow{P} X$.

(ii). Let $Y_n = X$ for all $n \in \mathbb{N}$. We apply (v): From $d(X_n, X) \xrightarrow{P} 0$ and $Y_n = X \xrightarrow{d} X$ it follows that $X_n \xrightarrow{d} X$.

(iii). This follows directly from (i) and (ii).

(iv). We assume that $X_n \xrightarrow{P} c$, then by (ii) we have $X_n \xrightarrow{d} c$ as well. For the other implication we assume that $X_n \xrightarrow{d} c$ and let $B(c, \epsilon)$ be the open ball with radius $\epsilon > 0$ around c such that $\mathbb{P}(X_n \notin B(c, \epsilon)) < \epsilon$. Note that the complement $B(c, \epsilon)^C$ of this open ball is closed. We can write

$$\mathbb{P}(d(X_n, c) \geq \epsilon) = \mathbb{P}(X_n \in B(c, \epsilon)^C) = 0$$

and using (vi) of the portmanteau lemma we find

$$\limsup \mathbb{P}(X_n \in B(c, \epsilon)^C) \leq \mathbb{P}(c \in B(c, \epsilon)^C) = 0$$

which implies that $\mathbb{P}(d(X_n, c) > \epsilon) \to 0$, so $X_n \xrightarrow{P} c$.

(v). Suppose $X_n \xrightarrow{d} X$ and $d(X_n, Y_n) \xrightarrow{P} 0$. Let f be a Lipschitz continuous function with range $[0, 1]$, so that f is bounded. Then

$$\begin{aligned} |\mathrm{E}(f(X_n)) - \mathrm{E}(f(Y_n))| &= |\mathrm{E}(f(X_n) - f(Y_n))| \\ &\leq \mathrm{E}(|f(X_n) - f(Y_n)|) \\ &\leq \mathrm{E}(d(X_n, Y_n)) \\ &= \mathrm{E}(d(X_n, Y_n) \mathbb{1}_{\{d(X_n, Y_n) \leq \epsilon\}}) + \mathrm{E}(d(X_n, Y_n) \mathbb{1}_{\{d(X_n, Y_n) > \epsilon\}}) \end{aligned} \tag{3.1}$$

The first term in the last expression can be rewritten as

$$\begin{aligned}\mathrm{E}(d(X_n,Y_n)\mathbb{1}_{\{d(X_n,Y_n)\leq\epsilon\}}) &\leq \epsilon\mathrm{E}(\mathbb{1}_{\{d(X_n,Y_n)\leq\epsilon\}})\\ &= \epsilon\mathbb{P}(d(X_n,Y_n)\leq\epsilon)\\ &\leq \epsilon\end{aligned}$$

and thus can be made arbitrarily small. For the second term in the last expression of (3.1) we find

$$\begin{aligned}\mathrm{E}(d(X_n,Y_n)\mathbb{1}_{\{d(X_n,Y_n)>\epsilon\}}) &\leq 2\mathrm{E}(\mathbb{1}_{\{d(X_n,Y_n)>\epsilon\}})\\ &= 2\mathbb{P}(d(X_n,Y_n)>\epsilon)\end{aligned}$$

because f is bounded. By Definition 3.10 it follows that this term converges to zero for every $\epsilon > 0$. Hence we conclude that $\mathrm{E}(f(X_n))$ and $\mathrm{E}(f(Y_n))$ have the same limit. Now, it follows from the Portmanteau Lemma (ii) that $Y_n \xrightarrow{d} X$.

(vi). We assume $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$. We use (v) again: from

$$d((X_n,Y_n),(X_n,c)) = d(Y_n,c) \xrightarrow{P} 0$$

it follows that we only have to prove that $(X_n,c) \xrightarrow{d} (X,c)$. Let $f:(x,y)\mapsto f(x,y)$ be some bounded continuous function, then $x\mapsto f(x,c)$ is bounded and continuous as well. Now, from $X_n \xrightarrow{d} X$ and (ii) of the portmanteau lemma it follows that $\mathrm{E}(f(X_n,c)) \to \mathrm{E}(f(X,c))$. Applying (ii) of the portmanteau lemma again gives $(X_n,c) \xrightarrow{d} (X,c)$, which proves that $(X_n,Y_n) \xrightarrow{P} (X,c)$.

(vii). Let $k_1,k_2\in\mathbb{N}$ such that k_1 is the dimension of the vector X_n and k_2 is the dimension of the vector Y_n. Let $d:\mathbb{R}^{k_1+k_2}\times\mathbb{R}^{k_1+k_2}\to\mathbb{R}^{k_1+k_2}$ be the metric we work with. Whenever a vector has dimension k_1 we consider the last k_2 coordinates to be equal to 0. Similarly, whenever a vector has dimension k_2, we consider the first k_1 coordinates to be 0. We assume the antecedent of our statements, which is equivalent to $\mathbb{P}(d(X_n,X)>\epsilon)\to 0$ and $\mathbb{P}(d(Y_n,Y)>\epsilon)\to 0$, we find that

$$\mathbb{P}(d(X_n,X)+d(Y_n,Y)>2\epsilon)\to 0. \tag{3.2}$$

By the triangle inequality we find that

$$d(x_1,x_2)+d(y_1,y_2)\geq d((x_1,y_1),(y_1,y_2)) \tag{3.3}$$

for all $x_1,x_2\in\mathbb{R}^{k_1}$ and $y_1,y_2\in\mathbb{R}^{k_2}$. Combining (3.2) and (3.3) gives us

$$\mathbb{P}(d((X_n,Y_m),(X,Y))>2\epsilon)\to 0,$$

which concludes our proof. □

The theorem above is very useful and we should examine it more closely. We remark that by statement (i) the statements (iv)-(vii) are also true if we replace convergence in probability with convergence almost surely.

The last statement of the theorem can be understood as follows: convergence in probability of a sequence of random vectors is equivalent to convergence of each of the components of the vectors. We should observe that this statement is not necessarily true for sequences of random vectors that converge in distribution. The separate components of a random vector can be dependent or independent on each other, this means that the distribution of all components taken separately does not determine the joint distribution. We illustrate this with an example.

3.16 Example. Let X, Y have the standard normal distribution $N(0,1)$. Define $X_n \sim N(0,1)$ for all $n \in \mathbb{N}$ and $Y_n = -X_n$. It is obvious that $X_n \xrightarrow{d} X$, and by symmetry of the normal distribution $Y_n \xrightarrow{d} Y$ as well. However, it is not true that $(X_n, Y_n) \xrightarrow{d} (X, Y)$ due to the dependence of Y_n on X_n. Instead, we find that $(X_n, Y_n) \xrightarrow{d} (X, -X)$. The distribution of $(X, -X)$ is very different from that of (X, Y). Thus the distributions of the components taken separately do not necessarily determine the joint distribution. □

Before moving on, we consider statement (vi) of Theorem 3.15 because it has some interesting implications. Let X_n and Y_n be sequences of random vectors, X a random vector and $c \in \mathbb{R}$ a constant such that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$. Also, let g be any function that is continuous on the subset of $\mathbb{R}^k \times \{c\}$ in which (X, c) takes its values. Then by the continuous mapping theorem we find that $g(X_n, Y_n) \xrightarrow{d} g(X, c)$. The following lemma is known as Slutsky's lemma and describes some of the applications of statement (vi).

3.17 Lemma (Slutsky's lemma)**.** *Let X_n and Y_n be sequences of random vectors, X a random vector and $c \in \mathbb{R}$ a constant. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$, then the following statements are true:*
(i) $X_n + Y_n \xrightarrow{d} X + c$.
(ii) $Y_n X_n \xrightarrow{d} cX$.
(iii) If $c \neq 0$, then $Y_n^{-1} X_n \xrightarrow{d} c^{-1} X$.

Some remarks are in order. In virtue of Theorem 3.15 (iv), the above lemma would still hold if we have $Y_n \xrightarrow{d} c$ instead of $Y_n \xrightarrow{P} c$. For (i) to be meaningful, the vector c must have the same dimension as X. A similar remark goes for (ii), only here c and Y_n must either be scalars or matrices

having dimensions such that cX makes sense. The same goes for (iii), as long as $\det(c) \neq 0$. All parts of the lemma above directly follow from Theorem 3.15 (vi) with subsequent application of the continuous mapping theorem (Theorem 3.13).

We conclude this section with an example that illustrates how the different theorems and lemmas of this chapter can be applied.

3.18 Example. If $X_1, X_2, \ldots$ are i.i.d. random variables with $\mathrm{E}(X_1) = \mu = 0$ and $\mathrm{E}(X_1^2) < \infty$, then the t-statistic $\sqrt{n}\frac{\overline{X_n}}{S_n}$, where S_n^2 is the sample variance, is asymptotically standard normal. That is, the t-statistic converges weakly to $N(0,1)$.

Using the lemma's and theorems from this chapter we can prove this claim. Firstly, we take a look at

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2\right),$$

by the weak law of large numbers and the continuous mapping theorem we know that $\overline{X_n}^2 \xrightarrow{P} E(X_i)^2 = \mu^2 = 0$. What remains within the parenthesis converges in probability as well, we can see this by applying the weak law of large numbers and the continuous mapping theorem again:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} \mathrm{E}(X_i^2).$$

Obviously $\frac{n}{n-1} \to 1$. By combining these three results, we find

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2\right) \xrightarrow{P} 1(\mathrm{E}(X_i^2) - \mathrm{E}(X_i)^2) = \mathrm{Var}(X_i) = \sigma^2.$$

By applying the continuous map theorem again we find that $S_n \xrightarrow{P} \sigma$. We shift our focus to $\sqrt{n}\overline{X_n}$. By the central limit theorem we find that $\sqrt{n}\overline{X_n} = \sqrt{n}(\overline{X_n} - \mu) \xrightarrow{d} N(0, \sigma^2)$. Using (iii) of slutsky's lemma we can combine these two limits and obtain

$$\sqrt{n}\frac{\overline{X_n}}{S_n} \xrightarrow{d} \frac{N(0,\sigma^2)}{\sigma} = N(0,1),$$

and thus the t-statistic is asymptotically standard normal. □

3.4 Stochastic order symbols

Quite often we encounter sequences of random vectors that converge to zero in probability or are uniformly tight. We can generalize the o and O symbols that we have for deterministic sequences to include sequences of random vectors as well.

3.19 Definition (Stochastic o and O symbols)**.** Let X_n, Y_n and R_n be sequences of random vectors, then
(i) $X_n = o_P(R_n)$ is equivalent to $X_n = Y_n R_n$ and $Y_n \xrightarrow{P} 0$.
(ii) $X_n = O_P(R_n)$ is equivalent to $X_n = Y_n R_n$ and Y_n is uniformly tight.

For instance, we denote a sequence of random vectors that converges to zero in probability as $o_P(1)$ and we denote a sequence of random vectors that is uniformly tight as $O_P(1)$. The following lemma states some equalities that hold for o and O in calculus and generalizes them for the stochastic o_P and O_P.

3.20 Lemma. *Let R_n be a sequence of random vectors. Then*
(i) $o_P(1) + o_P(1) = o_P(1)$;
(ii) $O_P(1) + O_P(1) = O_P(1)$;
(iii) $o_P(1) + O_P(1) = O_P(1)$;
(iv) $O_P(1)o_P(1) = o_P(1)$;
(v) $o_P(R_n) = R_n o_P(1)$;
(vi) $O_P(R_n) = R_n O_P(1)$ and;
(vii) $o_P(O_P(1)) = o_P(1)$.

Proof. In the following proofs let all X_n and Y_n be sequences of random vectors.

(i) Let $X_n \xrightarrow{P} 0$ and $Y_n \xrightarrow{P} 0$. Define $Z_n = X_n + Y_n$. Then $Z_n \xrightarrow{P} 0$ by the continuous mapping theorem.

(ii) Let X_n and Y_n be uniformly tight. Let $\epsilon > 0$, then there exist $M_x, M_y \in \mathbb{R}$ such that $\sup_n \mathbb{P}(||X_n|| > M_X), \sup_n \mathbb{P}(||Y_n|| > M_Y) < \epsilon$. By the subadditivity of norms we find that

$$\{||X_n|| \leq M_X\} \cap \{||Y_n|| \leq M_Y\} \subseteq \{||X_n + Y_n|| \leq M_X + M_Y\},$$

this implies that

$$\mathbb{P}(||X_n + Y_n|| > M_X + M_Y) \leq \mathbb{P}(||X_n|| > M_X) + \mathbb{P}(||Y_n|| > M_Y) < 2\epsilon.$$

Therefore we also have

$$\sup_n \mathbb{P}(||X_n + Y_n|| > M_X + M_Y) < 2\epsilon.$$

Hence $X_n + Y_n$ is uniformly tight as well.

(iii) Let $X_n \xrightarrow{P} 0$ and let Y_n be uniformly tight. By Theorem 3.15 and Prokhorov's Theorem, the sequence X_n is uniformly tight. The result now follows from (ii).

(iv) Let X_n be uniformly tight and let $Y_n \xrightarrow{P} 0$. We need to show that $\mathbb{P}(||X_nY_n|| > \epsilon) \to 0$. Suppose $\epsilon > 0$, we write

$$\mathbb{P}(||X_nY_n|| > \epsilon) = \mathbb{P}(||X_nY_n|| > \epsilon, ||X_n|| \leq M) + \mathbb{P}(||X_nY_n|| > \epsilon, ||X_n|| > M).$$

For the first term, we find

$$\mathbb{P}(||X_nY_n|| > \epsilon, ||X_n|| \leq M) \leq P(||MY_n|| > \epsilon) = \mathbb{P}(||Y_n|| > \frac{\epsilon}{M}) \to 0.$$

Notice here that we implicitly assume that $M > 0$. This is no problem because if $M \leq 0$, then $\mathbb{P}(||X_nY_n|| > \epsilon, ||X_n|| \leq M) = 0$. For the second term, we get

$$\mathbb{P}(||X_nY_n|| > \epsilon, ||X_n|| > M) \leq \mathbb{P}(||X_n|| > M) < \epsilon$$

by uniform tightness, so this term converges to zero as well. Hence $\mathbb{P}(||X_nY_n|| > \epsilon) \to 0$.

(v)-(vi) These follow directly from Definition 3.19.

(vii) By definition $o_P(O_P(1)) = Y_nO_P(1) = o_P(1)O_P(1)$ where $Y_n \xrightarrow{P} 0$. The result now follows immediately from (iv). □

The following lemma is a bit more intricate but will be used in the proof of the main result of the next chapter.

3.21 Lemma. *Let R be a function with domain $D \in \mathbb{R}^k$ and $R(0) = 0$. If X_n is a sequence of random vectors with $X_n \in D$ for all $n \in \mathbb{N}$ and $X_n \xrightarrow{P} 0$. Then for every $p > 0$ we have*
(i) if $R(h) = o(||h||^p)$ as $h \to 0$, then $R(X_n) = o_P(||X_n||^p)$;
(ii) if $R(h) = O(||h||^p)$ as $h \to 0$, then $R(X_n) = O_P(||X_n||^p)$.

Proof. Let R and X_n be as stated above. We define $g(h) := \frac{1}{||h||^p}R(h)$ for $h \neq 0$ and $g(0) = 0$, so that $R(X_n) = g(X_n)||X_n||^p$. We prove both statements:

(i) We assume the antecedent of (i). Since g is continuous at zero, application of the continuous mapping theorem gives us $g(X_n) \xrightarrow{P} g(0) = 0$. The desired result follows.

(ii) We assume the antecedent of (ii), then there exist M and $\delta > 0$ such that if $||h|| \leq \delta$, then $|g(h)| \leq M$. So $\mathbb{P}(|g(X_n)| > M) \leq \mathbb{P}(||X_n|| > \delta) \to 0$, which means that $g(X_n)$ is uniformly tight. Our result follows immediately. □

Chapter 4

The Delta Method

In this chapter we treat the Delta method and some of its applications. In the first section we will prove the Delta method and give some examples, including variance stabilizing transformations. In the second section we will briefly take a look at parametric models and how we should interpret the results of asymptotic statistics. Lastly, we take a look at a particular kind of estimator, namely moment estimators.

4.1 Main result

Suppose that we have an estimator T_n for some parameter θ on hand and we are interested in $\phi(\theta)$. By the continuous mapping Theorem we find that $\phi(T_n)$ converges to $\phi(\theta)$ in the same mode as T_n converges to θ, provided that ϕ is continuous.

A similar problem arises often: if we have a limit distribution to which some sequence $\sqrt{n}(T_n - \theta)$ converges, does $\sqrt{n}(\phi(T_n) - \phi(\theta))$ converge as well? According to the Delta method this is the case for convergence in distribution, provided that the function ϕ is differentiable at θ.

We will introduce, state and prove the Delta method and subsequently give a few examples of its application. Most of these examples will not directly concern parameter estimation. However, later on the Delta method will prove to be of great importance for the estimation of parameters. Therefore we treat it comprehensively.

The Delta method uses a Taylor expansion to approximate random vectors $\phi(T_n)$. Such an expansion looks like $\phi(\theta+\delta)-\phi(\delta) = \delta\phi'(\theta)+\ldots$, notice that we have left out the higher order terms of the Taylor expansion and are reducing it to a linear approximation in this way. Figure 4.1 graphically explains the Delta method.

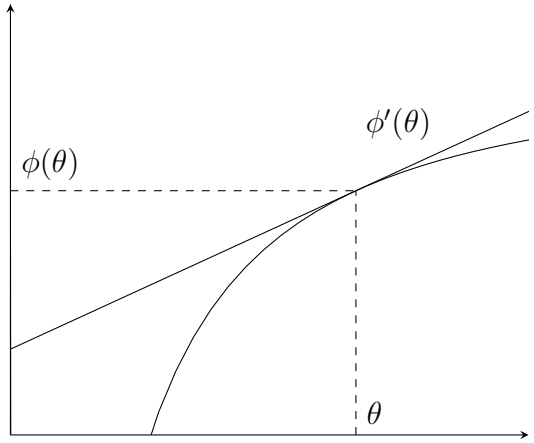

Figure 4.1: Graphical representation of the Delta method.

On the horizontal axis we find the true value of θ. For sufficiently large n, we find T_n in a small neighborhood of θ with high probability. The continuous function ϕ maps θ to a point on the vertical axis, by continuity the values of $\phi(T_n)$ are also arbitrarily close to $\phi(\theta)$ for sufficiently large n with a high probability. A measure for the difference of the convergence of $\phi(T_n)$ is given by the slope of the tangent $\phi'(\theta)$. This suggests that

$$\phi'(\theta)\sqrt{n}(T_n - \theta) \approx \sqrt{n}(\phi(T_n) - \phi(\theta).$$

The figure above depicts the one-dimensional case, but we are also interested in the case where T_n and θ are vector-valued. In order to do this, we need the notion of total differentiability. We mentioned this in the preliminaries and we assume that the reader is familiar with this concept. In addition to what is stated in the preliminaries, we like to mention that in this context it might be better to think of the total derivative as a linear approximation $h \mapsto \phi'(h)$ to the function $h \mapsto \phi(\theta + h) - \phi(\theta)$ than as a matrix of partial derivatives. We proceed to the actual theorem.

4.1 Theorem (The Delta method)**.** *Let $\theta \in D \subseteq \mathbb{R}^k$ and let $\phi : D \to \mathbb{R}^m$ be a function that is differentiable at θ. Let T_n be a sequence of random vectors taking their values in D and let r_n be a sequence of numbers in $\mathbb{R}$ with $r_n \to \infty$. If $r_n(T_n - \theta) \xrightarrow{d} T$, then $r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \phi'(\theta)T$. Moreover, $r_n(\phi(T_n) - \phi(\theta)) - \phi'(\theta)r_n(T_n - \theta)$ converges in probability to zero.*

Proof. Let ϕ, r_n and θ be as described above. Suppose that $r_n(T_n - \theta) \xrightarrow{d} T$. By Prokhorov's theorem the sequence $r_n(T_n - \theta)$ is uniformly tight. Because $r_n \to \infty$ we also know that $T_n - \theta \xrightarrow{d} 0$, if this would not have been the case

then $r_n(T_n - \theta)$ would diverge and that is a contradiction. Consequently, by Theorem 3.15 the sequence $T_n - \theta \xrightarrow{P} 0$ as well. Since ϕ is differentiable in θ, we find that for the function $R(h) = \phi(\theta + h) - \phi(\theta) - \phi'(\theta)h$ we have

$$\lim_{h \to 0} \frac{||\phi(\theta + h) - \phi(\theta) - \phi'(\theta)h||}{||h||} = 0,$$

that is, $R(h) = o(||h||)$ as $h \to \infty$. By Lemma 3.21 we have

$$R(T_n - \theta) = \phi(T_n) - \phi(\theta) - \phi'(\theta)(T_n - \theta) = o_P(||T_n - \theta||).$$

Multiplying on both sides with r_n results in

$$\begin{aligned} r_n R(T_n - \theta) &= r_n(\phi(T_n) - \phi(\theta) - \phi'(\theta)(T_n - \theta)) \\ &= r_n(\phi(T_n) - \phi(\theta)) - \phi'(\theta) r_n(T_n - \theta) \\ &= o_p(1) \\ &= o_P(O_P(1)) \\ &= o_P(r_n||T_n - \theta||) \\ &= r_n o_P(||T_n - \theta||) \end{aligned}$$

by Prokhorov's Theorem and Lemma 3.20 (vii). This proves that

$$r_n(\phi(T_n) - \phi(\theta)) - \phi'(\theta) r_n(T_n - \theta) \xrightarrow{P} 0,$$

or equivalently

$$d(r_n(\phi(T_n) - \phi(\theta)), \phi'(\theta) r_n(T_n - \theta)) \xrightarrow{P} 0.$$

Now, we only have to show that $r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \phi'(\theta)T$. Since the derivative map is by definition matrix multiplication and matrix multiplication is always continuous, applying the continuous mapping Theorem to our assumption gives $\phi'(\theta) r_n(T_n - \theta) \xrightarrow{d} \phi'(\theta)T$. Now, applying Theorem 3.15 (v) yields the desired result

$$r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \phi'(\theta)T.$$

□

Thus the Delta method can be used to turn one statement of weak convergence into another statement of weak convergence, much like the continuous mapping theorem or Slutsky's Lemma. Another way of viewing the Delta

method is as a way to obtain the limit distribution of a statistic that is a function of another statistic for which we already obtained the limit distribution. Most of the applications of the Delta method revolve around some normally distributed random variable or random vector with mean zero, which is obtained by applying the central limit theorem. Example 4.2 provides a simple example of this.

4.2 Example. Suppose that we are interested in the limit distribution of our sample variance. Let $X_1, \ldots, X_n$ be a sample consisting of n observations. We define the (biased) sample variance as $S_n^2 = \frac{1}{n+1}\sum_{i=1}^n (X_i - \overline{X_n})^2$. Notice that the bias here does not matter since $\frac{n}{n-1} \to 1$. Defining our sample variance like this makes it similar to the formula of the population variance, therefore the function $\phi(x, y) = y - x^2$ gives the sample variance if we take $\phi(\overline{X_n}, \overline{X_n^2})$. Remember that $\overline{X_n} = \frac{1}{n+1}\sum_{i=1}^n X_i$ by assumption. Now, suppose that our sample is taken from a distribution with $\mathrm{E}(X^4) < \infty$, and denote the first four moments as α_1, α_2, α_3 and α_4. By the multivariate central limit theorem we obtain

$$\sqrt{n}\left(\begin{pmatrix}\overline{X_n}\\ \overline{X_n^2}\end{pmatrix} - \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}\right) \xrightarrow{d} N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2\\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2\end{pmatrix}\right).$$

The map $\phi(x, y)$ is differentiable with derivative map $\phi'(x, y) = \begin{pmatrix}-2x & 1\end{pmatrix}$. In particular at (α_1, α_2) the derivative is given by $\begin{pmatrix}-2\alpha_1 & 1\end{pmatrix}$. Applying the Delta method gives us

$$\sqrt{n}(\phi(\overline{X_n}, \overline{X_n^2}) - \phi(\alpha_1, \alpha_2)) = \sqrt{n}(S_n^2 - \mathrm{Var}(X)) \xrightarrow{d} N(0, \sigma^2).$$

Since we are interested in the sample variance, we calculate σ^2 exactly:

$$\begin{aligned}\sigma^2 &= \begin{pmatrix}-2\alpha_1 & 1\end{pmatrix}\begin{pmatrix}\alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2\\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2\end{pmatrix}\begin{pmatrix}-2\alpha_1\\ 1\end{pmatrix}\\ &= \alpha_4 - 4\alpha_1\alpha_3 + 8\alpha_1^2\alpha_2 - \alpha_2^2 - 4\alpha_1^4.\end{aligned}$$

Alternatively, we could replace all X_i with the centered variables $Y_i = X_i - \alpha_1$. In this case, the first moment of Y would be zero. This would reduce our expression for the sample variance to $\sigma^2 = \mu_4 - \mu_2^2$, which can be found using the expression for σ^2 we obtained earlier by taking $\alpha_1 = 0$. Also, instead of remarking that $\frac{n}{n-1} \to 1$ we could also conclude that the same result is valid for the unbiased sample variance by applying Slutksy's lemma. □

The following example is an extension of Example 4.2 and utilizes the Delta method again.

4.3 Example. We are interested in the joint limit distribution of the sample variance and the T-statistic, i.e. the pair $(S_n^2, \overline{X_n}/S_n)$. We consider the same (biased) sample variance as in Example 4.2. We define

$$\phi(x,y) = \left(y - x^2, \frac{x}{\sqrt{y-x^2}} \right)$$

and observe that for this function $\phi(\overline{X_n}, \overline{X_n^2}) = (S_n^2, \overline{X_n}/S_n)$. The function ϕ is defined and differentiable on $\{(x,y) \mid y - x^2 > 0\} \subseteq \mathbb{R}^2$ with derivative map

$$\phi'(x,y) \begin{pmatrix} -2x & 1 \\ \frac{x^2}{(y-x^2)^{\frac{3}{2}}} + \frac{1}{\sqrt{y-x^2}} & \frac{-x}{2(y-x^2)^{\frac{3}{2}}} \end{pmatrix}.$$

Once again, let α_1, α_2, α_3 and α_4 be the first four moments of X respectively. Applying the Delta method gives us

$$\sqrt{n}(S_n^2 - \sigma^2, \frac{\overline{X_n}}{S_n} - \frac{\alpha_2}{\sigma}) \xrightarrow{d} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right)$$

where the covariance matrix Σ is given by

$$\phi'(\alpha_1, \alpha_2) \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} (\phi'(\alpha_1, \alpha_2))^T.$$

□

Variance stabilizing transformations are another useful application of the Delta method. The concept of variance stabilizing transformations is of less importance for the aim of this thesis. However, since this application illustrates the significance of the Delta method we treat it briefly. The following scenario exemplifies what such a transformation entails.

Suppose we have a sequence of estimators T_n such that $\sqrt{n}(T_n - \theta) \xrightarrow{d_\theta} N(0, \sigma^2(\theta))$ for all possible values of θ. Then asymptotic confidence intervals of level $1 - \alpha$ for θ are given by

$$\left(T_n - z_{\alpha/2}\frac{\sigma(\theta)}{\sqrt{n}}, T_n + z_{\alpha/2}\frac{\sigma(\theta)}{\sqrt{n}} \right).$$

However, this confidence interval is problematic: in order to obtain the boundaries of the confidence interval we need to know the value of θ, but we do not know θ. In other words; we cannot use θ in the procedure of estimating θ.

A possible solution to this problem is to replace $\sigma(\theta)$ with an estimator that is independent of θ. For a consistent sequence of estimators the confidence level will remain $1-\alpha$. Another solution is the application of a variance stabilizing transformation, which generally leads to better results.

Thus the aim of our variance stabilizing transformation is to transform $\sigma^2(\theta)$ such that it is independent of θ. We can do this by transforming θ to another parameter $\phi(\theta)$. An estimator for $\phi(\theta)$ is $\phi(T_n)$ provided ϕ is continuous. We choose ϕ such that it is differentiable and $\phi'(\theta)\sigma(\theta) = 1$. According to the Delta method $\sqrt{n}(\phi(T_n) - \phi(\theta)) \xrightarrow{d_\theta} N(0, \phi'(\theta)^2\sigma^2(\theta)) = N(0,1)$, which solves our initial problem.

The function

$$\phi(\theta) = \int \frac{1}{\sigma(\theta)} d\theta \tag{4.1}$$

induces a variance stabilizing transformation. If ϕ is well-defined, then it is also monotone, which implies that a confidence interval for $\phi(\theta)$ can be transformed to a confidence interval for θ, which is the parameter we were originally interested in. In the following example we illustrate this concept.

4.4 Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson distributed for some $\theta > 0$. By the central limit theorem we find that

$$\sqrt{n}(\overline{X_n} - \theta) \xrightarrow{d} N(0, \theta).$$

An asymptotic confidence interval of level $1-\alpha$ for the parameter θ is now given by

$$\left(\overline{X_n} - z_{\alpha/2}\sqrt{\frac{\theta}{n}}, \overline{X_n} + z_{\alpha/2}\sqrt{\frac{\theta}{n}} \right).$$

However, since the value of θ is unknown this confidence interval as well as the convergence is uninteresting. We apply a variance stabilizing transformation. Define the function $\phi(x) = 2\sqrt{x}$, for $x > 0$ this function is differentiable with derivative $\phi'(x) = \frac{1}{\sqrt{x}}$. By the Delta method we find that

$$\sqrt{n}(\phi(\overline{X_n}) - \phi(\theta)) \xrightarrow{d} \phi'(\theta)N(0,\theta) = \frac{1}{\sqrt{\theta}}N(0,\theta) = N(0,1)$$

and thus the variance of the limit distribution is now independent of θ. Notice that ϕ is defined as prescribed by (4.1). Now the confidence interval for $\phi(\theta) = 2\sqrt{\theta}$ is given by

$$\left(\overline{X_n} - z_{\alpha/2}\frac{1}{\sqrt{n}}, \overline{X_n} + z_{\alpha/2}\frac{1}{\sqrt{n}} \right).$$

Because $2\sqrt{x}$ is strictly increasing, we can use its inverse $\frac{x^2}{4}$ to obtain a confidence interval for θ that is independent of θ. We find that

$$\left(\frac{1}{4}\left(\overline{X_n} - z_{\alpha/2}\frac{1}{\sqrt{n}}\right)^2, \frac{1}{4}\left(\overline{X_n} + z_{\alpha/2}\frac{1}{\sqrt{n}}\right)^2\right)$$

is an asymptotic confidence interval for θ of level $1-\alpha$. □

4.2 Intermezzo: parametric models

In this section we will discuss parametric models. It serves to strengthen the mental framework in which subsequent sections and chapters can be embedded. Additionally, we will expose some of the limitations of the asymptotic theory we present in this thesis.

We will often assume that the distribution from which we sample is part of some kind of collection or family of distributions. We use the notion of parametric models to describe such collections.

4.5 Definition (parametric model)**.** A parametric model $\mathcal{P}$ is a collection of probability distributions on some sample space Ω. It is given by

$$\mathcal{P} := \{P_\theta \mid \theta \in \Theta\}.$$

The model $\mathcal{P}$ is indexed by Θ, which we call the parameter space. For all $\theta \in \Theta$ the corresponding P_θ is a probability distribution.

In this thesis $\mathcal{P}$ always precisely contains all the distributions from a particular family. For example, $\mathcal{P}$ may consist of all exponential distributions or of all normal distributions. However, it is possible to define $\mathcal{P}$ and Θ in such a way that $\mathcal{P}$ consists of multiple families of distributions or is a mix of different kind of distributions.

Suppose we have a model $\mathcal{P} := \{P_\theta \mid \theta \in \Theta\}$, i.e. we have some distribution P_θ in mind that depends on some unknown k-dimensional parameter θ. Given θ, our supposition implies that we exactly know with what kind of distribution P_θ we are dealing. When working with asymptotics we often work with consistent estimators for θ.

4.6 Definition (Consistent estimator)**.** Suppose we have some parametric model $\mathcal{P} := \{P_\theta \mid \theta \in \Theta\}$. An estimator T_n is a consistent estimator for some parameter θ if

$$T_n \xrightarrow{P} \theta$$

for every possible value of $\theta \in \Theta$.

Usually we will denote the true parameter value with θ_0. So, when doing asymptotic statistics, consistent estimators will give us the exact parameter value with a probability tending to one.

Another even more useful property of an estimator is asymptotic normality. This allows us to construct confidence intervals for the estimator and to conduct several statistical tests. A sequence of probability distributions is asymptotically normal whenever it converges in distribution to a normal distribution. In the next section and Section 5.3 we will extensively treat this property.

Consistency and asymptotic normality are very desirable, but we must be realistic with regard to its application. Such asymptotic properties are not helpful when our sample size n is not sufficiently large. For instance, in Chapter 5 we will see that maximum likelihood estimators, among other estimators, are asymptotically optimal. But it is possible that another estimator may work better when doing statistics on a finite data set, since they might converge faster to a parameter of interest θ. Thus when applying the concepts presented here we must always be careful to make sure that the errors due to our finite sample size are small enough.

As mentioned before, the assumption of some parametric model $\mathcal{P}$ implies that if we know the exact value of the parameter θ, then we know everything about the underlying distribution from which we sample. We must not confuse the certainty we have within a model with knowledge about the actual distribution from which we sample. So even though the asymptotic procedures we study may give us certainty with a probability tending to one, the parametric model we assumed may be wrong. In other words, the assumption of a parametric model introduces a bias which we cannot erase from consequent procedures.

Many parameter spaces Θ we study are of infinite size and the assumption of a corresponding parametric model might not seem like a big restriction. However, even though a lot of parameter spaces are infinitely large, a parametric model covers a mere fraction of all the possible distributions that exist. When the actual distribution of interest is not contained within a parametric model then no matter how large the sample size n is, we will never obtain the actual distribution of interest. We should always keep the bias introduced by parametric models in mind.

4.3 Moment estimators

This section will be devoted to moment estimators. As a consequence of the Delta method treated in Section 4.1, this relatively simple method of

estimation can often lead to useful results. The method of moments uses a set of equations for which the solution is an estimator for some parameter θ. The parameter θ does not necessarily have to be a one-dimensional vector.

We start by giving a simple example, which we will generalize into the method of moments. After that we prove a theorem that gives us conditions under which the method of moments leads to desirable results.

4.7 Example. Suppose we have $X_1, \ldots, X_n$ i.i.d. samples from a $N(\mu, \sigma^2)$ distribution. The parameter of interest is

$$\theta = (\mu, \sigma) \in \{(\mu, \sigma) \in \mathbb{R}^2 \mid \mu \in \mathbb{R}, \sigma > 0\}.$$

Let $X \sim N(\mu, \sigma^2)$, then $\mathrm{E}(X) = \mu$ and $\mathrm{E}(X^2) = \mathrm{E}(X)^2 + \mathrm{Var}(X) = \mu^2 + \sigma^2$. By the law of large numbers we expect that if n is sufficiently large, then $\overline{X_n} \approx \mu$ and $\overline{X_n^2} \approx \mu^2 + \sigma^2$. We construct a system of equations for which the solution gives us an estimation of θ:

$$\begin{aligned} \overline{X_n} &= \frac{1}{n}\sum_{i=1}^{n} X_i = \mathrm{E}(X) = \mu \\ \overline{X_n^2} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mathrm{E}(X^2) = \mu^2 + \sigma^2 \end{aligned} \tag{4.2}$$

Solving this system gives us $\mu = \overline{X_n}$ and $\sigma = \sqrt{\overline{X_n^2} - (\overline{X_n})^2}$. An estimation of θ given a sample of size n would thus be $\hat{\theta}_n = \left(\overline{X_n}, \sqrt{\overline{X_n^2} - (\overline{X_n})^2}\right)$. □

We give a general definition for what a moment estimator is.

4.8 Definition. Let $X_1, \ldots, X_n$ be a sample from a distribution that depends on a parameter $\theta \in \Theta$ and let $f_1, \ldots, f_k$ be functions. A method of moments estimate is the solution of the system of equations that we obtain by ranging the equation

$$\frac{1}{n}\sum_{i=1}^{n} f_j(X_i) = \mathrm{E}_\theta(f_j(X)) \tag{4.3}$$

over the functions $f_1, \ldots, f_k$.

Notice that in the definition above we implicitly assume that all functions f_j are defined in a way that includes all possible outcomes of the random vector X in their domain. We can state the problem in Example 4.7 in terms of Definition 4.8. If we define $f_1(x) = x$ and $f_2(x) = x^2$, then the system of

equations given in Definition 4.8 leads to the same set of equations (4.2) as in Example 4.7.

The method of moments thus consists of matching theoretical moments to sample moments. Most of the time it is sufficient to match k moments if the estimated parameter θ is a k-dimensional. In Example 4.7 this is also the case. In its simplest form the method of moments utilizes the functions $f_j = x^j$, to which the method owes its name. In this case we can reduce the system of equations presented in Definition 4.8 to

$$\frac{1}{n}\sum_{i=1}^{n} X_i^j = \overline{X_n^j} = \mathrm{E}_\theta(X^j) \text{ for } j = 1, \ldots, k. \tag{4.4}$$

4.9 Example (Gamma distribution)**.** Let $X_1, \ldots, X_n$ be i.i.d. samples from a gamma distribution with parameter $\theta = (\alpha, \beta)$. We are interested in estimating θ. We let $f_j(x) = x^j$ for $j = 1, 2$, such that the system of equations we need to solve reduces to (4.4) with $k = 2$. Suppose X has a gamma distribution with parameter $\theta = (\alpha, \beta)$. We find that $\mathrm{E}_\theta(X) = \frac{\alpha}{\beta}$ and that $\mathrm{E}_\theta(X^2) = \frac{\alpha(\alpha+1)}{\beta^2}$. Hence, we want to solve the system of equations

$$\begin{aligned} \overline{X_n} &= \frac{\alpha}{\beta} \\ \overline{X_n^2} &= \frac{\alpha(\alpha+1)}{\beta^2}. \end{aligned} \tag{4.5}$$

We can rewrite the second equation to $\overline{X_n^2} = \overline{X_n}^2 + \frac{\overline{X_n}}{\beta}$, which gives us

$$\beta = \frac{\overline{X_n}}{\overline{X_n^2} - \overline{X_n}^2}.$$

Now, using the first equation we find

$$\alpha = \beta\overline{X_n} = \frac{\overline{X_n}^2}{\overline{X_n^2} - \overline{X_n}^2}.$$

We conclude that our method of moments estimate for θ is

$$\hat{\theta}_n = \left(\frac{\overline{X_n}^2}{\overline{X_n^2} - \overline{X_n}^2}, \frac{\overline{X_n}}{\overline{X_n^2} - \overline{X_n}^2} \right).$$

□

Sometimes moment estimators are not the ideal choice, but in appropriate circumstances they have convergence rate $\sqrt{n}$ and are asymptotically normal. Theorem 4.10 will specify these circumstances using the Delta method, but in order to state and prove the theorem we have to approach the method of moments from a slightly different angle.

We define $f = (f_1, \ldots, f_k)$ and $e : \Theta \to \mathbb{R}^k : \theta \mapsto \mathrm{E}_\theta(f(x))$. Note that $\mathrm{E}_\theta(f(x))$ is a vector consisting of all the expectations $\mathrm{E}_\theta(f_j(x))$ for $j = 1, \ldots, k$. We can rewrite the system of equations for which the solution is the method of moments estimate as

$$\overline{f_n} := \frac{1}{n}\sum_{i=1}^{n} f(X_i) = e(\theta) := \mathrm{E}_\theta(f(X)). \tag{4.6}$$

The vector $\overline{f_n}$ and e should be related to each other in a specific way in order for (4.6) to yield a unique solution. To have any solution at all, the vector $\overline{f_n}$ should be in the range of $e(\theta)$. Moreover, if the function $e(\theta)$ is a bijection, then the solution of (4.6) is unique with $\hat{\theta}_n = e^{-1}(\overline{f_n})$. This implies that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \sqrt{n}(e^{-1}\left(\overline{f_n}\right) - e^{-1}(\mathrm{E}_{\theta_0}(\overline{f_n}))\big). \tag{4.7}$$

Theorem 4.10 is stated in terms of this function e and vector f.

4.10 Theorem. *Let $e(\theta) = \mathrm{E}_\theta(f(X))$ be a bijection on an open set $\Theta \subseteq \mathbb{R}^k$, furthermore, let $e(\theta)$ be continuously differentiable at θ_0 with nonsingular derivative e'_{θ_0} and let $\mathrm{E}_{\theta_0}(||f(X)||^2) < \infty$. Then moment estimators $\hat{\theta}_n$ exist with probability tending to one and*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, e'^{-1}_{\theta_0}\Sigma_{\theta_0}(e'^{-1}_{\theta_0})^T\right).$$

Proof. Suppose $e(\theta) = E_\theta(f(X))$ is a bijection on an open set $\Theta \in \mathbb{R}^k$ that is continuously differentiable at θ_0 with nonsingular derivative e'_{θ_0} and suppose $\mathrm{E}_{\theta_0}(||f(X)||^2) < \infty$. Our assumption that e is continuously differentiable implies that e is differentiable in a neighborhood of θ_0. Furthermore, the assumed continuity of e' in θ_0 combined with the nonsingularity of e'_{θ_0} implies nonsingularity in a neighborhood of θ_0. It follows from the inverse function theorem that there exists an open neighborhood U of θ_0 on which $e : U \to e(U)$ is bijective with a differentiable inverse $e^{-1} : e(U) \to U$. As a consequence, the range $e(U)$ is an open neighborhood of $e(\theta_0)$.

By the law of large numbers the sequence $\overline{f_n}$ converges in probability to $e(\theta_0)$, which implies that the probability that $\overline{f_n}$ is contained in $e(U)$ tends to one. As a consequence, $e^{-1}(\overline{f_n})$ is uniquely determined and the moment estimators $\hat{\theta}_n$ exist with probability tending to one.

We are left with proving the last part of the theorem. By the central limit theorem the sequence $\sqrt{n}(\overline{f_n} - \mathrm{E}_{\theta_0}(\overline{f_n}))$ is asymptotically normal. Observe that e^{-1} is differentiable at θ_0. Then by the Delta method we find that

$$\sqrt{n}(e^{-1}(\overline{f_n}) - e^{-1}(\mathrm{E}_{\theta_0}(\overline{f_n}))) \xrightarrow{d} N\left(0, e'^{-1}_{\theta_0}\Sigma_{\theta_0}(e'^{-1}_{\theta_0})^T\right)$$

which we can rewrite as

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d_\theta} N\left(0, e'^{-1}_{\theta_0}\Sigma_{\theta_0}(e'^{-1}_{\theta_0})^T\right).$$

by equation (4.7). Lastly, we should observe that in the above equation Σ_{θ_0} represents the covariance matrix of the vector $f(X)$ under θ_0. □

As mentioned before, the theorem above gives conditions for which moment estimators are asymptotically normal and have convergence rate $\sqrt{n}$. We take another look at Example 4.9. The function $e : \Theta \to \mathbb{R}^k : \theta \mapsto \mathrm{E}_\theta(f(x))$ is continuously differentiable with nonsingular derivative e'_{θ_0} since both the right-hand sides of the equations in (4.5) describe smooth functions. Therefore by the theorem above we find that the moment estimators $\hat{\theta}_n$ of Example 4.9 are asymptotically normal.

We conclude this chapter with another example in which Theorem 4.10 is useful.

4.11 Example. We consider the uniform distribution with parameter $\theta = (\alpha, \beta)$, that is, we consider the uniform distribution on the interval $[\alpha, \beta]$ with $\alpha < \beta$. Note that the parameter space $\{\theta \in \mathbb{R}^2 \mid \alpha < \beta\}$ is an open set. Once more, we let $f_j(x) = x^j$ for $j = 1, 2$. Then the theoretical moments of X are given by

$$\mathrm{E}_\theta(X) = \frac{\alpha + \beta}{2}$$
$$\mathrm{E}_\theta(X^2) = \frac{\alpha^2 + \alpha\beta + \beta^2}{3}$$

The right-hand side of these equations are both continuously differentiable with nonsingular derivative, hence by Theorem 4.10 the moment estimators $\hat{\theta}_n$ exist and are asymptotically normal. □

Chapter 5

M-estimators

In this chapter we study M-estimators and their asymptotic behaviour. The first section serves as an introduction to M-estimators and mainly consists of examples. In the second section we look into the consistency of M-estimators in general and derive some conditions for consistency. The subsequent section treats asymptotic normality. We conclude the chapter with a section dedicated to the maximum likelihood estimator, which is the most important type of M-estimator.

5.1 Introduction to M-estimators

In the previous chapters we have developed methods to fluently work with stochastic convergence of sequences of random vectors. We have also looked at the estimation of parameters, most notably by use of moment estimators. Throughout this chapter we assume that we have some parametric model $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ containing the distribution we are sampling from.

Like moment estimators, M-estimators use a preset method in order to find an estimate $\hat{\theta}_n$. Let $X_1, \ldots, X_n$ be an i.i.d. sample from some distribution p_θ and let χ denote the set consisting of all values the random variables X_i can take. An M-estimator maximizes a criterion function $M_n(\theta)$ that is expressed in terms of known functions $m_\theta : \chi \to \overline{\mathbb{R}}$. This criterion function M_n is given by

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i).$$

M-estimators owe their name to this maximization procedure. Often the maximum of $M_n(\theta)$ will be determined by solving the set of equations obtained by setting the partial derivatives equal to zero. In this case, these estimators are sometimes referred to as Z-estimators, where the 'Z' stands

for 'zero'. Van der Vaart is the most notable user of this name, but most of the literature simply refers to them as M-estimators, as will we.

We come to the following formal definition.

5.1 Definition (M-estimators)**.** If $\hat{\theta}_n$ is the maximum of some function

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i)$$

where $m_\theta : \chi \to \overline{\mathbb{R}}$ are known functions, or if $\hat{\theta}_n$ is the solution of

$$0 = \Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i), \tag{5.1}$$

where ψ_θ are known vector-valued functions, then we say that $\hat{\theta}_n$ is an M-estimator.

Some remarks are in order. If the functions m_θ are smooth and concave in θ and if $\psi_{\theta,i}(x) = \frac{\partial}{\partial\theta_i} m_\theta(x)$ then the two M-estimators are equivalent. As mentioned earlier, this is often the case. Furthermore, this particular situation is the biggest motivation for including M-estimators that are zeros in the definition.

Also notice that (5.1) represents a system of equations. Alternatively, given that the functions ψ_θ are k-dimensional, we can write this system as

$$0 = \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i) = \sum_{i=1}^{n} \psi_{\theta,j}(X_i) \quad \text{for } j = 1, \ldots, k. \tag{5.2}$$

We dropped the $\frac{1}{n}$ in the equation above since the expression is set equal to zero. We will sometimes refer to these equations or to (5.1) as estimating equations.

When saying 'the maximizer' in the definition above there are two problems that may arise. First, a maximizer does not always exist. Secondly, if a maximizer exists it is not necessarily unique. The first problem can be solved by simply saying that there is no estimate if there is no maximum. The second problem is not a problem at all, if we just choose one of the maxima found then this works as a proper M-estimator. So we do not have to concern ourselves with these potential problems, and in subsequent theorems they will be taken into account.

The most important type of M-estimator is the maximum likelihood estimator. The following example makes it explicit why these types of estimators are M-estimators, after which we will not concern ourselves with them until section 5.4.

5.2 Example (Maximum likelihood estimator)**.** Let $X_1, \ldots, X_n$ be samples from some distribution $P_\theta \in \mathcal{P}$ and let p_θ denote the corresponding probability density function. Then the maximum likelihood estimator maximizes the likelihood function $\prod_{i=1}^n p_\theta(X_i)$. Since the log function is strictly increasing, the maximum of the likelihood function is equivalent to the maximum of the log likelihood function

$$\sum_{i=1}^n \log(p_\theta(X_i)).$$

Hence the maximum likelihood estimator is an M-estimator with $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log(p_\theta(X_i))$ and $m_\theta(x) = \log(p_\theta(x))$. In the special case of the maximum likelihood estimator we will write ℓ_θ instead of m_θ. So $\ell_\theta(x) = \log(p_\theta(x))$. We write it this way because now our notation is the same as for the score defined in Definition 2.14. We can see this in the following way. If the density function p_θ is partially differentiable with respect to θ for each x, then we can write the maximum likelihood estimator in the form of (5.1) by setting $\psi_\theta = \ell'_\theta$, which is the vector-valued function defined as

$$\ell'_{\theta,j}(x) = \frac{\partial}{\partial \theta_j} \log(p_\theta(x)).$$

In other words, ℓ'_θ is the same as the score. Most of the time it is easier to compute $\hat{\theta}$ using the score, but sometimes the log likelihood function is not smooth and the partial derivatives in ℓ'_θ may not exist. Thus the Maximum likelihood estimator as a maximum is more fundamental then the maximum likelihood estimator as a zero. □

We might wonder what kind of reasonable M-estimators exist besides maximum likelihood estimators. Location estimators provide an example of this. Note that 'location' can be taken to mean different things like the sample mean, sample median or centre of symmetry. The following example generalizes a few of these location estimators and in subsequent sections we will return to some of these examples.

5.3 Example (Location estimators)**.** We start with considering the sample mean and the sample median. We can obtain an estimation for the sample mean by solving $\sum_{i=1}^n (X_i - \theta) = 0$. Likewise, we can obtain an estimation of the sample median by solving $\sum_{i=1}^n \operatorname{sign}(X_i - \theta) = 0$. We implicitly assume here that there are no tied observations at the median. These are of the same form as the estimating equations (5.2), hence both are M-estimators. Furthermore, both are of the form

$$\psi(x - \theta) = 0 \tag{5.3}$$

for some function ψ that is monotone and odd around zero. M-estimators of this form can be called 'location estimators' since $\hat{\theta}_n + \alpha$ would solve $\sum_{i=1}^{n} \psi(X_i + \alpha - \theta) = 0$, that is, shifting the data with an amount α would result in shifting the estimate with α as well. This property is called location equivariance, but we won't study it further in this thesis.

Another example of a location estimator in the sense described above are Huber estimators, which can be obtained by letting

$$\psi(x) = \begin{cases} -k & \text{if } x < -k \\ x & \text{if } |x| \leq k \\ k & \text{if } x > k \end{cases} \tag{5.4}$$

for some $k \in \mathbb{R}$ in (5.3). Huber estimators can be useful when having a dataset containing a few extreme points that have a significant influence on the estimate. For the median these extreme values do not really matter, therefore the sample median can be considered a robust estimator. However, they do matter for the sample mean, which we consider non-robust for that reason. A Huber estimator can be used as a robust estimator for the sample mean. For large k a Huber estimator behaves like the sample mean, while for small k it behaves like the sample median.

Quantiles are another instance of a type of location estimator that is also an M-estimator. The pth quantile of a sample is approximately equal to the point θ such that pn observations are less than θ. Naturally, this implies that approximately $((1-p)n)$ observations are greater than θ. We do not treat the quantiles any further, but merely wanted to make the reader aware of the fact that they are M-estimators as well. □

In the next two sections we will encounter more examples of M-estimators, the examples given in this section should suffice in order for us to look at properties of M-estimators in general.

5.2 Consistency

When estimating while using large datasets or when doing asymptotic statistics it is desirable that an estimator $\hat{\theta}_n$ is consistent for the parameter of interest θ. Remember that we defined the concept of a consistent estimator in Definition 4.6. In this section we will study the conditions under which M-estimators can be taken to be consistent.

We have encountered numerous consistent estimators in previous chapters. For example, by the law of large numbers the sample mean $\overline{X_n}$ is asymptotically consistent for the population mean, provided that $\mathrm{E}(|X|) < \infty$.

Similarly, we could show that the sample median is a consistent estimator for the population mean.

Remember consistency means that $d(\hat{\theta}_n, \theta_0) \xrightarrow{P} 0$. So in order to prove an estimator to be consistent, we need to have a metric d on the set Θ. Throughout this chapter we assume that we have such a metric d.

Suppose we have a criterion function $M_n(\theta)$ corresponding to an M-estimator, which is maximized by $\hat{\theta}_n$. Then the asymptotic behaviour of $\hat{\theta}_n$ is determined by the asymptotic behaviour of the criterion functions $M_n(\theta)$, since the values of $\hat{\theta}_n$ depend on the maxima of $M_n(\theta)$. We often find that $M_n(\theta)$ converges in probability to some non-random function $M(\theta)$, provided that there is some kind of appropriate normalization. Let θ_0 be the maximizer of $M(\theta)$. Now, if $M_n(\theta) \xrightarrow{P} M(\theta)$ for every θ, then we might presume that the maximizer of $M_n(\theta)$ converges to the maximizer of $M(\theta)$, that is, $\hat{\theta}_n \xrightarrow{P} \theta_0$.

However, this is not always the case. It turns out that the pointwise convergence

$$M_n(\theta) \xrightarrow{P} M(\theta) \quad \text{for all } \theta \in \Theta$$

is too weak to guarantee the convergence of $\hat{\theta}_n \to \theta_0$, since $\hat{\theta}_n$ depends on the entire function $M_n(\theta)$. The focus of this section lies on finding conditions that are strong enough to guarantee this convergence.

First we will strengthen the pointwise convergence by assuming uniform convergence of M_n to M. We will see in Theorem 5.6 and Theorem 5.7 that this uniform convergence is strong enough to guarantee the consistency of $\hat{\theta}_n$. However, we will see that it is too strong in the sense that weaker assumptions are sufficient for the consistency of $\hat{\theta}_n$ as well. Lemma 5.8 will provide an example of one of the numerous weaker alternatives.

We define two properties of M_n and M. Both of these will be used in the two subsequent theorems.

5.4 Definition (Near maximization)**.** Let $M_n : \theta \mapsto M_n(\theta)$ be a function. We say that estimators $\hat{\theta}_n$ nearly maximize M_n if

$$M_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} M_n(\theta) - o_P(1).$$

Remember, we use θ_0 to denote the true parameter value. Definition 5.4 implies that $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$ and this near maximization will turn out to be enough to guarantee convergence in probability of $\hat{\theta}_n$ to θ_0.

5.5 Definition (Well-separated maximum)**.** Let $M : \Theta \to \overline{\mathbb{R}}$ be a nonrandom map. If M attains its unique maximum at θ_0 and if only θ close to θ_0

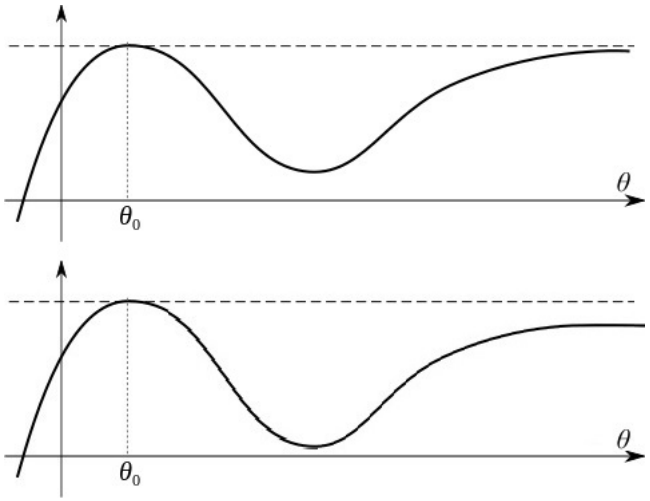

Figure 5.1: The upper graph represents a maximum that is not well-seperated: there are θ not close to θ_0 giving values $M(\theta)$ close to $M(\theta_0)$. For the lower graph this is not the case, therefore we see a well-separated maximum there. The original and unedited figure can be found on https://www.wikiwand.com/en/Extremum_estimator (accessed on 25-12-2019).

give values $M(\theta)$ close to the maximum $M(\theta_0)$, that is,

$$\sup_{\theta \in \{\theta | d(\theta, \theta_0) \geq \epsilon\}} M(\theta) < M(\theta_0),$$

then we say that θ_0 is a well-separated maximum of M.

Figure 5.1 gives an illustration of what a well-separated maximum is, and what is not.

5.6 Theorem. *Suppose M_n are random functions of θ and M is a deterministic function of θ. Suppose that for every $\epsilon > 0$*

$$\begin{aligned} &\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0; \\ &\sup_{\theta \in \{\theta | d(\theta, \theta_0) \geq \epsilon\}} M(\theta) < M(\theta_0). \end{aligned} \tag{5.5}$$

Then a sequence of estimators $\hat{\theta}_n$ that nearly maximize M_n converges in probability to θ_0.

Proof. Suppose M_n are random functions and that M is a deterministic function, and suppose that (5.5) holds for these functions. Additionally we

assume that $\hat{\theta}_n$ is a sequence of estimators that nearly maximizes M_n. By definition 5.4 we have

$$\begin{aligned} M_n(\hat{\theta}_n) &\geq \sup_{\theta \in \Theta} M_n(\theta) - o_P(1) \\ &\geq M_n(\theta_0) - o_P(1) \\ &= M(\theta_0) - o_P(1) \end{aligned}$$

The last equality in this equation follows from the first statement of (5.5). Using the inequality above and the first statement of (5.5) again, we obtain

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}_n) &\leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_P(1) \\ &\leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| + o_P(1) \xrightarrow{P} 0. \end{aligned}$$

Thus we have established that

$$M(\theta_0) - M(\hat{\theta}_n) \xrightarrow{P} 0. \tag{5.6}$$

It follows from the second statement of (5.5) that if $d(\theta, \theta_0) > \epsilon$, then $M(\theta) < M(\theta_0) - \eta$ for some $\eta > 0$. Substituting $\hat{\theta}_n$ for θ gives us

$$\begin{aligned} \mathbb{P}(d(\hat{\theta}_n, \theta_0) > \epsilon) &\leq \mathbb{P}(M(\hat{\theta}_n) < M(\theta_0) - \eta) \\ &= \mathbb{P}(M(\theta_0) - M(\hat{\theta}_n) > \eta) \to 0. \end{aligned}$$

The convergence to zero is a consequence of (5.6). We conclude that $\hat{\theta}_n$ converges in probability to θ_0. □

Thus the theorem above states that if there is uniform convergence of M_n in Θ, and if $\hat{\theta}_n$ nearly maximizes M_n, and if M has a well-separated maximum, then $\hat{\theta}_n$ is consistent.

This theorem only applies in the case where our M-estimator maximizes a criterion function. A similar theorem holds true in the case where our M-estimator $\hat{\theta}_n$ is the zero of some criterion function $\Psi_n(\theta)$. Again, there is a deterministic function to which our random functions converge in probability; i.e. $\Psi_n \xrightarrow{P} \Psi$. Analogous to the preceding theorem, we might expect that a sequence of zeros (or near-zeros) of Ψ_n converge in probability to a zero of Ψ. Under some restrictions that are much like the ones in theorem 5.6, this turns out to be the case.

5.7 Theorem. *Suppose Ψ_n are random vector-valued functions of θ and Ψ is a deterministic vector-valued function of θ such that for every $\epsilon > 0$*

$$\begin{aligned} &\sup_{\theta \in \Theta} ||\Psi_n(\theta) - \Psi(\theta)|| \xrightarrow{P} 0; \\ &\inf_{\theta \in \{\theta | d(\theta, \theta_0) \geq \epsilon\}} ||\Psi(\theta)|| > 0 = ||\Psi(\theta_0)||. \end{aligned} \tag{5.7}$$

Then a sequence of estimators $\hat{\theta}_n$ with $\Psi_n(\hat{\theta}n) = o_P(1)$ converges in probability to θ_0.

Proof. This proof is essentially an application of the previous theorem. Let Ψ_n be random vector-valued functions of θ and let Ψ be a deterministic vector-valued function of θ for which (5.7) holds. Define $M_n(\theta) = -||\Psi_n(\theta)||$ and $M(\theta) = -||\Psi(\theta)||$. Then by (5.7) and the reverse triangle inequality we find

$$\begin{aligned}\sup_{\theta\in\Theta} ||\Psi_n(\theta) - \Psi(\theta)|| &\geq \sup_{\theta\in\Theta} \big|||\Psi_n(\theta)|| - ||\Psi(\theta)||\big| \\ &= \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0\end{aligned}$$

and also

$$\begin{aligned}\inf_{\theta\in\{\theta|d(\theta,\theta_0)\geq\epsilon\}} ||\Psi(\theta)|| &= \sup_{\theta\in\{\theta|d(\theta,\theta_0)\geq\epsilon\}} -||\Psi(\theta)|| \\ &= \sup_{\theta\in\{\theta|d(\theta,\theta_0)\geq\epsilon\}} M(\theta) \\ &< 0 \\ &= ||\Psi(\theta_0)|| \\ &= -||\Psi(\theta_0)|| \\ &= M(\theta_0).\end{aligned}$$

So (5.5) is true for M_n and M. Now, suppose that $\hat{\theta}_n$ is a sequence of estimators such that $\Psi_n(\hat{\theta}_n) = o_P(1)$. Then $\hat{\theta}_n$ nearly maximizes $M_n(\hat{\theta}) = -||\Psi_n(\hat{\theta}_n)||$. Hence by theorem 5.6 we find that $\hat{\theta}_n \xrightarrow{P} \theta_0$. □

Thus, theorems 5.6 and 5.7 can help us in verifying that M-estimators are consistent and are therefore very useful. The main difficulty is to establish that the conditions under which we can apply these theorems hold.

The deterministic condition imposed upon M and Ψ is implied by uniqueness of θ_0 as a maximizer of M or a zero of Ψ, if Θ is a compact set and M or Ψ continuous. This provides a relatively easy way to verify the deterministic condition.

The stochastic condition, i.e. the uniform convergence imposed upon M_n and Ψ_n is equivalent to the set of functions $\{m_\theta \mid \theta \in \Theta\}$ or $\{\psi_\theta \mid \theta \in \Theta\}$ being Glivenko Cantelli. A possible set of sufficient conditions for these sets of functions to be Glivenko Cantelli is that Θ is compact and that the functions $m_\theta(x)$ and $\psi_\theta(x)$ are continuous and dominated by an integrable function. Further exploration of these sets of functions is beyond the scope of this thesis, but the reader should be aware of their existence.

Earlier we mentioned that the condition of uniform convergence is quite strong and that many weaker alternatives exist. The following lemma is an instance of one of these weaker alternatives.

5.8 Lemma. *Let $\Theta \subseteq \mathbb{R}$ and let Ψ_n be random functions of θ such that for every $\theta \in \Theta$ we have $\Psi_n(\theta) \xrightarrow{P} \Psi(\theta)$, where Ψ is a fixed function. If each function Ψ_n is nondecreasing and $\Psi_n(\hat{\theta}_n) = o_P(1)$, and if for every $\epsilon > 0$ the point θ_0 is such that $\Psi(\theta_0 - \epsilon) < 0 < \Psi(\theta_0 + \epsilon)$. Then $\hat{\theta}_n$ is a consistent estimator for θ_0.*

Proof. We suppose that we have Θ, Ψ_n and Ψ as in the lemma above. We divide the proof into two cases; the case where Ψ_n has a unique zero at $\hat{\theta}_n$ and the case where $\hat{\theta}_n$ is near a zero.

(i) We assume that Ψ_n has a unique zero at $\hat{\theta}_n$, then for all $\theta_1 < \theta_2$ with $\Psi_n(\theta_1) < 0 < \Psi_n(\theta_2)$, the zero must be between θ_1 and θ_2, hence $\theta_1 < \hat{\theta}_n < \theta_2$. So, for all $\epsilon > 0$ we have

$$\{\Psi_n(\theta_0 - \epsilon) < 0, \Psi_n(\theta_0 + \epsilon) > 0\} \subseteq \{\theta_0 - \epsilon < \hat{\theta}_n < \theta_0 + \epsilon\}. \tag{5.8}$$

It follows from $\Psi_n(\theta_0 - \epsilon) \xrightarrow{P} \Psi(\theta_0 - \epsilon)$ that $\mathbb{P}(\Psi_n(\theta_0 - \epsilon) < 0) \to 1$. Similarly, $\Psi_n(\theta_0 + \epsilon) \xrightarrow{P} \Psi(\theta_0 + \epsilon)$ implies $\mathbb{P}(\Psi_n(\theta_0 + \epsilon) > 0) \to 1$. Therefore the probability of the event on the left in (5.8) converges to one. As a consequence, the probability of the event on the right has to converge to one as well. We conclude that $\hat{\theta}_n \xrightarrow{P} \theta_0$, so $\hat{\theta}_n$ is consistent for θ_0.

(ii) Suppose $\hat{\theta}_n$ is a near zero. We adjust the proof of (i) to include this case. For every $\epsilon, \eta > 0$ we have

$$\begin{aligned} &\{\Psi_n(\theta_0 - \epsilon) < -\eta, \Psi_n(\theta_0 + \epsilon) > \eta\} \\ \subseteq &\{\theta_0 - \epsilon < \hat{\theta}_n < \theta_0 + \epsilon\} \cup \{\Psi_n(\hat{\theta}_n) \notin [-\eta, \eta]\} \end{aligned} \tag{5.9}$$

From $\Psi_n(\theta) \xrightarrow{P} \Psi(\theta)$ for all θ it follows that $\mathbb{P}(\Psi_n(\theta_0 - \epsilon) < -\eta) \to 1$ and $\mathbb{P}(\Psi_n(\theta_0 + \epsilon) > \eta) \to 1$. Now, if we take $\eta > 0$ sufficiently small, then the left-hand side of (5.9) converges to one. From the assumption that $\Psi_n(\hat{\theta}_n) = o_P(1)$ it follows that $\mathbb{P}(\Psi_n(\hat{\theta}_n) \notin [-\eta, \eta])$ converges to zero. Consequently, $\mathbb{P}(\theta_0 - \epsilon < \hat{\theta}_n < \theta_0 + \epsilon)$ converges to one and we conclude that $\hat{\theta}_n$ is a consistent estimator. □

Notice that the lemma above can also be applied if all functions Ψ_n are non-increasing, since we can just take the negative of Ψ_n and Ψ and all other assumptions would still hold. The following example illustrates how the theorems and lemma presented in this section can be applied.

5.9 Example (Sample median)**.** We consider the sample median from example 5.3 again. We expect that estimator $\hat{\theta}_n$ is consistent. The estimator $\hat{\theta}_n$ is a zero or near a zero of the function

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \text{sign}(X_i - \theta). \tag{5.10}$$

For every $\theta \in \Theta$ we find that

$$\Psi(\theta) = \text{E}(\text{sign}(X - \theta)) = \mathbb{P}(X > \theta) - \mathbb{P}(X < \theta)$$

using the law of large numbers. Since we are setting Ψ to zero, we expect that $\hat{\theta}_n$ converges in probability to a point θ_0 such that $\mathbb{P}(X > \theta_0) = \mathbb{P}(X < \theta_0)$, i.e. if $\hat{\theta}_n$ is consistent it converges in probability to the population median.

Consistency of $\hat{\theta}_n$ would follow from an application of theorem 5.7, but verifying that Ψ_n uniformly converges to Ψ is difficult and therefore we prefer to apply lemma 5.8. We just saw that the law of large numbers implies $\Psi_n(\hat{\theta}_n) = o_P(1)$. We can see that the function Ψ_n in 5.10 is non-increasing. If the population median θ_0 is unique, which we assumed in 5.3, then for all $\epsilon > 0$ we have that

$$\mathbb{P}(X < \theta_0 - \epsilon) < \frac{1}{2} < \mathbb{P}(X < \theta_0 + \epsilon).$$

From this it follows that $\Psi(\theta_0 - \epsilon) > 0 > \Psi(\theta_0 + \epsilon)$. By lemma 5.8 we conclude that $\hat{\theta}_n \xrightarrow{P} \theta_0$. □

5.3 Asymptotic normality

In the previous section we investigated under which conditions a sequence of estimators $\hat{\theta}_n$ is consistent. Given consistency of a sequence of estimators, the next question we ought to ask ourselves is at what rate the difference between $\hat{\theta}_n$ and θ_0 converges to zero. Quite often this rate is $\frac{1}{\sqrt{n}}$ and as a result the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ neither diverges nor collapses to zero. Moreover, we are also interested in whether the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges to a normal distribution or not, i.e. whether the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal. In this section we will attempt to answer both of these questions with regard to M-estimators.

There are many different sets of conditions we do not mention in this section which also yield asymptotic normality. The purpose of this section is to give the reader an impression of what some of these sets of conditions

look like and how they can be used to establish asymptotic normality. Thus contrary to our treatment in previous chapters and sections we will not be able to prove all theorems we posit, nor will we prove every claim we make.

Instead of an arbitrary metric we will only consider the Euclidean metric d in this chapter. Furthermore, we let Θ be an open subset of a Euclidean space throughout this section.

In definition 5.1 we characterised M-estimators in two different ways: as maximum of a function M_n and as zero of some function Ψ_n. In our treatment of asymptotic normality of M-estimators we will mostly work with the latter characterization. This choice might weaken our results in some instances, because the characterization as a maximum is sometimes more fundamental than the characterization as a zero, as we saw in example 5.2. However, the consequences of this choice are minimal and should not be a problem.

Throughout the remainder of this chapter we will suppose that we have an i.i.d. sample from some distribution p_θ.

Remember that definition 5.1 stated that

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i)$$

and that by the law of large numbers $\Psi(\theta) = \mathrm{E}(\psi_\theta(X_i))$, such that the zeros $\hat{\theta}_n$ of Ψ_n converge to the zero θ_0 of Ψ. $\hat{\theta}_n$ First, we only consider the situation where the parameter of interest θ is one-dimensional. We obtain the following theorem.

5.10 Theorem. *Suppose that the function $\psi_\theta(x)$ is twice continuously differentiable in some neighborhood B of θ_0. For every $\theta \in B$ and every fixed x, suppose the derivatives $\psi'_\theta(x)$ and $\psi''_\theta(x)$ are such that $|\psi''_\theta(x)| \leq \psi''(x)$ for a function ψ'' with $\mathrm{E}(\psi''(X_i)) < \infty$. Also suppose that $\mathrm{E}(\psi^2_{\theta_0}(X_i))$, $E(|\psi'_{\theta_0}(X_i)|) < \infty$ and $\mathrm{E}(\psi'_{\theta_0}(X_i)) \neq 0$. If $\hat{\theta}_n$ are zeros of Ψ_n that are consistent for a zero θ_0 of Ψ, then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{\mathrm{E}(\psi^2_{\theta_0}(X_i))}{\mathrm{E}(\psi'_{\theta_0}(X_i))^2}\right).$$

In order to prove this theorem we need the following lemma, which we will prove before giving a proof of theorem 5.10.

5.11 Lemma. *Under the same assumptions as in theorem 5.10, and if $\hat{\theta}_n$ are zeros of Ψ_n that are consistent for a zero θ_0 of Ψ, then the following statements are true:*
(i) $\sqrt{n}\Psi_n(\theta_0) \xrightarrow{d} N\left(0, \mathrm{E}(\psi^2_{\theta_0}(X_i))\right)$;

(ii) $\Psi_n'(\theta_0) \xrightarrow{P} \mathrm{E}(\psi_{\theta_0}'(X_i))$*;*
(iii) $\Psi_n''(\tilde{\theta}_n) = O_p(1)$*, where* $\tilde{\theta}_n$ *denotes a point between* $\hat{\theta}_n$ *and* θ_0 *for each* $n \in \mathbb{N}$*.*

Proof. We make the same assumptions as in theorem 5.10 and prove each statement.

(i) Because

$$\sqrt{n}\Psi_n(\theta_0) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{\theta_0}(X_i)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_0}(X_i)$$

and $\mathrm{E}(\psi_{\theta_0}(X_i)) = 0$, application of the central limit theorem yields $\sqrt{n}(\Psi_n(\theta_0)) \xrightarrow{d} N(0, \mathrm{E}(\psi_{\theta_0}^2(X_i)))$.

(ii) We can write $\Psi_n'(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\psi_{\theta_0}'(X_i)$, so $\Psi_n'(\theta_0)$ is essentially an average of all $\psi_{\theta_0}'(X_i)$. This enables us to apply the law of large numbers, by which we obtain our desired result $\Psi_n'(\theta_0) \xrightarrow{P} \mathrm{E}(\psi_{\theta_0}'(X_i))$.

(iii) As in (ii), $\Psi_n''(\tilde{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n}\psi_{\tilde{\theta}_n}''(X_i)$ is an average of all $\psi_{\tilde{\theta}_n}''(X_i)$. Our result would immediately follow from the law of large numbers, but the terms $\psi_{\tilde{\theta}_n}''(X_i)$ are dependent through $\tilde{\theta}_n = \tilde{\theta}_n(X_1, \ldots, X_n)$, therefore we cannot simply apply the law of large numbers.

We define the event $A_n = \{\tilde{\theta}_n \in B\}$ for all $n \in \mathbb{N}$. By assumption A_n happens with probability tending to one. By the triangle inequality and our assumptions we have

$$|\Psi_n''(\tilde{\theta}_n)| \leq \frac{1}{n}\sum_{i=1}^{n}|\psi_{\tilde{\theta}_n}''(X_i)| \leq \frac{1}{n}\sum_{i=1}^{n}\psi''(X_i).$$

Now, we can apply the law of large numbers to $\frac{1}{n}\sum_{i=1}^{n}\psi''(X_i)$, since there is no dependency between the different terms. This yields $\frac{1}{n}\sum_{i=1}^{n}\psi''(X_i) \xrightarrow{P} \mathrm{E}(\psi''(X_i)) < \infty$. By theorem 3.15 (ii) and Prokhorov's theorem the sequence $\sum_{i=1}^{n}\psi''(X_i)$ is uniformly tight. For $M \in \mathbb{R}$ we find that

$$\mathbb{P}\left(|\Psi_n''(\tilde{\theta}_n) > M\right) \leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\psi''(X_i) > M\right) + \mathbb{P}(A_n^C)$$

By assumption $\mathbb{P}(A_n^C)$ converges to zero and by uniform tightness we have an $M \in \mathbb{R}$ for every $\epsilon > 0$ such that $\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}\psi''(X_i) > M) < \epsilon$. Hence $\mathbb{P}(|\Psi_n''(\tilde{\theta}_n)| > M) < \epsilon$ as well, and thus $\Psi_n''(\tilde{\theta}_n) = O_P(1)$. □

Proof (of theorem 5.10)**.** We assume the antecedents of the theorem. Additionally assume that $\hat{\theta}_n$ are zeros of Ψ_n that are consistent for a zero θ_0 of Ψ. This means that $\hat{\theta}_n \to \theta_0$ and therefore it is reasonable to construct a Taylor expansion of $\Psi_n(\hat{\theta}_n)$ around θ_0. Let $\tilde{\theta}_n$ be a point between $\hat{\theta}_n$ and θ_0, then

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + (\hat{\theta}_n - \theta_0)\Psi_n'(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2\Psi_n''(\tilde{\theta}_n).$$

We rewrite this equation in order to get an expression for $\sqrt{n}(\hat{\theta}_n - \theta_0)$, first we write

$$-\Psi_n(\theta_0) = (\hat{\theta}_n - \theta_0)\left(\Psi_n'(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\Psi_n''(\hat{\theta}_n)\right)$$

which leads to

$$(\hat{\theta}_n - \theta_0) = \frac{-\Psi_n(\theta_0)}{\Psi_n'(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\Psi_n''(\hat{\theta}_n)}.$$

Now, multiplying by $\sqrt{n}$ gives us

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\sqrt{n}\Psi_n(\theta_0)}{\Psi_n'(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\Psi_n''(\hat{\theta}_n)}. \tag{5.11}$$

By lemma 5.11 we have $\Psi_n''(\tilde{\theta}_n) = O_p(1)$ and by assumption we have $\hat{\theta}_n \xrightarrow{P} \theta_0$. Combining these gives us

$$\frac{1}{2}(\hat{\theta}_n - \theta_0)\Psi_n''(\tilde{\theta}_n) = o_P(1)O_P(1) = o_P(1)$$

by lemma 3.20 (iv).

We consider the denominator of (5.11), the left term converges in probability to $\mathrm{E}(\psi_{\theta_0}'(X_i))$ and the right term converges to zero in probability. By Slutsky's lemma (i) we conclude that the denominator converges to $\mathrm{E}(\psi_{\theta_0}'(X_i))$ in distribution. The numerator of (5.11) converges in distribution to

$$N(0, \mathrm{E}(\psi_{\theta_0}^2(X_i))).$$

Note that $E(|\psi_{\theta_0}'(X_i)|) < \infty$ and $\mathrm{E}(\psi_{\theta_0}'(X_i)) \neq 0$ by assumption. Now, application of (iii) from Slutsky's lemma yields the desired result

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{\mathrm{E}(\psi_{\theta_0}^2(X_i))}{\mathrm{E}(\psi_{\theta_0}'(X_i))^2}\right).$$

□

Theorem 5.10 can easily be generalized to higher-dimensional θ. Suppose that the parameter of interest θ is k-dimensional. Instead of a single estimating equation, we would have k estimating equations. Remember that the estimating equations are given by (5.1), which can be found in definition 5.1. Or equivalently by (5.2), which can be found right underneath definition 5.1. Then $\Psi_n : \mathbb{R}^k \to \mathbb{R}^k$ and the sequence of derivatives $\Psi_n'(\theta_0)$ consists of square matrices of dimension k that converge to the matrix $\mathrm{E}(\psi_{\theta_0}'(X_i))$, which we assume to be invertible. The entries of this matrix are given by

$$\mathrm{E}(\psi_{\theta_0}'(X_i))_{i,j} = \mathrm{E}\left(\frac{\partial}{\partial\theta_j}\psi_{\theta_0,i}(X_i)\right).$$

With the exception of replacing ordinary multiplication with matrix multiplication and division multiplication with taking the inverse the proof would remain the same. So Theorem 5.10 is also true for higher-dimensional θ. Thus in the k-dimensional case we would obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_k\left(0, \left(\mathrm{E}\left(\psi_{\theta_0}'(X_i)\right)\right)^{-1} \mathrm{E}\left(\psi_{\theta_0}(X_i)\psi_{\theta_0}^T(X_i)\right) \left(\mathrm{E}\left(\psi_{\theta_0}'^T(X_i)\right)\right)^{-1}\right)$$

when applying Theorem 5.10.

If we are able to verify that all the conditions are met, Theorem 5.10 gives us a desirable result. However, not all the assumptions in the theorem are easy to verify or very common properties. The strongest condition required by the theorem is that uniformly for all θ in a neighborhood B of θ_0, the function $\psi_\theta''(x)$ can be bounded from above by some integrable function ψ''. Firstly, we should observe that the double prime in ψ'' is only used to make clear that this function is connected to the second derivative $\psi_\theta''(x)$, thus it does not signify that it is a second derivative of some function ψ. Secondly, we take a look at how we could find such an integrable function ψ''. The most obvious choice is $\psi'' = \sup_{\theta\in B}|\psi_{\theta_0}''|$. Since for this function we have

$$\sup_{\theta\in B}\mathrm{E}(|\psi_{\theta_0}''(X_i)|) \leq \mathrm{E}\left(\sup_{\theta\in B}|\psi_\theta''(X_i)|\right), \tag{5.12}$$

and thus gives us a desirable result. However, we still require $\mathrm{E}\left(\psi''(X_i)\right) < \infty$. Quite often it is hard to determine the exact value of the expression on the right-hand side in (5.12). In these cases we are better off by bounding it with some simpler expression to ensure that $\mathrm{E}\left(\psi''(X_i)\right) < \infty$.

5.12 Example (Cauchy likelihood)**.** Let us have some i.i.d. sample from a Cauchy distribution with location $\theta \in \Theta$. We consider the log likelihood function

$$\theta \mapsto M_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log(1 + (x_i - \theta)^2).$$

The maximum likelihood estimator $\hat{\theta}_n$ will then be a zero of the function $\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \psi_\theta(X_i)$, where

$$\psi_\theta(x) = \frac{x-\theta}{1+(x-\theta)^2}.$$

Alternatively, we could say that $\psi_\theta(x) = \psi(x-\theta)$ for

$$\psi(x) = \frac{x}{1+x^2}$$

The first two derivatives of ψ are

$$\psi'(x) = \frac{1-x^2}{(1+x^2)^2} \quad \text{and} \quad \psi''(x) = \frac{8x^3}{(1+x^2)^3} - \frac{6x}{(1+x^2)^2},$$

which are both continuous and have limit zero at $\pm\infty$. This implies that there is a constant $L > 0$ such that

$$|\psi(x)|, |\psi'(x)|, |\psi''(x)| \leq L \quad \text{for all } x \in \mathbb{R},$$

hence these function are uniformly bounded. This means that

$$\mathrm{E}\left(\psi''(X_i)\right), \mathrm{E}\left(\psi_{\theta_0}^2(X_i)\right), \mathrm{E}\left(\psi'_{\theta_0}(X_i)\right) < \infty.$$

Now, if additionally $\mathrm{E}\left(\psi'_{\theta_0}(X_i)\right) \neq 0$ then all conditions in Theorem 5.10 are satisfied. If this is the case, then for any $\hat{\theta}_n$ that is consistent for a zero θ_0 of $\Psi(\theta)$ the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal. □

Another condition of Theorem 5.10 that is often problematic is that $\psi_\theta(x)$ is required to be twice continuously differentiable. An instance where this leads to problems is the function $\psi_\theta(x) = \operatorname{sign}(x-\theta)$ from Example 5.3 and Example 5.9, which gives the sample median. This function is clearly not twice continuously differentiable and thus theorem 5.10 cannot be applied. However, the M-estimator corresponding to the function $\psi_\theta(x)$ is asymptotically normal. For many M-estimators this is the case: Theorem 5.10 is not applicable but they are asymptotically normal. Thus we are forced to develop other methods of establishing asymptotic normality. The following theorem is an alternative of Theorem 5.10 and assumes less than one derivative.

5.13 Theorem. *Let $\theta \in \Theta$ and let B be a neighborhood of θ_0. Let $\psi_\theta(x)$ be a vector-valued function such that for all $\theta_1, \theta_2 \in B$ and for some function ψ' with $\mathrm{E}\,(\psi'^2(X_i)) < \infty$ we have*

$$||\psi_{\theta_1}(x) - \psi_{\theta_2}(x)|| \leq \psi'(x)||\theta_1 - \theta_2||. \tag{5.13}$$

Furthermore, let $\mathrm{E}\left(||\psi_{\theta_0}(X_i)||^2\right) < \infty$ *and let the map* $\theta \mapsto \mathrm{E}\left(\psi_\theta(X_i)\right)$ *be differentiable at a zero* θ_0 *with nonsingular derivative matrix* V_{θ_0}.

If

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\hat{\theta}_n}(X_i) = o_P\left(\frac{1}{\sqrt{n}}\right)$$

and $\hat{\theta}_n \xrightarrow{P} \theta_0$, *then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_0}(X_i) + o_P(1).$$

Moreover, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *is asymptotically normal with mean zero and covariance matrix*

$$V_{\theta_0}^{-1}\mathrm{E}\left(\psi_{\theta_0}(X_i)(\psi_{\theta_0}(X_i))^T\right)(V_{\theta_0}^{-1})^T.$$

Proof. For a proof, see [6], Theorem 5.21, p.52-53. □

We illustrate the use of the theorem above by giving an example.

5.14 Example (Huber estimator)**.** We consider the Huber function (5.4) from Example 5.3. It is clear that this function $\psi(x)$ is differentiable for all x except at $x = \pm k$, since the left and right derivatives differ from each other in these points. Therefore the function $\theta \mapsto \psi_\theta(x) = \psi(x - \theta)$ with ψ the Huber function is differentiable at all θ except $\theta = x \pm k$. Notice that the derivative of this function is either equal to one or equal to zero. It follows that

$$|\psi(x - \theta_1) - \psi(x - \theta_2)| \leq |\theta_1 - \theta_2|$$

for every pair $\theta_1, \theta_2 \in \Theta$. We conclude that ψ is Lipschitz with Lipschitz constant $\psi'(x) = 1$.

We assume that the corresponding probability measure $\mathbb{P}$ on $\mathbb{R}$ has a density function p that is differentiable with respect to θ with nonsingular derive matrix $V_\theta = \int \psi(x)p'(x + \theta)dx$. We observe that for the density p corresponding to the probability measure $\mathbb{P}$ we have

$$\mathrm{E}\left(\psi_\theta(X_i)\right) = \int \psi(x - \theta)p(x)dx = \int \psi(x)p(x + \theta)dx.$$

Now all the conditions of Theorem 5.13 are satisfied, hence we conclude that the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix

$$V_{\theta_0}^{-1}\mathrm{E}\left(\psi_{\theta_0}(X_i)(\psi_{\theta_0}(X_i))^T\right)(V_{\theta_0}^{-1})^T.$$

□

Theorem 5.13 is better applicable than Theorem 5.10 while still being relatively simple. However, the asymptotic normality of the sample median still cannot be proven, since the Lipschitz condition in (5.13) does not hold for the function $\theta \mapsto \text{sign}(x-\theta)$. So the conditions in Theorem 5.13 are still stronger than necessary. Motivated by finding a suitable method of proving the asymptotic normality of the sample median we will state another theorem. The following theorem is based on the characterization of M-estimators as maximizers of M_n. It is very similar to the previous theorem, but it includes the sample median.

5.15 Theorem. *Let B be a neighborhood of θ_0 and let $m_\theta(x)$ be a function for all $\theta \in \Theta$ such that $\theta \mapsto m_\theta(x)$ is differentiable at θ_0 for almost every x, with derivative $m'_{\theta_0}(x)$. Furthermore, let $m'(x)$ be a measurable function with $\mathrm{E}\,(m'^2(X_i)) < \infty$ such that for all $\theta_1, \theta_2 \in B$*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq m'(x)||\theta_1 - \theta_2||.$$

Moreover, suppose that the map $\theta \mapsto \mathrm{E}\,(m_\theta(X_i))$ admits a second-order Taylor expansion at a point of maximum θ_0 with nonsingular symmetric second derivative matrix V_{θ_0}. If

$$\frac{1}{n}\sum_{i=1}^{n} m_{\hat{\theta}_n}(X_i) \geq \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i) - o_P\left(\frac{1}{n}\right)$$

and $\hat{\theta}_n \xrightarrow{P} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m'_{\theta_0}(X_i) + o_P(1).$$

Moreover, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix

$$V_{\theta_0}^{-1}\mathrm{E}\left(m'_{\theta_0}(X_i)(m_{\theta_0}(X_i))^T\right)V_{\theta_0}^{-1}.$$

Proof. For a proof, see [6], Theorem 5.23, p.54. □

Equipped with the theorem above we could prove that the sample median is asymptotically normal. However, proving asymptotic normality with the theorems above is rather tedious as we have seen in Example 5.12 and Example 5.14. Furthermore, the proof of the asymptotic normality of the sample median is very much alike these examples in spirit. Therefore we omit its proof.

5.4 Maximum Likelihood Estimators

Most of the subjects we treated thus far have been leading us towards the maximum likelihood estimator. In some sense, this section is the theoretical pinnacle of this thesis. Yet we do not really introduce any complex new matter here, we merely narrow the very general theorems from the previous sections down to the maximum likelihood estimator. The reason for this is that the maximum likelihood estimator is very important and therefore we ought to treat it comprehensively. This estimator is of grave importance due to its frequent application: it is popular in both parametric statistics as well as non-parametric statistics. We will look at both the consistency and the asymptotic normality of the maximum likelihood estimator.

Let $X_1, \ldots, X_n$ be a random sample from some density p_θ. Earlier, in example 5.2 we stated that the criterion function of a maximum likelihood estimator is equal to

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log(p_\theta(X_i)), \tag{5.14}$$

it will turn out to be convenient to subtract the constant $\sum_{i=1}^{n} \log(p_{\theta_0}(X_i))$ from this criterion function, which yields

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}\right). \tag{5.15}$$

We should observe that subtraction of a constant does not affect the maximum of a function, hence the maxima of (5.14) and (5.15) are equivalent. We let $\log(0) = -\infty$, this ensures that (5.15) is well-defined provided p_{θ_0} is the true density. The asymptotic function to which $M_n(\theta)$ converges in probability is

$$M(\theta) = \mathrm{E}_{\theta_0}\left(\log\left(\frac{p_\theta(X)}{p_{\theta_0}(X)}\right)\right)$$

by the law of large numbers. The number $-M(\theta)$ is called the Kullback-Leibler divergence of p_θ and p_{θ_0}. The minus is introduced because the Kullback-Leibler was first defined as

$$\mathrm{E}_{\theta_0}\left(\log\left(\frac{p_{\theta_0}(X)}{p_\theta(X)}\right)\right),$$

interchanging the numerator and the denominator requires the introduction of the minus sign in order to keep both expectations equal. We can consider the Kullback-Leibler divergence as a measure of how different p_θ is from p_{θ_0}.

We can also consider it as a measure of distance between two distributions, but this only works intuitively since the Kullback-Leibler divergence does not have all the properties of a metric.

Remember that in the section about consistency we established that the maximum of $M_n(\theta)$ converges to the maximum of $M(\theta)$ under suitable conditions. So, we expect that the maximum of $M_n(\theta)$ converges to the maximum of $M(\theta)$ for maximum likelihood estimators as well. If we consider the Kullback-Leibler divergence $-M(\theta)$ as a distance between two distributions, then the maximum of $M(\theta)$ minimizes the distance between p_{θ_0} and p_θ. The upcoming lemma will prove this and will also show that this maximum is unique if the true density p_{θ_0} is identifiable.

5.16 Definition (Identifiability)**.** A true density p_{θ_0} is identifiable if $p_\theta \neq p_{\theta_0}$ for all $\theta \neq \theta_0$.

In other words, the true density is identifiable within a parametric model if there is exactly one distribution in that model having a density equal to p_{θ_0}, namely the distribution corresponding to θ_0 itself. Identifiability of the parameter is necessary: if the true parameter is not identifiable then a consistent estimator cannot exist for it. We arrive at the following lemma.

5.17 Lemma. *Let $\mathcal{P} = \{p_\theta \mid \theta \in \Theta\}$ be a parametric model in which the true parameter θ_0 is identifiable.*

$$M(\theta) = \mathrm{E}_{\theta_0}\left(\log\left(\frac{p_\theta}{p_{\theta_0}}(X)\right)\right)$$

has a unique maximum at θ_0.

Proof. We observe that

$$M(\theta_0) = \mathrm{E}_{\theta_0}\left(\log\left(\frac{p_{\theta_0}}{p_{\theta_0}}(X)\right)\right) = \mathrm{E}_{\theta_0}\left(\log(1)\right) = 0.$$

Therefore θ_0 is the unique maximum whenever $M(\theta) < 0$ for all $\theta \in \Theta$ with $\theta \neq \theta_0$.

We know that for all $x \geq 0$ the inequality $\log(x) \leq x-1$ holds. This means that $\log(\sqrt{x}) \leq \sqrt{x}-1$ as well, from which we deduce that $\log(x) \leq 2(\sqrt{x}-1)$. Thus we also know that

$$\mathrm{E}_{\theta_0}\left(\log\left(\frac{p_\theta}{p_{\theta_0}}(X)\right)\right) \leq 2\mathrm{E}_{\theta_0}\left(\sqrt{\frac{p_\theta}{p_{\theta_0}}(X)} - 1\right). \tag{5.16}$$

For the term on the right-hand side of the inequality above we find

$$\begin{aligned}
2\mathrm{E}_{\theta_0}\left(\sqrt{\frac{p_\theta}{p_{\theta_0}}(X)}-1\right) &= 2\mathrm{E}_{\theta_0}\left(\frac{\sqrt{p_\theta}-\sqrt{p_{\theta_0}}}{\sqrt{p_{\theta_0}}}\right) \\
&= 2\left(\int p_{\theta_0}(x)\frac{\sqrt{p_\theta(x)}-\sqrt{p_{\theta_0}(x)}}{\sqrt{p_{\theta_0}(x)}}dx\right) \\
&= 2\left(\int \sqrt{p_\theta(x)p_{\theta_0}(x)}-\sqrt{p_{\theta_0}(x)}^2 dx\right) \\
&= 2\left(\int \sqrt{p_\theta(x)p_{\theta_0}(x)}dx-1\right) \\
&= -\left(2-\int 2\sqrt{p_\theta(x)p_{\theta_0}}dx\right) \\
&= -\int \sqrt{p_\theta(x)}^2+\sqrt{p_{\theta_0}(x)}^2-2\sqrt{p_\theta(x)p_{\theta_0}(x)}dx \\
&= -\int\left(\sqrt{p_\theta(x)}-\sqrt{p_{\theta_0}(x)}\right)^2 dx.
\end{aligned} \tag{5.17}$$

Observe that the integral at the bottom of the display above is strictly negative for $\theta \neq \theta_0$, because by assumption $p_\theta \neq p_{\theta_0}$ for all $\theta \neq \theta_0$. Combining (5.16) and (5.16) gives us

$$\mathrm{E}_{\theta_0}\left(\log\left(\frac{p_\theta}{p_{\theta_0}}(X)\right)\right) \leq -\int\left(\sqrt{p_\theta(x)}-\sqrt{p_{\theta_0}(x)}\right)^2 dx < 0 \quad \text{if } \theta \neq \theta_0.$$

We conclude that θ_0 is the unique maximum. □

Furthermore, we conclude that under the same regularity conditions as in section 5.2 the maximum likelihood estimator is consistent for any identifiable θ_0.

5.18 Example (Misspecified model)**.** This example is an extension of Section 4.2. Suppose we have a sample $X_1, \ldots, X_n$ and suppose we have a parametric model $\mathcal{P} = \{p_\theta | \theta \in \Theta\}$. We would like to estimate θ to find the underlying distribution from which we sampled. However, suppose that this underlying distribution p_{θ_0} is not included in the parametric model, i.e. there is no $\theta \in \Theta$ such that p_θ is the underlying distribution. We are interested in what would happen if we estimate θ by maximizing the log likelihood $\sum_{i=1}^n \log(p_\theta(X_i))$.

We might expect that $\hat{\theta}_n$ is unpredictable and behaves in an erratic way. However, the opposite is true. Assuming that $\hat{\theta}_n$ satisfies the conditions to be consistent, we expect that $\hat{\theta}_n$ converges to a value $\theta_0 \in \Theta$, which is the

maximum of the function $\theta \to \mathrm{E}(\log(p_\theta(X_i)))$. Note that the expectation is taken under the true underlying distribution. We find that θ_0 maximizes the Kullback-Leibler divergence

$$-M(\theta) = -\mathrm{E}_{\theta_0}\left(\log\left(\frac{p_\theta(X)}{p_{\theta_0}(X)}\right)\right)$$

that we discussed earlier in this section. We also mentioned that we could consider the Kullback-Leibler divergence as a measure of distance in some sense. Hence we can intuitively think about p_{θ_0} as a projection of the true underlying distribution onto our parametric model $\mathcal{P}$. However, we should always remember that the Kullback-Leibler divergence is not a metric and for that reason p_{θ_0} is not really a projection.

Under appropriate circumstances the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal as well.

We might ask what the practical value is of the estimator $\hat{\theta}_n$ when the underlying distribution is not included in the parametric model we assume. The answer to this question differs depending on the situation. If the parametric model comes close to the truth, $\hat{\theta}_n$ may be a valuable estimator. Conversely, if, for instance, we let $\mathcal{P}$ be the family of uniform distributions and our true underlying distribution is an exponential distribution, then $\hat{\theta}_n$ may not be of any help. Besides the "distance" between p_{θ_0} and the true underlying distribution p one should also consider the context in which one is estimating. □

Next, we shall take a look at the asymptotic normality of the maximum likelihood estimator. We start of with an informal treatment and subsequently we will formally state a theorem. However, as in the previous section, we lack the means to give a formal proof of the theorem we present, therefore we omit the proof.

We assume suitable regularity of our criterion function. As we saw earlier in Example 5.2, the maximum likelihood estimator solves the set of equations

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log\left(p_\theta(X_i)\right) = 0$$

Remember that in Example 5.2 we set $\psi_{\theta,j}$ equal to $\ell'_{\theta,j}(x) = \frac{\partial}{\partial\theta_j}\log(p_\theta(x))$. The results of the theorems in the previous section makes us expect that $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal with mean zero and covariance matrix

$$\left(\mathrm{E}(\ell''_{\theta_0}(X_i))\right)^{-1}\mathrm{E}\left(\ell'_{\theta_0}(X_i)\ell'^{T}_{\theta_0}(X_i)\right)\left(\mathrm{E}(\ell''^{T}_{\theta_0}(X_i))\right)^{-1}. \tag{5.18}$$

When comparing the covariance matrix above with the higher-dimensional version of Theorem 5.10 we see that the display above has second derivatives where the result of Theorem 5.10 has first derivatives. Furthermore, (5.18) has first derivatives where our earlier result just contained the function. This is a consequence of the equality $\psi_{\theta,j}(x) = \ell'_{\theta,j}(x)$, which we defined in example 5.2.

If we make a few additional assumptions it turns out that the covariance matrix (5.18) reduces to the Fisher information matrix

$$I_{\theta_0} = \mathrm{E}\left(\ell'_{\theta_0}(X_i)\ell'^T_{\theta_0}(X_i)\right).$$

We explicate the reason for this in the one-dimensional case. We differentiate $\int p_\theta(x)d\mu(x) = 1$ twice with respect to θ. Suppose that the order of integration and differentiation can be reversed, we then find $\int p'_\theta(x)d\mu(x) = \int p''_\theta(x)d\mu(x) = 0$. Combining this with

$$\ell'_\theta(x) = \frac{p'_\theta(x)}{p_\theta(x)} \quad \text{and} \quad \ell''_\theta(x) = \frac{p''_\theta(x)}{p_\theta(x)} - \left(\frac{p'_\theta(x)}{p_\theta(x)}\right)^2$$

gives us $\mathrm{E}\left(\ell'_\theta(X_i)\right) = 0$ and $\mathrm{E}\left(\ell''_\theta(X_i)\right) = -I_\theta$. This means that the curvature of the likelihood is equal to minus the Fisher information I_θ. Substituting these identities into (5.18) reduces the covariance matrix to I_θ^{-1}.

In the higher-dimensional case this result follows in a similar way, where we should interpret $\mathrm{E}\left(\ell'_\theta(X_i)\right)$ and $\mathrm{E}\left(\ell''_\theta(X_i)\right)$ as a vector and a matrix respectively.

So, maximum likelihood estimators satisfy

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I_{\theta_0}}\right) \tag{5.19}$$

under suitable regularity conditions. In light of the Cramér-Rao inequality (Theorem 2.16) this result is very important. It implies that for large n the estimator $\hat{\theta}_n$ has distribution

$$N\left(\theta_0, \frac{1}{nI_{\theta_0}}\right).$$

In other words, $\hat{\theta}_n$ is asymptotically unbiased and has asymptotic variance $(nI_{\theta_0})^{-1}$. The Cramér-Rao theorem states that the variance of an unbiased estimator is at least $(nI_{\theta_0})^{-1}$, which is the Cramér-Rao bound. Hence (5.19) seems to imply that the maximum likelihood estimator is asymptotically the most efficient unbiased estimator we can obtain. However, we must be

sceptical towards this conclusion. The reasoning above is not mathematically rigorous and the conclusion therefore cannot be simply accepted. We lack a proper asymptotic version of the Cramér-Rao theorem and hence the Cramér-Rao bound is useless when considering asymptotic normality. Yet the conclusion that maximum likelihood estimators are asymptotically efficient is true, we simply lack the means to prove this formally. The reasoning above merely serves to give the reader a basic notion of why the maximum likelihood estimator is asymptotically efficient.

At the end of this section we formally state a theorem containing a set of conditions for asymptotic efficiency. We will see that the conditions for asymptotic efficiency presented in the reasoning above can even be relaxed in some respects. For instance, the restriction of two derivatives for p_θ can be relaxed to weaker regularity conditions.

Another reason why we should be cautious when claiming asymptotic efficiency for maximum likelihood estimators is that for some frequently occurring distributions, like the uniform distribution, this is not the case. The example below illustrates this fact.

5.19 Example. Suppose we have a uniform distribution on $[0, \theta]$ from which we obtain a sample $X_1, \ldots, X_n$. The maximum likelihood estimator is given by $\hat{\theta}_n = \max\{X_1, \ldots, X_n\}$. The variance of the sequence of estimators $\hat{\theta}_n$ is of order $O(n^{-2})$, hence the norming rate is not $\sqrt{n}$, but n. The distribution of $\hat{\theta}_n$ is given by

$$F_{\hat{\theta}_n}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \left(\frac{x}{\theta}\right)^n & \text{if } 0 < x < \theta \\ 1 & \text{if } x \geq \theta \end{cases}$$

For all $x < 0$ we find

$$\begin{aligned} \mathbb{P}_{\theta_0}\left(-n(\hat{\theta}_n - \theta_0) \leq x\right) &= \mathbb{P}_{\theta_0}\left(\hat{\theta}_n \geq -\frac{x}{n} + \theta_0\right) \\ &= 1 - \mathbb{P}_{\theta_0}\left(\hat{\theta}_n \leq \theta_0 - \frac{x}{n}\right) \\ &= 1 - (1 - \frac{x/\theta_0}{n})^n \to 1 - e^{-\frac{1}{\theta_0}x}, \end{aligned}$$

which means that $-n(\hat{\theta}_n - \theta_0)$ converges in distribution to an exponential distribution with parameter θ_0. As a consequence, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges to zero in probability. Hence the conclusions we informally inferred are not true for the uniform distribution with parameter $[0, \theta]$.

We should observe that a lot of the operations used in the informal inference are not possible on the uniform distribution, so this conclusion should not come as a surprise. □

The following theorem concludes this chapter and makes the informal treatment of the asymptotic normality of the maximum likelihood estimator rigorous. The theorem follows from theorem 5.15, or from theorem 5.13 by setting $\ell'_\theta = \psi_\theta$ as we did earlier. However, the following theorem only applies to maximum likelihood estimators, and not to M-estimators in general. As we mentioned earlier, regularity conditions can be less restrictive: we do not require a second derivative. Nevertheless, the following theorem somehow guarantees that $\mathrm{E}\left(\ell''_\theta(X_i)\right) = -I_\theta$ without assuming second derivatives.

5.20 Theorem. *Let $\mathcal{P} = \{p_\theta \mid \theta \in \Theta\}$ be a parametric model with $\Theta \in \mathbb{R}^k$ open. Let $B \subseteq \Theta$ be a neighborhood of θ_0. Suppose $p_\theta(x)$ is a probability density function such that $\theta \mapsto \log(p_\theta(x))$ is continuously differentiable for every x and such that for all $\theta_1, \theta_2 \in B$ we have the inequality*

$$|\log(p_{\theta_1}(x)) - \log(p_{\theta_2}(x))| \leq \ell'(x)||\theta_1 - \theta_2||,$$

where ℓ' is a function satisfying $\mathrm{E}_{\theta_0}\left(\ell'^2(X_i)\right) < \infty$. Additionally, assume that the fisher information matrix $I_\theta = \mathrm{E}_\theta\left(\ell'_\theta(X_i)\ell'^T_\theta(X_i)\right)$ is nonsingular and continuous on Θ. Then for the maximum likelihood estimator $\hat{\theta}_n$ that is consistent the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $I_{\theta_0}^{-1}$.

Proof. For a proof, see [6], Theorem 5.39, p.39. □

Chapter 6

Simulations

In this chapter we study how well the results presented in this thesis can be applied in actual statistical research. We will do some simulations and compare the outcomes to what we would expect in light of the theoretical results we obtained in previous chapters. We will do four different simulations. Firstly, we will simulate a moment estimator and study how fast it converges to the actual parameter. In the second simulation we will consider a maximum likelihood estimator $\hat{\theta}_n$ and investigate for what sample size the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ starts tending to a normal distribution. The third section focusses on the asymptotic normality of the sample median. In the final section we will focus on a parametric model that does not contain the distribution from which we sample and look at what happens when we start estimating its parameter. The codes used for our simulations can be found in Appendix B.

6.1 Moment estimator of Gamma distribution

We consider the moment estimator for the Gamma distribution with parameter $\theta = (\alpha, \beta)$. In example 4.4 we found that the moment estimator of θ is given by

$$\hat{\theta}_n = \left(\frac{\overline{X_n}^2}{\overline{X_n^2} - \overline{X_n}^2}, \frac{\overline{X_n}}{\overline{X_n^2} - \overline{X_n}^2} \right).$$

This estimator does not meet the conditions of theorem 4.10, hence we do not expect our estimator to be asymptotically normal. However, we are not interested in the normality of our estimator for large samples, but in how accurate the estimate is for different sample sizes.

We let $\theta = (\alpha, \beta) = (7, 2)$ and simulate the parameter estimates for α and β for different sample sizes n. The figures below contain the results.

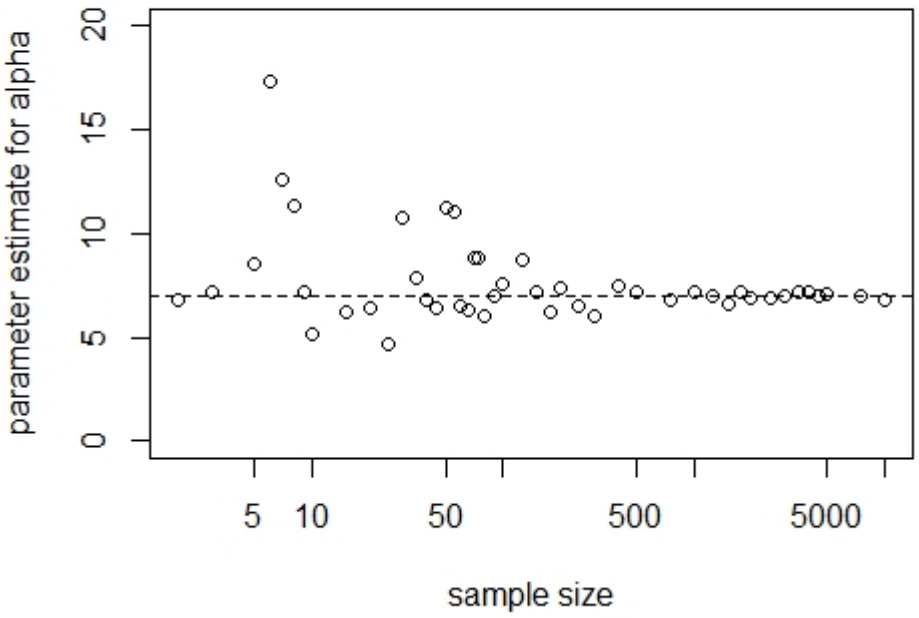

Figure 6.1: Estimates for α for different sample sizes.

We should note that the range of the vertical axis in the plot of the estimate of α is much greater than in the plot of the estimator of β. Hence the values of the estimates for α are much more scattered, therefore the estimate of β seems to be more accurate than the estimate of α. However, for sufficiently large n, say $n > 1000$, the estimates of α and β are both very close to the actual value. A sample of this size is often easy to obtain given the modern techniques at our disposal for collecting data.

Even though the theoretical framework in which we develop our methods supposes a sample size that tends to infinity, we can see that the developed methods can already be useful for relatively small sample sizes of $n = 1000$.

We should be very cautious when interpreting the figures above. In this particular case $n > 1000$ might be a sufficiently large sample size, but this is not necessarily true for moment estimators from every other possible distribution. Moreover, it might not even be the case for other Gamma distributions.

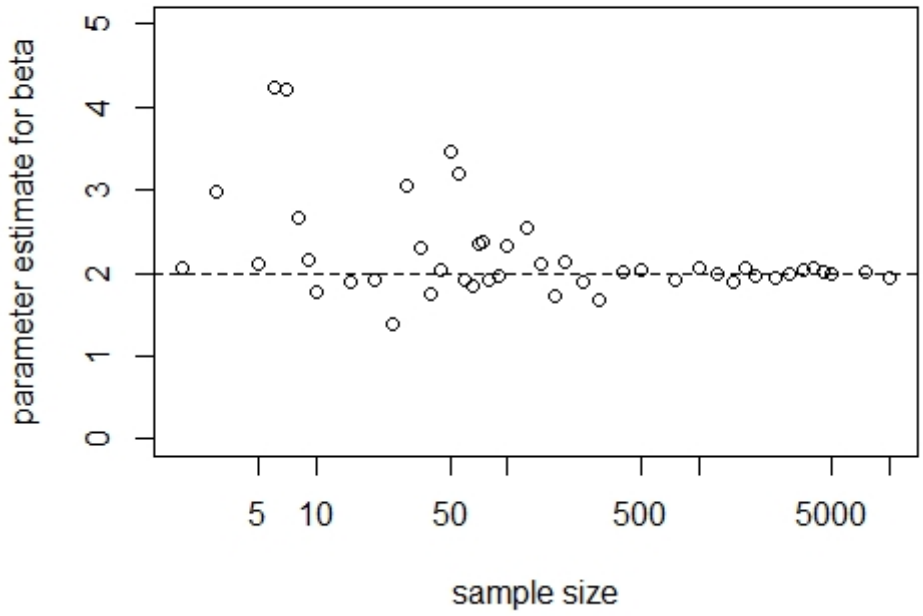

Figure 6.2: Estimates for β for different sample sizes.

Maybe some possible parameters θ cause bigger fluctuations in the estimates, demanding a bigger sample size to obtain a reliable estimate. We investigate this by doing another similar simulation, but now for $\theta = (\alpha, \beta) = (3, 5)$. The results are shown in Figure 6.3 and Figure 6.4.

We can see that the rate of convergence of the estimator for α in Figure 6.3 is very similar to the estimator in Figure 6.1 if we take the different scales of the vertical axis into account. For the second simulation a sample size of $n > 1000$ seems to be large enough to ensure a reliable estimate of α. However, for the estimate of β in Figure 6.4 we see that for some $n > 1000$ the estimate is not as close to the actual value as for the other three estimates. This is not surprising since the estimates for β in Figure 6.4 are also more scattered than all the other estimates. An explanation for this is that the shape of the density of the Gamma(3,5) distribution is very different from that of the Gamma(7,2) distribution, as can be seen in Figure 6.5. As a result, a sample from the Gamma(3,5) probably has a greater variance than a sample from the Gamma(7,2) distribution.

Nonetheless, for $n > 1000$ we still have an estimate for β that would be

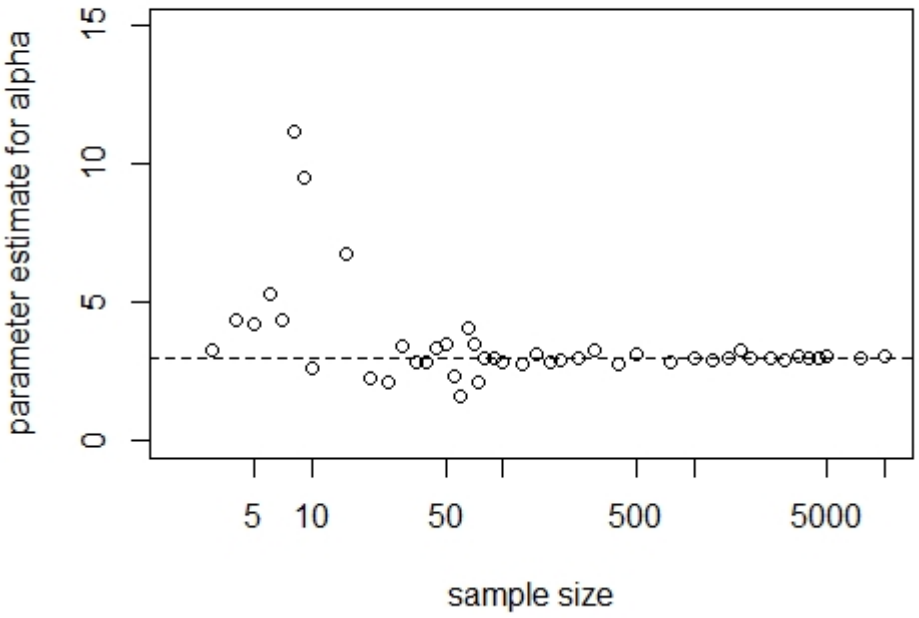

Figure 6.3: Estimates for α for different sample sizes.

sufficiently accurate for most applications. Furthermore, for larger n such as $n > 5000$ the estimate seems to be as reliable as all the other estimates. So consistency still holds, but we must conclude that the parameter values influence the number of samples needed before obtaining a reliable estimate. We should always keep this in mind. Because contrary to the discussion above, when doing actual statistical research we do not know the actual distribution from which we are sampling. Hence we cannot determine beforehand how large a sample size should be. So the example given here merely serves to demonstrate that it might be possible that a sample size of $n = 1000$ is sufficiently large to reap the benefits of asymptotic statistics. In practice, we should always remain cautious when assuming that n is sufficiently large.

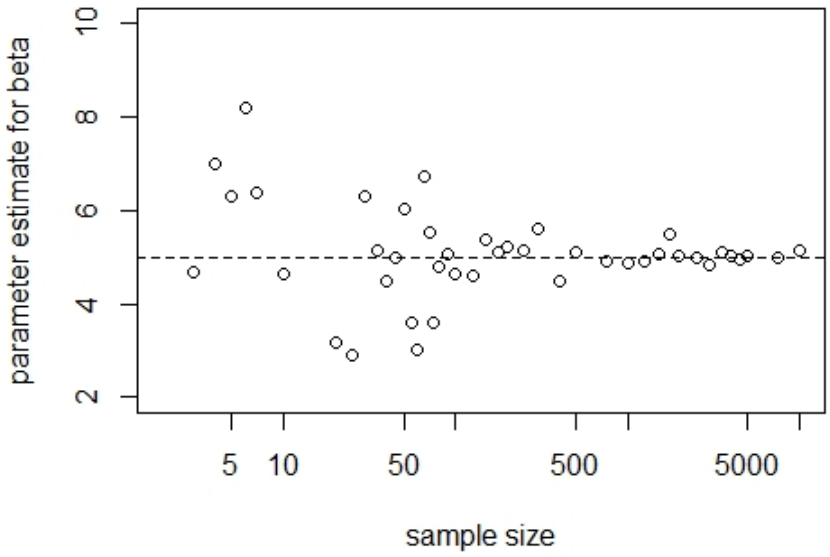

Figure 6.4: Estimates for β for different sample sizes.

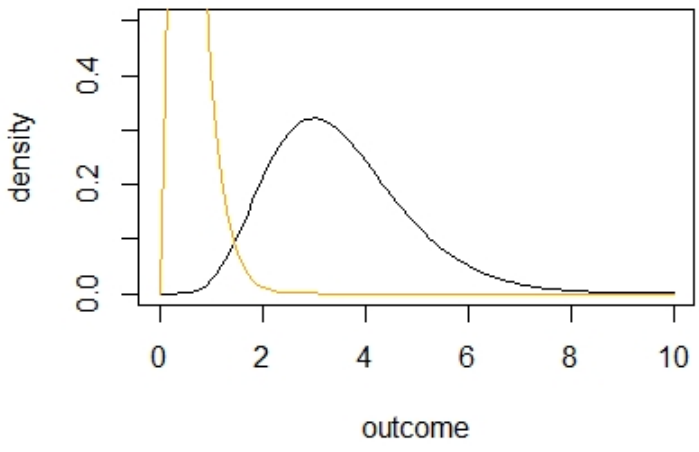

Figure 6.5: Comparison of the densities of the Gamma(7,2) distribution and the Gamma(3,5) distribution.

6.2 Asymptotic normality of the maximum likelihood estimator

In this section we will consider an exponential distribution with parameter θ and the maximum likelihood estimator $\hat{\theta}_n$. Given the results of Section 5.4 we expect that for large n the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ tends to a normal distribution

$$N\left(0, \frac{1}{I_{\theta_0}}\right).$$

Note that we expect this because the density function of the exponential distribution is smooth and complies with the regularity conditions described earlier in Section 5.4. Remember that I_{θ_0} denotes the Fisher information of the random variable having an exponential distribution with θ_0 as its parameter.

Given an exponential distribution with parameter θ and a sample of size n, the likelihood function is given by

$$\begin{aligned} L(\theta \mid X_1, \ldots, X_n) &= \prod_{i=1}^{n} \theta e^{-\theta X_i} \\ &= \theta^n e^{-\theta \sum_{i=1}^{n} X_i}. \end{aligned}$$

Then the log-likelihood equals

$$l(\theta \mid X_1, \ldots, X_n) = nlog(\theta) - \theta \sum_{i=1}^{n} X_i.$$

Differentiating the log-likelihood and setting it equal to zero gives us

$$0 = \frac{n}{\theta} - \sum_{i=1}^{n} X_i,$$

from which the maximum likelihood estimator easily follows:

$$\hat{\theta}_n = \frac{n}{\sum_{i=1}^{n} X_i}.$$

Next, we compute the Fisher information of an exponentially distributed random variable. The score of the random variable is equal to

$$\begin{aligned} \frac{d}{d\theta} \log\left(\theta e^{-\theta x}\right) &= \frac{d}{d\theta} \left(\log(\theta) - \theta x\right) \\ &= \frac{1}{\theta} - x \end{aligned}$$

Since the exponential distribution is sufficiently regular, we know that

$$\begin{aligned} I(\theta) &= \mathrm{E}_\theta \left((\frac{1}{\theta} - x)^2 \right) \\ &= \mathrm{E} \left(\frac{(1 - x\theta)^2}{\theta^2} \right) \\ &= \frac{1}{\theta^2}. \end{aligned}$$

Thus we have found the Fisher information.

Now, suppose we set $\theta_0 = 2$, so that we are sampling from an exponential distribution with rate 2. Then

$$I_{\theta_0} = \frac{1}{2^2} = \frac{1}{4},$$

from which it follows that

$$\frac{1}{I_{\theta_0}} = 4.$$

So, using the results that we computed above, we expect that

$$\sqrt{n} \left(\frac{n}{\sum_{i=1}^n X_i} - 2 \right) \xrightarrow{d} N(0, 4). \tag{6.1}$$

We simulate the distribution on the left-hand side of (6.1) for different sample sizes n. Figure 6.3 displays the results. The histogram represents the simulated distribution for each sample size, while the red line represents the normal distribution $N(0, 4)$. As we can see, for $n = 5$ the histogram does not resemble the normal distribution $N(0, 4)$ at all: it is asymmetric and the shape of the histogram does not correspond to that of the normal distribution. However, a sample size of 5 is almost always too low to do any sensible statistical analysis. We have merely included $n = 5$ to illustrate our point. Looking at the next plot we see that for $n = 20$ the histogram resembles the normal distribution a lot more, but still it is asymmetric and its values do not correspond very well to the red line. For $n = 50$, which still is a relatively low sample size, we see that the histogram starts resembling a normal distribution, even though it still is a bit asymmetric and the left side does not correspond to the red line very well. Lastly, for a much larger sample size such as $n = 10000$, we see that the histogram has become almost completely symmetric and has the same shape as the normal distribution $N(0, 4)$. This confirms our expectation that (6.1) holds.

A few comments should be made. The histogram of $n = 10000$ is not very different from the histogram of $n = 50$, while the histogram of $n = 50$ differs

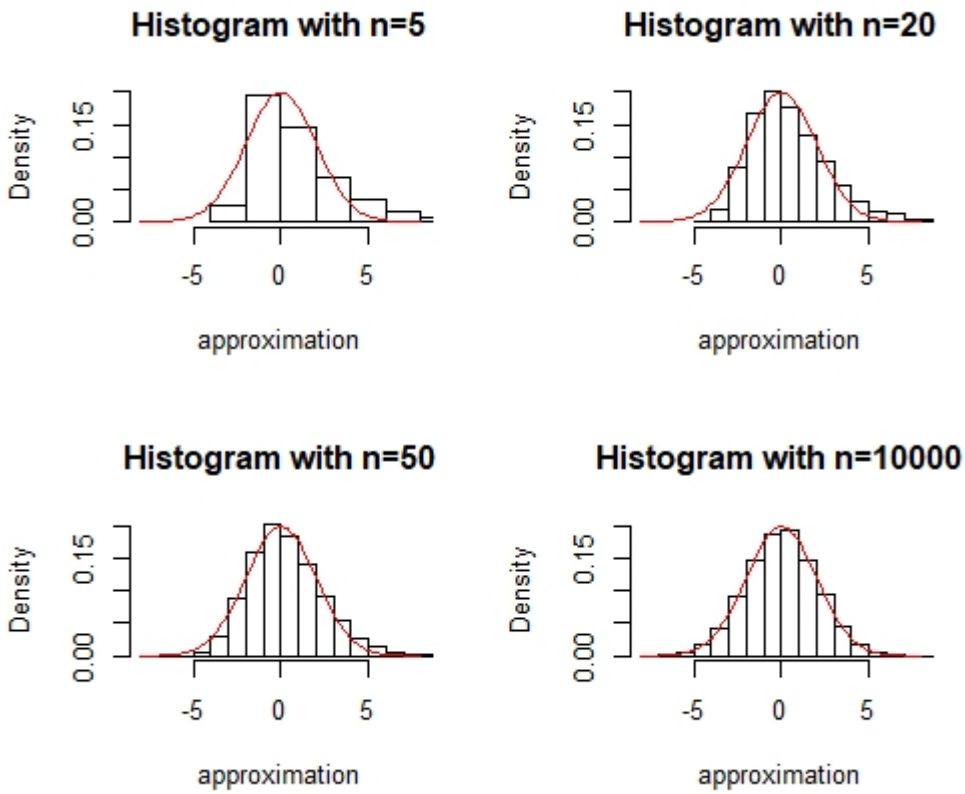

Figure 6.6: Comparison between the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ and $N(0, I_{\theta_0})$ for different sample sizes n.

a lot from those of $n = 5$ and $n = 20$. So we see that at first the distribution tends to a normal distribution relatively fast, but a much larger sample size is needed before it closely resembles one. Nonetheless, if the context does not require a lot of accuracy we can use relatively small sample sizes, say $n > 50$, to obtain estimates that are close to the actual parameter value with a high probability. Such estimates may seem like they are not very desirable, but sometimes the context in which we use statistical procedures does not allow large sample sizes. Furthermore, in a lot of situations the maximum likelihood estimator is an optimal estimator, especially when regularity conditions apply.

We conclude that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I_{\theta_0}}\right) \tag{6.2}$$

holds when sampling from the exponential distribution. Observe that the display above is the same as (5.19) in Section 5.4. Lastly, since (6.2) holds we know that the maximum likelihood estimator is also consistent in this example.

6.3 Asymptotic normality of the sample median

At the end of Section 5.3 we claimed that as a result of Theorem 5.15 the sample median is asymptotically normal, but we did not give a proof. In this section we will attempt to demonstrate this fact using a simulation.

We will demonstrate, using an example, that under the regularity conditions of Theorem 5.15 the sample median is asymptotically normal with asymptotic variance

$$\frac{1}{(2f(\theta_0))^2}, \tag{6.3}$$

where f denotes the density of the distribution under consideration and θ_0 is the median of this distribution.

Suppose that we have a Beta-distribution with parameter $\theta = (\alpha, \beta) = (2, 5)$ from which we sample. Note that this distribution meets the conditions of Theorem 5.15. Let Y_n denote the sample median from a sample of size n. The median of this distribution is approximately equal to

$$\theta_0 \approx \frac{\alpha - \frac{1}{3}}{\alpha + \beta - \frac{2}{3}} = \frac{\frac{5}{3}}{\frac{19}{3}} = \frac{5}{19} \approx 0.263158.$$

We will first look at the estimate of the population median given different sample sizes. Figure 6.7 shows our results. We can see that the sample median seems to be consistent: it converges to the population median at a rate at least as good as the estimates we saw for the moment estimators in Section 6.1. This is not surprising, since we have already proven that the sample median is consistentin Example 5.9.

We continue by investigating the distribution of $\sqrt{n}(Y_n - \theta_0)$ for different n. Our expectation is that this distribution tends to a normal distribution with its variance equal to (6.3). Given that $\theta_0 \approx \frac{5}{19}$, we find that

$$f(\theta_0) = \frac{\frac{5}{19}(1 - \frac{5}{19})^4}{B(2, 5)} = \frac{\frac{5}{19}(\frac{14}{19})^4}{\frac{1}{30}} \approx 2.3272,$$

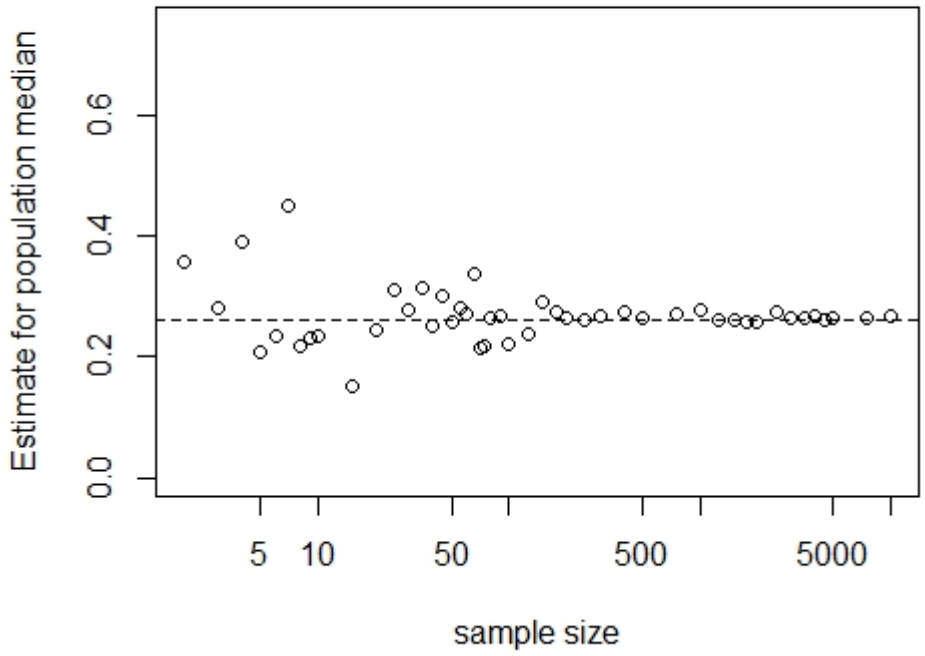

Figure 6.7: Estimates for the population median for different sample sizes.

where B denotes the Beta function. Therefore, the asymptotic variance we expect to find is equal to

$$\frac{1}{(2 \cdot 2.3272)^2} \approx 0.04616.$$

A simulation similar to the one in the previous section leads to the results shown in Figure 6.8.

We can see that for $n = 5$ the sample median already vaguely resembles a normal distribution. At $n = 25$ this resemblance is even clearer. Taking larger sample sizes such as $n = 100$ does not affect the distribution of $\sqrt{n}(Y_n - \theta_0)$ very much: the histogram in the middle and the one on the bottom are very similar. For $n = 25$ and $n = 100$ the corresponding histograms tend to the normal distribution indicated by the red line. We conclude that the sample median is indeed asymptotically normal.

Any reader that is interested in a rigorous proof of the asymptotic normality of the sample median is referred to [6], Example 5.24, p.54-55.

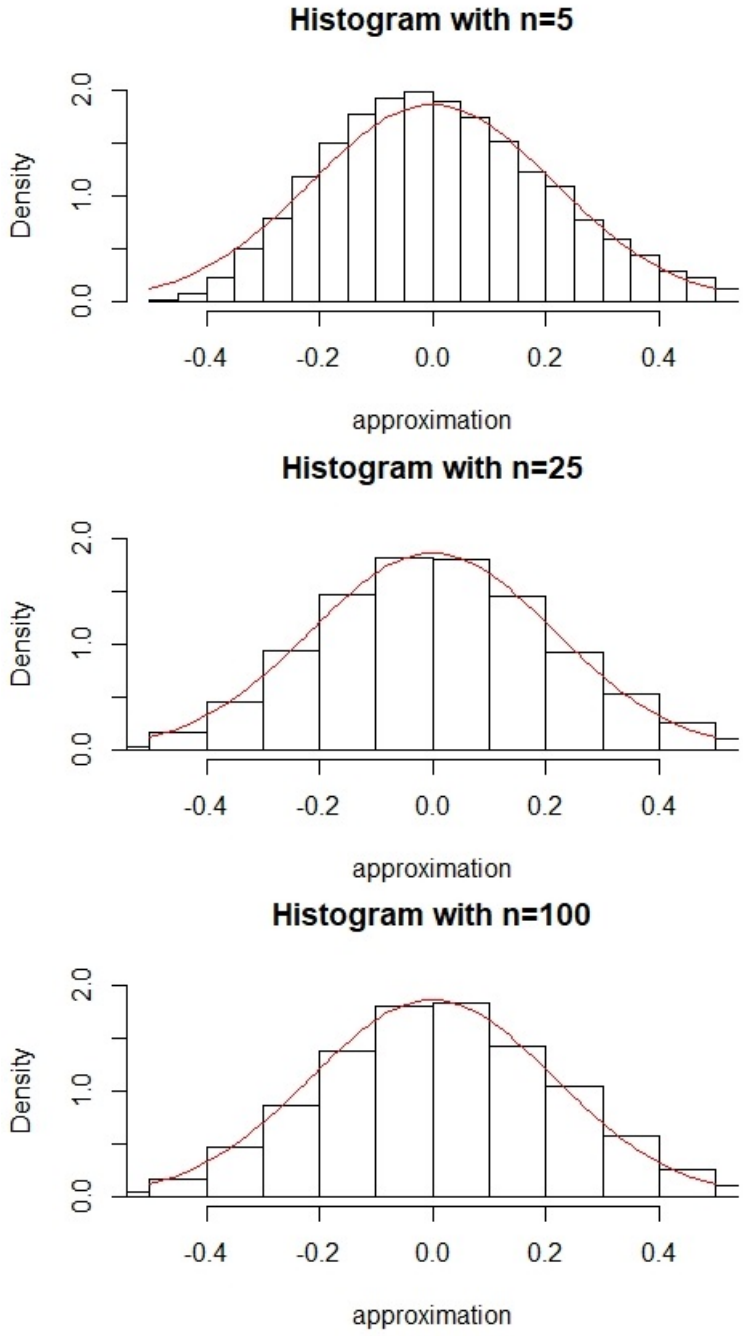

Figure 6.8: Comparisons between the distribution of $\sqrt{n}(Y_n - \theta_0)$ and the Normal distribution with mean 0 and variance equal to 0.04616.

6.4 Misspecified model

In this last simulation we take a look at what happens when we have a misspecified model, i.e. when the actual distribution from which we sample is not contained in our parametric model. Suppose that our parametric model consists of all exponential distributions and that we are sampling from a uniform distribution working on the interval $[0, 2]$. We are interested in finding the actual distribution of the population from which we sample. As we saw in the previous section, the maximum likelihood estimator is suitable for such an estimation. But remember, our estimation gives us a parameter θ for which the corresponding distribution p_θ is an exponential distribution, so with this parametric model it is impossible to find the actual distribution. Figure 6.4 gives us the results of our maximum likelihood estimate for different sample sizes n.

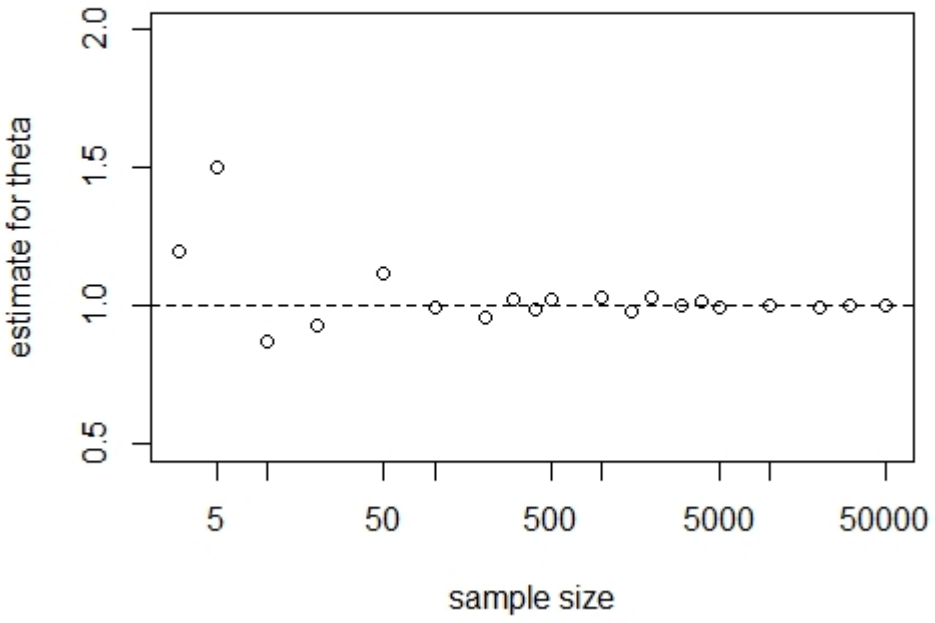

Figure 6.9: Estimates for θ for different sample sizes.

One might expect that such an estimator would behave erratically, but the opposite is true: the maximum likelihood estimator seems to converge to 1 as n gets large. Moreover, it seems to converge at a rate very similar to the

convergence of the moment estimators we saw in Section 6.1. This illustrates the danger of being overconfident in our estimation if we have a large sample size: our estimated distribution p_θ looks nothing like the actual distribution. Figure 6.10 compares our estimated distribution and the actual distribution.

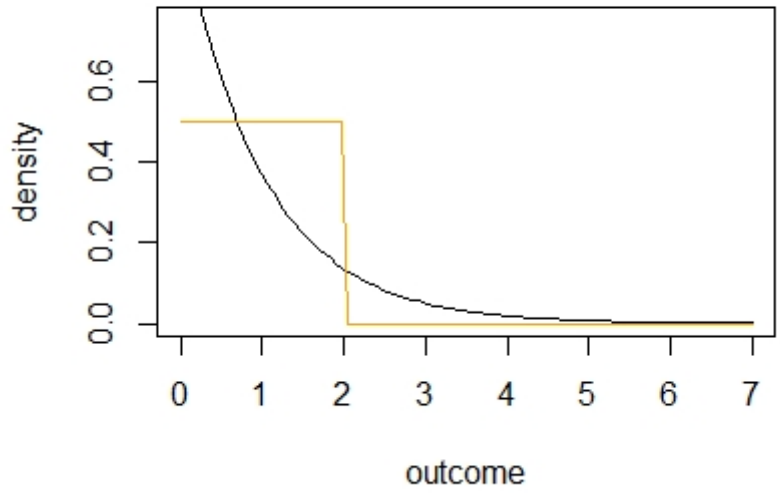

Figure 6.10: Comparison of the uniform and exponential distributions under consideration.

Figure 6.10 speaks for itself, the represented distributions do not look like each other at all. The attentive reader might argue that this example is ridiculous: if every sample is smaller than 2, then we can be almost certain that we are not dealing with an exponential distribution with rate 1. This is true, but the example we give here is merely to illustrate what happens when our actual distribution is not contained in our parametric model. What we should learn from this is that even though our asymptotic tool kit functions properly, our results do not necessarily correspond to reality. Thus we should always exercise caution when accepting a parametric model as a basis for statistical analysis.

Chapter 7

Conclusion

In Chapter 2 and Chapter 3 we developed the necessary knowledge and methods to start investigating moment estimators and M-estimators. We distinguished between three modes of convergence: convergence in distribution, convergence in probability and convergence almost surely. This distinction helped us to better understand important statistical results like the central limit theorem and the laws of large numbers, allowing us to work more fluently with these theorems. Moreover, these different concepts of stochastic convergence were fundamental in the definitions of important concepts like consistency and asymptotic normality.

Chapter 4 focussed on the Delta method. We looked at some applications of the Delta method including finding the limit distribution of the sample variance and variance stabilizing transformations. The intermezzo on parametric models did not contain any mathematical results in the form of a theorem or lemma, but nonetheless, this short section turned out to be of vital importance. Parametric models are essential when estimating parameters using moment estimators or M-estimators. In fact, in subsequent chapters we saw that the practical value of an estimator largely depends on the choice of the parametric model within the estimation is performed.

The last section of Chapter 4 focussed on moment estimators. We gave a general definition and also established conditions for which moment estimators are asymptotically normal in Theorem 4.10. In several examples we saw how moment estimators could be used. Additionally, in the first section of Chapter 6 we demonstrated what an actual estimation using a moment estimator would look like.

The last purely theoretical chapter was about M-estimators and formed the theoretical peak of this thesis. We examined a few existing M-estimators other than maximum likelihood estimators: the sample mean, the sample median, Huber estimators and quantiles. Using the concepts of well-seperated

maxima and near maximization we proved in Theorem 5.6 and Theorem 5.7 that in many cases M-estimators are consistent. In the subsequent section we investigated under what conditions an M-estimator is asymptotically normal. We found that it is hard to find a satisfying set of conditions that implies asymptotic normality because either the conditions are too strong to allow estimators like the sample mean and sample median or the conditions are not very practical. We presented three sets of conditions that imply asymptotic normality, these are given in Theorem 5.10, Theorem 5.13 and Theorem 5.15. The conditions for asymptotic normality stated in these theorems all revolve around regularity of the functions $\psi_\theta(x)$ and $m_\theta(x)$.

The last section of Chapter 5 treated the most important kind of M-estimator: the maximum likelihood estimator. In particular, besides the results that hold for all M-estimators we have proven that under sufficient regularity conditions the maximum likelihood estimator is an optimal estimator when the sample size n tends to infinity.

Lastly, Chapter 6 served to demonstrate that the theoretical results could also be applied in cases where we have a large finite sample size. We showed that the moment estimator of the Gamma distribution presented in Example 4.9 gives an accurate estimate for large sample sizes. Moreover, we saw that the maximum likelihood estimator is indeed asymptotically normal and optimal. The third section focussed on the asymptotic normality of the sample median and we concluded that this M-estimator is asymptotically normal as well.

The last simulation showed that our asymptotic results do not always lead to desirable results. Specifically, in the preceding sections we learned that our theorems seem to work well within a parametric model, but that this does not guarantee good results because our actual distribution may not be contained in our parametric model.

To conclude, when having a large sample size asymptotic statistics can provide very useful methods for parameter estimation. However, we should always be cautious with regard to assuming a parametric model within which we perform such estimations.

Appendix A

References

This appendix contains references to the consulted literature and used sources. For chapters 2-5 there is a list containing references to the relevant literature. These references do not mean that we have exactly copied them from their sources. In many cases we have expanded or improved what was given in the literature. This is the case for almost all proofs from [6]. In the case that we have literally adopted something from the literature, it is stated in the list underneath.

The first paragraph of Chapter 1 is loosely based on the introduction from [6], the remainder of that chapter is original. The same goes for Chapter 6, whenever we use things from previous chapters we mention it in the text.

All figures have been made by us using R or pgfplots, with the exception of Figure 5.1, which is an edited figure. The original figure can be found on https://www.wikiwand.com/en/Extremum_estimator.

Whenever '-' appears in the column on the right it means that we've come up with it ourselves: no literature was used.

Chapter 2: Preliminaries

2.1 Definition	[3], Definition 1.1, p.3.
2.2 Definition	[3], Definition 1.2, p.4.
2.3 Definition	[3], Definition 1.3, p.5.
2.4 Definition	[3], p.9.
2.5 Definition	[3], p.9.
2.6 Definition	-
2.7 Definition	[7], p.13.
2.8 Definition	[7], p.13.
2.9 Lemma	[7], Lemma 2.1, p.13-14. We adopted this lemma in its entirety.
2.10 Definition	[7], Definition 2.2, p.15.
2.11 Theorem	[5], Theorem A, p.178.
2.12 Theorem	[5], Theorem B, p.184.
2.13 Theorem	[2], Theorem 29.5, p.409.
2.14 Definition	-
2.15 Definition	-
2.16 Theorem	[5], Theorem A, p.300-301.

Chapter 3: Stochastic Convergence

3.1 Definition	[6], p.5.
3.2 Example	-
3.3 Example	[6], Example 2.1, p.6.
3.4 Lemma	[6], Lemma 2.2, p.6-7.
3.5 Example	[2], Example 25.6, p.352.
3.6 Definition	[6], p.8.
3.7 Proposition	-
3.8 Theorem	[6], Theorem 2.3, p.7-8.
3.9 Lemma	[7], Lemma 2.5, p.9.
3.10 Definition	[6], p.5.
3.11 Definition	[6], p.6.
3.12 Example	[6], Example 2.1, p.6.
3.13 Theorem	[6], Theorem 2.3, p.7-8.
3.14 Example	-
3.15 Theorem	[6], Theorem 2.7, p.10.
3.16 Example	-
3.17 Lemma	[6], Lemma 2.8, p.11.
3.18 Example	[6], Example 2.9, p.11.
3.19 Definition	[6], p.12.
3.20 Lemma	[6], p.12-13. Proof completely by myself.
3.21 Lemma	[6], Lemma 2.12, p.13.

Chapter 4: The Delta Method

4.1 Theorem	[6], Theorem 3.1, p.26.
4.2 Example	[6], Example 3.2, p.26-27.
4.3 Example	[6], Example 3.4, p.28.
4.4 Example	[6], Lemma 2.2, p.6-7.
4.5 Definition	-
4.6 Definition	-
4.7 Example	[5], Example B, p.263.
4.8 Definition	[6], p.35.
4.9 Example	[5], Example C, p.263-264.
4.10 Theorem	[6], Theorem 4.1, p.36.
4.11 Example	-

Chapter 5: M-estimators

5.1 Definition	[6], p.41.
5.2 Example	[6], Example 5.3, p.42.
5.3 Example	[6], Example 5.4, p.42-44.
5.4 Definition	[6], p.45.
5.5 Definition	[6], p.45.
5.6 Theorem	[6], Theorem 5.7, p.45-46.
5.7 Theorem	[6], Theorem 5.9, p.46.
5.8 Lemma	[6], Lemma 5.10, p.47.
5.9 Example	[6], Example 5.11, p.47.
5.10 Theorem	[6], p.51.
5.11 Lemma	[6], p.51.
5.12 Example	[7], Example 4.12, p.48.
5.13 Theorem	[6], Theorem 5.21, p.52-53.
5.14 Example	[7], Example 4.14, p.45-46
5.15 Theorem	[6], Theorem 5.23, p.53-54.
5.16 Definition	[6], p.62.
5.17 Lemma	[6], Lemma 5.35, p.62.
5.18 Example	[6], Example 5.25, p.55-56.
5.19 Example	[6], Example 5.37, p.64.
5.20 Theorem	[7], Theorem 4.21, p.56.

Appendix B

Codes for simulations

This appendix contains the code for the simulations we used to create the figures in Chapter 6.

Code used in Section 6.1

The following code is used to obtain the simulations shown in Figure 6.1 and Figure 6.2. The code used to obtain the simulations of Figure 6.3 and Figure 6.4 can be obtained by changing the values of 7 and 2 to 3 and 5 respectively. Also, in both plots the ylim parameter should be changed to c(0, 15) and c(2, 10) respectively.

```
#In this simulation we will estimate the parameter of a Gamma
#distribution for different sample sizes.

seq = c(2,3,4,5,6,7,8,9,10,15,20,25,30,35,40,45,
50,55,60,65,70,75,80,90,100,125,150,175,
200,250,300,400,500,750,1000,1250,1500,
1750,2000,2500,3000,3500,4000,4500,5000,
7500,10000) #vector containing all the sample sizes
#for which we want to estimate the parameter of interest.

sim <- rep(0,length(seq))
sim2 <- rep(0,length(seq))

for(n in 0:47) { #for-loop that determines all estimates
X <- rgamma(seq[n],7,2)
sim[n] <- mean(X)^2/(mean(X^2)-mean(X)^2)
sim2[n] <- mean(X)/(mean(X^2)-mean(X)^2)
```

```
}

plot(seq,sim,xlab="sample size",ylab="parameter estimate
for alpha", ylim=c(0, 20), log = 'x',) #plot containing
#estimates of alpha

abline(h=7,lty=44) #dotted line signifying the actual
#value of the parameter

plot(seq,sim2,xlab="sample size",ylab="parameter estimate
for beta", ylim=c(0, 5), log = 'x') #plot containing
#estimates of beta

abline(h=2,lty=44) #dotted line signifying the actual
#value of the parameter

#The code underneath plots the densities of both Gamma
#functions under consideration
plot(seq(0,10,length=100),dgamma(seq(0,10,length=100),
7,2),type='l',ylim=c(0,0.5),ylab='density',xlab='outcome')
lines(seq(0,10,length=100),dgamma(seq(0,10,length=100),3,5),
col='orange')
```

Code used in Section 6.2

The following code is used to obtain the simulations shown in Figure 6.6.

```
#In this simulation we will show that the maximum
#likelihood estimator of the exponential distribution
#tends to a normal distribution

theta = 2 #rate of the exponential distribution

par(mfrow=c(2,2)) #this makes sure that the four
#plots are compiled into one image

#the following for-loop creates histograms that
#approximate a normal distribution for four different
#sample sizes.
for (n in c(5,20,50,10000)) {
approximation <- c();
```

```
for(i in 0:10000) { #for-loop that creates
#the data on which the histograms are based
theta_sc <- n/rgamma(1,rate=theta,shape=n)
approximation <- c(approximation,sqrt(n)*
(theta_sc - theta))
}

#The following code plots the histograms and the
#normal distribution.
hist(approximation,breaks=20,probability=T,main=
paste("Histogram with n",n,sep="="),xlim=
c(-8,8), ylim=c(0,0.2))

lines(seq(-8,8,length=100),dnorm(seq(-8,8,length=
100),mean=0,sd=theta),col='red')
}
```

Code used in Section 6.3

The following code contains the code used to create Figure 6.7 and Figure 6.8.

```
#In this simulation we will show that the sample median is
#consistent and asymptotically normal. We will also look at
#the numerical errors with regard to the population mean.

median = 0.263158 #actual value of the median of the Beta
#distribution with parameter (2,5).

seq = c(2,3,4,5,6,7,8,9,10,15,20,25,30,35,40,45,
50,55,60,65,70,75,80,90,100,125,150,175,
200,250,300,400,500,750,1000,1250,1500,
1750,2000,2500,3000,3500,4000,4500,5000,
7500,10000) #vector containing all the sample sizes
#for which we want to estimate the median of the distribution.

sim <- rep(0,length(seq))

for(n in 0:47) { #for-loop that determines all estimates
#and generates a list of the errors of the estimates
```

```
Y <- rbeta(seq[n],2,5)
sim[n] <- median(Y)
print(sim[n]-median)
}

plot(seq,sim,xlab="sample size",
ylab="Estimate for population median",
ylim=c(0, 0.75), log = 'x',) #plot containing estimates
#for population median

abline(h=median,lty=44) #this creates a dotted line
#signifying the actual value of the median

#the following for-loop creates histograms that
#approximate a normal distribution for four different
#sample sizes.
for (n in c(5,25,100)) {
approximation <- c();
for(i in 0:10000) { #for-loop that creates
#the data on which the histograms are based
medianestimate <- median(rbeta(n,2,5))
approximation <- c(approximation,sqrt(n)*
(medianestimate - median))
}
#The following code plots the histograms and the normal
#distribution.
hist(approximation,breaks=20,probability=T,main=
paste("Histogram with n",n,sep="="),xlim=
c(-0.5,0.5), ylim=c(0,2))

lines(seq(-0.5,0.5,length=100),dnorm(seq(-0.5,0.5,
length=100),mean=0,sd=0.2146),col='red')
}
```

Code used in section 6.4.

The following code is used to obtain Figure 6.9 and Figure 6.10.

```
#In this simulation we illustrate what happens when we
#have a misspecified model. We will sample from a uniform
#distribution and try to estimate the parameter of an
```

```
#exponential distribution.

seq = c(3,5,10,20,50,100,200,300,400,500,1000,1500,2000,
3000,4000,5000,10000,20000,30000,50000) #vector
#containing all the sample sizes for which we want
#to estimate the parameter of interest.
sim <- rep(0,length(seq))

for(n in 0:20) { #for-loop creating the estimate for all
#sample sizes.
X <- runif(seq[n], min=0,max=2)
sim[n] <- seq[n]/sum(X)
}

plot(seq,sim,xlab="sample size",ylab="estimate for theta",
ylim=c(0.5, 2), log = 'x',) #plot of the estimations
abline(h=1,lty=44) #this dotted line signifies the value
#to which the estimator seems to converge

#the following code plots the densities of the exponential
#and uniform distribution.
plot(seq(0,7,length=100),dexp(seq(0,7,length=100),1),
type='l',ylim=c(0,0.75), ylab= 'density',
xlab = 'outcome')
lines(seq(0,7,length=100),dunif(seq(0,7,length=100),
min=0,max=2),col='orange')
```

Appendix C

Bibliography

[1] Apostol, Tom M. *Mathematical Analysis Second Edition.* Menlo Park, California: Addison-Wesly Publishing Company, 1974.

[2] Billingsley, Patrick. *Probability and Measure.* Hoboken, New York: John Wiley & Sons, Inc., 1974.

[3] Dirksen, Sjoerd. *Stochastische Processen Lecture Notes, WISB362, Universiteit Utrecht.* 2019.

[4] Munkres, James. *Topology.* New York: Pearson, 2000.

[5] Rice, John A. *Mathematical Statistics and Data Analysis, Third Edition* Belmont, California: Brooks/Cole, 2007.

[6] Vaart, van der, A.W. *Asymptotic Statistics.* New York: Cambridge University Press, 1998.

[7] Vaart, van der, A.W. *Asymptotic Statistics Lecture Notes, 5374ASST8Y, University of Amsterdam.* 1998.