

Eine Einführung in die Schließende Statistik

An Introduction to Inferential Statistics

Christina Andersson

Gerald Kroisandt

2003

Vorwort

Preface

Dieses Skript wendet sich an die Studenten, die an der Vorlesung „Introduction to Mathematical Statistics“ an der Universität Kaiserslautern teilnehmen. Dieser Kurs gehört zum Studienprogramm „Mathematics International“, weshalb die Zuhörerschaft nicht nur unterschiedliche Vorkenntnisse, sondern auch verschiedene Muttersprachen hat. Um die ausländischen Studenten zu ermuntern, ihre Deutschkenntnisse zu verbessern, entschieden wir uns, dieses Skript zweisprachig zu verfassen. Ein weiterer Grund ist, dass die deutschen Studenten nicht nur die englischen, sondern auch die deutschen Fachbegriffe lernen.

Das Skript baut auf der Vorlesung „Stochastische Methoden“ an der Universität Kaiserslautern oder einem entsprechenden Grundlagenkurs in Wahrscheinlichkeitstheorie auf. Deshalb beinhaltet das einführende Kapitel lediglich eine kurze Zusammenfassung der Grundbegriffe aus der Statistik und Wahrscheinlichkeitstheorie.

Im zweiten Kapitel wird die Punktschätzung betrachtet und wünschenswerte Eigenschaften eingeführt. Maximum–Likelihood–, Minimax– und Bayes–Schätzer werden untersucht und durch Beispiele illustriert.

Im dritten Kapitel werden Konfidenzintervalle besprochen.

Das letzte Kapitel widmet sich Hypothesentests.

These lecture notes are addressing students attending the course “Introduction to Mathematical Statistics” at the University of Kaiserslautern. This course is offered within the programme “Mathematics International”. Therefore, the audience does not only have a mixed background and previous knowledge in statistics, but also different mother tongues. To encourage the foreign students to improve and use their German skills, we decided to develop these bilingual lecture notes. Another reason is that the German students should not only learn the English, but also the German technical terms.

The prerequisites assumed are the course „Stochastische Methoden“ at the University of Kaiserslautern or another basic course in probability. Therefore, the introductory chapter contains only a short review of the very basics in statistics and probability theory.

In the second chapter, point estimation is treated. Desirable properties of estimators are introduced. The maximum likelihood–, minimax– and Bayes’ estimators are described and illustrated by examples.

In Chapter 3, confidence intervals are discussed.

The main topic of the last chapter is statistical hypothesis testing.

Contents

1 Einführung			
Introduction		1	
1.1 Grundlegende Wahrscheinlichkeits- und Statistikkonzepte			
Basic concepts of probability and statistics	4		
1.1.1 Grundlagen der Wahrscheinlichkeitstheorie			
The basics of probability	5		
1.1.2 Erwartungswert und (Ko-)Varianz			
Expectation and (Co-)variance	11		
1.1.3 Beschreibende Statistik			
Descriptive statistics	15		
1.1.4 Nützliche Ungleichungen			
Useful inequalities	17		
1.1.5 Konvergenzarten			
Concepts of convergence	18		
1.1.6 Gesetze der großen Zahlen und Grenzwertsätze			
Laws of large numbers and limit theorems	20		
1.2 Beispiele von Verteilungen			
Examples of distributions	22		
1.3 Zufallsvektoren			
Random vectors	29		
2 Punktschätzung			
Point estimation		33	
2.1 Wünschenswerte Eigenschaften von Schätzern			
Desirable properties of estimators	35		
2.1.1 Konsistenz, Verlust und Risiko			
Consistency, loss and risk	35		
2.1.2 Unverzerrtheit			
Unbiasedness	39		
2.1.3 Quadratisches Risiko in Beziehung zu Bias und Varianz			
Quadratic risk in connection to bias and variance	42		
2.1.4 Suffizienz und Vollständigkeit			
Sufficiency and Completeness	44		
2.2 Bayes-Schätzer			
Bayes estimators	47		
2.3 Minimaxschätzer			
Minimax estimators	56		
2.4 Maximum-Likelihood-Schätzer			
Maximum likelihood estimators	61		
2.4.1 Grundlegende Idee und Definitionen			
Basic idea and definitions	61		
2.4.2 Beispiele			
Examples	63		
2.4.3 Eigenschaften			
Properties	66		
2.4.4 Interessante Beispiele von Maximum-Likelihood-Schätzern			
Interesting examples of maximum likelihood estimators	77		
2.5 Konsistenz und asymptotische Normalität von M-Schätzern			
Consistency and asymptotic normality of M-estimators	80		
3 Intervallschätzung			
Interval estimation		93	
3.1 Einige Verteilungen und nützliche Eigenschaften			
Some distributions and properties needed	94		
3.1.1 Die Dichten der χ^2 , t und F-Verteilung			
The densities of the χ^2 , t and F-distribution	98		
3.1.2 Schätzer der Parameter einer Normalverteilung und ihre Verteilungen			
Estimators of the parameters of a normal distribution and their distributions	101		
3.2 Konfidenzintervalle für die Parameter einiger gewöhnlichen Verteilungen			
Confidence intervals for the parameters of some common distributions	104		

4	Testtheorie	
	Test theory	115
4.1	Eine allgemeine Beschreibung von Hypothesentests A general description of hypothesis testing	115
4.2	Tests für normalverteilte Daten Tests for normally distributed data	122
4.3	Likelihood–Quotienten–Tests Likelihood ratio tests	133
4.3.1	Gleichmäßig beste Tests Uniformly most powerful tests	133
4.3.2	Allgemeine Likelihood–Quotienten–Tests General likelihood ratio tests	141
4.4	χ^2 –Tests χ^2 –tests	151
4.4.1	Herleitung des χ^2 –Tests Derivation of the χ^2 –test	151
4.4.2	Goodness–of–Fit Tests Goodness–of–Fit tests	161
4.4.3	Unabhängigkeitstest Test of independence	164
5	Literatur	
	Literature	169

1 Einführung

Introduction

Die Theorie und Praxis der schließenden Statistik handeln von Problemen der folgenden Art:

Basierend auf einem beobachteten Datenvektor $\mathbf{x} = (x_1, \dots, x_n)$, der eine Realisierung der Zufallsvariable \mathbf{X} ist, wollen wir Schlussfolgerungen über die Verteilung von \mathbf{X} , $\mathcal{L}(\mathbf{X})$ genannt, ziehen. Unser Wissen über $\mathcal{L}(\mathbf{X})$ ist unvollständig und wir benutzen ein statistisches Modell um Schlussfolgerungen über $\mathcal{L}(\mathbf{X})$ zu ziehen.

Wir unterscheiden zwischen parametrischen und nicht-parametrischen statistischen Modellen:

Definition 1.0.1 (Parametrisches statistisches Modell)

Unser statistisches Modell ist, dass die Beobachtungen $\mathbf{x} = (x_1, \dots, x_n)$ Realisierungen eines Zufallsvektors \mathbf{X} (stetig oder diskret), mit einer Verteilung $\mathcal{L}(\mathbf{X})$, sind. Für die Verteilung $\mathcal{L}(\mathbf{X})$ wird eine bekannte funktionale Form angenommen. Diese hängt nur vom endlich-dimensionalen Parametervektor ϑ aus dem Parameterraum Θ ab, d.h.

$$\mathcal{L}(\mathbf{X}) \in \{\mathcal{P}_\vartheta; \vartheta \in \Theta\}.$$

The theory and practice of statistical inference are concerned with problems of the following kind:

Based on an observed data vector $\mathbf{x} = (x_1, \dots, x_n)$, which is a realization of the random variable \mathbf{X} , we want to draw conclusions about the distribution of \mathbf{X} , called $\mathcal{L}(\mathbf{X})$. Our knowledge of $\mathcal{L}(\mathbf{X})$ is incomplete and we use a statistical model to conclude about $\mathcal{L}(\mathbf{X})$.

We distinguish between parametric and non-parametric statistical models:

Definition 1.0.1 (Parametric statistical model)

Our statistical model is that the observations $\mathbf{x} = (x_1, \dots, x_n)$, are realizations of a random vector \mathbf{X} (continuous or discrete) with a distribution $\mathcal{L}(\mathbf{X})$. The distribution $\mathcal{L}(\mathbf{X})$ is assumed to have a known functional form. This depends only on the finite-dimensional parameter vector ϑ from the parameter space Θ , i.e.

Definition 1.0.2 (Nicht-parametrisches statistisches Modell)

In einem nicht-parametrischen statistischen Modell ist die funktionale Form der Verteilung des Zufallsvektors \mathbf{X} vollkommen unbekannt.

Bemerkung 1.0.3

Dieses Vorlesungsskript konzentriert sich auf parametrische statistische Modelle.

Anhand eines Beispiels betrachten wir die drei statistischen Hauptprobleme, die wir durchgängig in diesem Vorlesungsskript antreffen werden:

Beispiel 1.0.4 (Die Binomialverteilung)

10 StudentInnen des Studiengangs „Mathematics International“ der Universität Kaiserslautern wundern sich über die deutsche Gewohnheit, im Pfälzer Wald bei der Universität zu wandern. Sie wollen dies auch mal machen und entscheiden sich zu einer Tour zum Humbergturm und zurück zur Universität.

Definition 1.0.2 (Non-parametric statistical model)

In a non-parametric statistical model, the functional form of the distribution of the random vector \mathbf{X} is totally unknown.

Remark 1.0.3

These lecture notes concentrate on parametric statistical models.

We use an example to describe which three statistical main problems we are going to meet throughout these lecture notes:

Example 1.0.4 (The binomial distribution)

10 mathematics international students from the University of Kaiserslautern wonder at the German habit to go hiking in the Palatinate forest in the surroundings of the university. They also want to test this and decide to hike to the Humberg tower and back to the university again.

Bevor sie aufbrechen, erkundigen sie sich ängstlich über die Möglichkeit, wilde Tiere im Wald zu treffen. Ein deutscher Statistikstudent sagt ihnen, dass es keine gefährlichen Tiere im Wald gibt, und dass das interessanteste Tier, das sie antreffen könnten, die sogenannte Elwedritsche ist. Dies ist ein Märchentier, das nur im Pfälzer Wald „antreffbar“ ist.

Before they leave, they anxiously ask about the possibility to meet wild animals in the forest. A German statistics student tells them that there are no dangerous animals in the forest and that the most interesting animal, which they might meet, is the so-called elwedritsche. This is a fairy-tale animal, only to be “found” in the forest of Palatinate.

Nun will der Student ihnen „natürlich“ auch noch seine Statistikkennnisse näher bringen und lehrt sie die Binomialverteilung: Wenn sie insgesamt n Tiere während ihrer Wanderung beobachten, dann kann die Anzahl der beobachteten Elwedritschen als eine Zufallsvariable X modelliert werden, die binomialverteilt ist mit Parameter p , d.h. $X \sim \mathcal{B}(n, p)$ und

Now “of course” the statistics student wants to tell them about his statistics knowledge and he teaches them the binomial distribution: If they totally observe n animals during their hike, then the number of elwedritches observed can be modelled as a random variable X , which is binomially distributed with parameter p , i.e. $X \sim \mathcal{B}(n, p)$ and

$$\mathcal{P}(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}, i = 1, 2, \dots$$

Er fügt hinzu, dass er in der Formel die „bekannte“ Tatsache berücksichtigt hat, dass die Elwedritsche kein Rudeltier ist und somit jede Elwedritsche unabhängig von den anderen auftritt.

He adds, that in the formula, he has taken into account the “well-known” fact that the elwedritsche is no herd animal, so each elwedritsche appears independently from the others.

Der Parameter p , $0 \leq p \leq 1$ ist die Wahrscheinlichkeit, dass ein beobachtetes Tier eine Elwedritsche ist. Akzeptiert man dieses statistische Modell, so ist die Verteilung der Anzahl der Elwedritschen unter n beobachteten Tieren bis auf den Parameter p bekannt.

The parameter p , $0 \leq p \leq 1$ is the probability that one observed animal is an elwedritsche. So, accepting this statistical model, the distribution of the number of elwedritches among n observed animals is known except from the parameter p .

Die StudentInnen wollen nun statistische Methoden benutzen, um Schlussfolgerungen über p zu ziehen. Sie kommen zu dem Schluss, dass folgende drei Fragen beantwortet werden müssen:

The students want to use statistical methods to draw conclusions about p . They arrive at the conclusion that the following three questions must be answered:

i) Punktschätzung

Wie groß ist p ? Welche sind die Methoden um einen geschätzten Wert für p zu erhalten?

i) Point estimation

How big is p ? What are the methods to obtain an estimated value for p ?

ii) Intervallschätzung

Wie können wir ein Intervall finden, worin p mit einer gewissen Mindestwahrscheinlichkeit liegt?

ii) Interval estimation

How can we find an interval where p is located with at least a certain probability?

iii) Hypothesentest

Wie können wir entscheiden, ob eine gewisse Behauptung, wie z.B. „ $p \leq p_0$ “, wahr oder falsch ist?

iii) Hypothesis testing

How can we conclude about the truth or falsehood of a certain statement like e.g. “ $p \leq p_0$ ”?

1.1 Grundlegende Wahrscheinlichkeits- und Statistikkonzepte Basic concepts of probability and statistics

Hier werden wir die Grundlagen der Wahrscheinlichkeitstheorie und Statistik wiederholen. Wir nehmen an, dass die Sätze bekannt sind und präsentieren sie deshalb ohne Beweise.

Here, we review the very basics of probability and statistics. We assume that the theorems are well-known and they are therefore presented without proofs.

1.1.1 Grundlagen der Wahrscheinlichkeitstheorie
The basics of probability

Wir fangen mit einigen wichtigen Definitionen an:

Definition 1.1.1 (σ -Algebra)

Sei Ω eine Menge.
 Eine σ -Algebra \mathfrak{A} ist ein Teilmengensystem von Ω mit folgenden Eigenschaften:

1. $\emptyset \in \mathfrak{A}$
2. $A \in \mathfrak{A} \Rightarrow A^c \in \mathfrak{A}$
3. $A_1, A_2, \dots \in \mathfrak{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathfrak{A}$

Das folgende Lemma zeigt, dass man aus einem gegebenen Teilmengensystem immer eine kleinste σ -Algebra erhalten kann.

Lemma 1.1.2

Sei \mathfrak{S} ein beliebiges Teilmengensystem von Ω .
 Unter allen σ -Algebren, die \mathfrak{S} beinhalten, gibt es eine kleinste.
 Diese bezeichnen wir mit $\sigma(\mathfrak{S})$.

In \mathbb{R}^N gibt es eine besondere σ -Algebra, die sogenannte Borel- σ -Algebra.

We start with some important definitions:

Definition 1.1.1 (σ -algebra)

Let Ω be some set.
 A σ -algebra \mathfrak{A} is a system of subsets of Ω with the following properties:

1. $\emptyset \in \mathfrak{A}$
2. $A \in \mathfrak{A} \Rightarrow A^c \in \mathfrak{A}$
3. $A_1, A_2, \dots \in \mathfrak{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathfrak{A}$

The following lemma shows the possibility to obtain the smallest σ -algebra for a given system of subsets.

Lemma 1.1.2

Let \mathfrak{S} be an arbitrary system of subsets of Ω .
 Under all σ -algebras including \mathfrak{S} , there is a smallest one.
 We denote this σ -algebra by $\sigma(\mathfrak{S})$.

In \mathbb{R}^N , we have a special σ -algebra, the so-called Borel- σ -algebra.

Definition 1.1.3 (Borel- σ -Algebra)

Die Borel- σ -Algebra \mathcal{B} von \mathbb{R}^N ist die kleinste σ -Algebra, die alle offenen Teilmengen des \mathbb{R}^N enthält.

Definition 1.1.4 (Wahrscheinlichkeitsmaß)

Sei (Ω, \mathfrak{A}) ein Maßraum.
 $\mathcal{P}: \mathfrak{A} \rightarrow [0, 1]$ ist ein Wahrscheinlichkeitsmaß auf (Ω, \mathfrak{A}) , falls folgende Eigenschaften erfüllt sind:

1. $\mathcal{P}(\emptyset) = 0, \mathcal{P}(\Omega) = 1,$
2. σ -Additivität:
 $A_1, A_2, \dots \in \mathfrak{A}$ paarweise disjunkt
 $\Rightarrow \mathcal{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathcal{P}(A_n),$
3. $\mathcal{P}(A^c) = 1 - \mathcal{P}(A)$ für $A \in \mathfrak{A}.$

Satz 1.1.5 (Bayes' Satz)

Sei A_1, \dots, A_n eine disjunkte Partition von Ω .
 Dann gilt für jede Menge $B \in \mathfrak{A}$

$$\mathcal{P}(B) = \sum_{i=1}^n \mathcal{P}(B|A_i) \cdot \mathcal{P}(A_i),$$

wobei $\mathcal{P}(B|A_i)$ die bedingte Wahrscheinlichkeit von B unter der Bedingung, dass A_i eintritt, ist.

Definition 1.1.3 (Borel- σ -algebra)

The Borel- σ -algebra \mathcal{B} of \mathbb{R}^N is the smallest σ -algebra including all open subsets of \mathbb{R}^N .

Definition 1.1.4 (Probability measure)

Let (Ω, \mathfrak{A}) be a measurable space.
 $\mathcal{P}: \mathfrak{A} \rightarrow [0, 1]$ is a probability measure on (Ω, \mathfrak{A}) if the following properties are fulfilled:

1. $\mathcal{P}(\emptyset) = 0, \mathcal{P}(\Omega) = 1,$
2. σ -additivity:
 $A_1, A_2, \dots \in \mathfrak{A}$ pairwise disjoint
 $\Rightarrow \mathcal{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathcal{P}(A_n),$
3. $\mathcal{P}(A^c) = 1 - \mathcal{P}(A)$ for $A \in \mathfrak{A}.$

Theorem 1.1.5 (Bayes' Theorem)

Let A_1, \dots, A_n be a disjoint partition of Ω .
 Then it holds for every set $B \in \mathfrak{A}$

where $\mathcal{P}(B|A_i)$ denotes the conditional probability of B under the condition that A_i occurs.

Definition 1.1.6 (Zufallsvariable)

Sei $(\Omega, \mathfrak{A}, \mathcal{P})$ ein Wahrscheinlichkeitsraum und $(\mathbf{X}, \mathfrak{C})$ ein messbarer Raum, der sogenannte Beobachtungsraum.

Eine Zufallsvariable $X : \Omega \rightarrow \mathbf{X}$ ist eine messbare Abbildung vom Wahrscheinlichkeitsraum in den Beobachtungsraum.

Das induzierte (Wahrscheinlichkeits-)Maß auf dem Beobachtungsraum wird ebenfalls \mathcal{P} genannt.

In dem reellen Fall erhalten wir:

Definition 1.1.7 (Zufallsvariable)

Eine Zufallsvariable $X : \Omega \rightarrow \mathbb{R}$ ist eine messbare Funktion, die zu jedem Element $\omega \in \Omega$ eine reelle Zahl zuordnet so dass für jedes $x \in \mathbb{R}$ die Wahrscheinlichkeit $\mathcal{P}(X(\omega) \leq x)$ definiert ist.

Bemerkung 1.1.8

Es ist also ein enormer Unterschied zwischen einer Zufallsvariable und ihre Realisation, die ein festes Element aus dem Beobachtungsraum ist!

In diesem Vorlesungsskript benutzen wir große Buchstaben für Zufallsvariablen X und kleine Buchstaben x für die Realisierung.

Definition 1.1.6 (Random variable)

Let $(\Omega, \mathfrak{A}, \mathcal{P})$ be a probability space and $(\mathbf{X}, \mathfrak{C})$ a measurable space, the so-called observation space.

A random variable $X : \Omega \rightarrow \mathbf{X}$ is a measurable mapping from the probability space into the observation space.

The induced (probability) measure on the observation space will be called \mathcal{P} , too.

In the real case, we obtain:

Definition 1.1.7 (Random variable)

A random variable $X : \Omega \rightarrow \mathbb{R}$ is a measurable function assigning a real number to each element $\omega \in \Omega$ such that for every $x \in \mathbb{R}$ the probability $\mathcal{P}(X(\omega) \leq x)$ is defined.

Remark 1.1.8

Thus, there is an enormous difference between a random variable and its realization, which is a fixed element from the observation space!

In these lecture notes, we use capital letters for random variables X and a lower case x for the realization.

Beispiel 1.1.9 (Zufallsvariable kontra Realisierung)

Die MathematikstudentInnen vom Beispiel 1.0.4 auf Seite 2 wollen etwas zum Essen bevor sie aufbrechen. Sie gehen deshalb zur Mensa, die heute Schnitzel anbietet.

Sei X_i die Zufallsvariable, die das Gewicht X_i des hypothetischen Schnitzels i repräsentiert, das den StudentInnen verkauft wird. Dann wird der/die zehnte StudentIn der Schlange das zehnte Schnitzel mit gemessenem Gewicht x_{10} erhalten. Dieses beobachtete Gewicht, z.B. $x_{10} = 165\text{g}$ ist dann die beobachtete Realisierung der Zufallsvariable X_{10} .

Um das Verhalten einer Zufallsvariablen zu studieren, brauchen wir den Begriff der Verteilung:

Definition 1.1.10 (Verteilung)

Ein Wahrscheinlichkeitsmaß \mathcal{P} wird Verteilung der Zufallsvariable $\mathbf{X} : \Omega \rightarrow \mathbb{R}^N$ genannt wenn

$$\mathcal{P}(\mathbf{X} \in B) = \mathcal{P}(B) \quad \forall B \in \mathfrak{B}.$$

Bemerkung 1.1.11

Wir benutzen die Notation $\mathcal{P} = \mathcal{L}(\mathbf{X})$.

Example 1.1.9 (Random variable versus realization)

The mathematics students from Example 1.0.4 on page 2 want some food before they leave.

Therefore, they go to the student's refectory, today offering steaks.

Let X_i be the random variable representing the weight of a hypothetical steak i offered to the students. Then the tenth student of the queue will get the tenth steak produced with the recorded weight x_{10} . This observed weight, e.g. $x_{10} = 165\text{g}$ is then the observed realization of the random variable X_{10} .

To study how the random variable behaves, we need the concept of distribution:

Definition 1.1.10 (Distribution)

A probability measure \mathcal{P} is called distribution of the random variable $\mathbf{X} : \Omega \rightarrow \mathbb{R}^N$ if

Remark 1.1.11

We use the notation $\mathcal{P} = \mathcal{L}(\mathbf{X})$.

Definition 1.1.12 (Verteilungsfunktion)

Sei \mathbf{X} eine Zufallsvariable und \mathcal{P} ein Wahrscheinlichkeitsmaß.
Dann ist $F(\mathbf{y}) = \mathcal{P}(\mathbf{X} \leq \mathbf{y})$ die Verteilungsfunktion von \mathbf{X} .

Wir unterscheiden zwischen stetigen und diskreten Zufallsvariablen:

Definition 1.1.13 (Absolut-stetige Verteilung)

$\mathcal{L}(\mathbf{X})$ ist absolut-stetig wenn eine Dichte $f(\mathbf{x}) \geq 0$ existiert, so dass

1. $\int_{\mathbb{R}^N} f(\mathbf{x}) d\mathbf{x} = 1$
2. $\mathcal{P}(B) = \int_B f(\mathbf{x}) d\mathbf{x}$
3. $F(\mathbf{y}) = \int_{-\infty}^{\mathbf{y}} f(\mathbf{x}) d\mathbf{x}$
4. $f(x) = F'(x)$ fast überall, wenn $x \in \mathbb{R}^1$.

Definition 1.1.14 (Diskrete Verteilung)

$\mathcal{L}(\mathbf{X})$ ist diskret, wenn die Zufallsvariable \mathbf{X} abzählbar viele unterschiedliche Werte $\mathbf{x}_1, \mathbf{x}_2, \dots$ annimmt, mit Gewichten $p_n \geq 0$, so dass:

Definition 1.1.12 (Distribution function)

Let \mathbf{X} be a random variable and \mathcal{P} a probability measure.
Then $F(\mathbf{y}) = \mathcal{P}(\mathbf{X} \leq \mathbf{y})$ is the distribution function of \mathbf{X} .

We distinguish between continuous and discrete random variables:

Definition 1.1.13 (Absolutely continuous distribution)

$\mathcal{L}(\mathbf{X})$ is absolutely continuous if there is a density function $f(\mathbf{x}) \geq 0$ such that

1. $\int_{\mathbb{R}^N} f(\mathbf{x}) d\mathbf{x} = 1$
2. $\mathcal{P}(B) = \int_B f(\mathbf{x}) d\mathbf{x}$
3. $F(\mathbf{y}) = \int_{-\infty}^{\mathbf{y}} f(\mathbf{x}) d\mathbf{x}$
4. $f(x) = F'(x)$ almost everywhere if $x \in \mathbb{R}^1$.

Definition 1.1.14 (Discrete distribution)

$\mathcal{L}(\mathbf{X})$ is discrete if the random variable \mathbf{X} takes at most countably many different values $\mathbf{x}_1, \mathbf{x}_2, \dots$ with weights $p_n \geq 0$ such that:

$$1. \sum_{n=1}^{\infty} p_n = 1$$

$$2. \mathcal{P}(B) = \sum_{n=1}^{\infty} p_n \cdot 1_B(\mathbf{x}_n)$$

$$3. \mathcal{P}(\{\mathbf{x}_n\}) = p_n = \mathcal{P}(\mathbf{X} = \mathbf{x}_n)$$

wobei $1_B(\mathbf{x}_n)$ die Indikatorfunktion der Menge B ist.

Wenn die Verteilung unbekannt ist, benutzen wir die Daten, um die empirische Verteilungsfunktion zu konstruieren:

Definition 1.1.15 (Empirische Verteilungsfunktion)

Für $n \in \mathbb{N}$ und $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^N$ ist die empirische Verteilungsfunktion $F_n(\cdot; \mathbf{x}_1, \dots, \mathbf{x}_n) : \mathbb{R}^N \rightarrow [0, 1]$ definiert als

$$F_n(\mathbf{z}; \mathbf{x}_1, \dots, \mathbf{x}_n) = \frac{1}{n} \cdot \sum_{i=1}^n 1_{x_i \leq \mathbf{z}}, \quad \mathbf{z} \in \mathbb{R}^N.$$

Satz 1.1.16 (Glivenko-Cantelli)

Wenn X_1, X_2, \dots eine Folge von unabhängigen Zufallsvariablen ist, dann gilt

$$\mathcal{P} \left(\lim_{n \rightarrow \infty} \sup_{\mathbf{z} \in \mathbb{R}^N} |F_n(\mathbf{z}; \mathbf{x}_1, \dots, \mathbf{x}_n) - F(\mathbf{z})| = 0 \right) = 1.$$

In vielen Situationen wird von unabhängigen und identisch verteilten Zufallsvariablen ausgegangen.

$$1. \sum_{n=1}^{\infty} p_n = 1$$

$$2. \mathcal{P}(B) = \sum_{n=1}^{\infty} p_n \cdot 1_B(\mathbf{x}_n)$$

$$3. \mathcal{P}(\{\mathbf{x}_n\}) = p_n = \mathcal{P}(\mathbf{X} = \mathbf{x}_n)$$

where $1_B(\mathbf{x}_n)$ is the indicator function of the set B .

If the distribution is unknown, we use the data to construct the empirical distribution function:

Definition 1.1.15 (Empirical distribution function)

For $n \in \mathbb{N}$ and $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^N$ is the empirical distribution function $F_n(\cdot; \mathbf{x}_1, \dots, \mathbf{x}_n) : \mathbb{R}^N \rightarrow [0, 1]$ defined as

Theorem 1.1.16 (Glivenko-Cantelli)

If X_1, X_2, \dots is a sequence of independent random variables, then it holds that

In many situations, we assume to have independent and identically distributed random variables.

Definition 1.1.17 (Unabhängige Zufallsvariablen)

Seien X_1, \dots, X_n Zufallsvariablen. Sie sind unabhängig wenn

$$\mathcal{P}(X_1 \in B_1, \dots, X_n \in B_n) = \prod_{i=1}^n \mathcal{P}(X_i \in B_i) \quad \forall B_1, \dots, B_n \in \mathfrak{B}.$$

Definition 1.1.17 (Independent random variables)

Let X_1, \dots, X_n be random variables. They are independent if

Definition 1.1.18 (Identisch verteilte Zufallsvariablen)

Seien X_1, \dots, X_n Zufallsvariablen. Sie sind identisch verteilt, falls

$$\mathcal{L}(X_k) = \mathcal{L}(X_1), \quad k = 2, \dots, n.$$

Definition 1.1.18 (Identically distributed random variables)

Let X_1, \dots, X_n be random variables. They are identically distributed if

Bemerkung 1.1.19

Dass die Zufallsvariablen X_1, \dots, X_n unabhängig und identisch verteilt sind, wird mit u.i.v. bezeichnet.

Remark 1.1.19

That random variables X_1, \dots, X_n are independent and identically distributed is denoted by i.i.d.

1.1.2 Erwartungswert und (Ko-)Varianz Expectation and (Co-)variance

Oft werden die ersten und zweiten Momente der Verteilung, d.h. der Erwartungswert und die Varianz, benutzt um die Zufallsvariable genauer zu beschreiben.

Frequently, the first and second moments of the distribution, i.e. the expectation and the variance, are used to describe the random variable closer.

Definition 1.1.20 (Erwartungswert)

Der Erwartungswert einer absolut-stetigen Zufallsvariable $X : \Omega \rightarrow \mathbb{R}$ ist

$$\mathcal{E}(X) = \int_{\mathbb{R}} x d\mathcal{P}(x) = \int_{\mathbb{R}} x dF(x) = \int_{\mathbb{R}} x \cdot f(x) dx.$$

Definition 1.1.20 (Expectation)

The expectation of an absolutely continuous random variable $X : \Omega \rightarrow \mathbb{R}$ is

Bemerkung 1.1.21

Analog wird der Erwartungswert für eine diskrete Zufallsvariable X definiert:

$$\mathcal{E}(X) = \sum_{i=1}^{\infty} x_i \cdot \mathcal{P}(X = x_i) = \sum_{i=1}^{\infty} x_i \cdot p_i.$$

Remark 1.1.21

The expectation is defined in an analogous manner for a discrete random variable X :

Bemerkung 1.1.22 (Linearität des Erwartungswerts)

Seien X und Y Zufallsvariablen. Dann gilt

$$\mathcal{E}(a \cdot X + b \cdot Y) = a \cdot \mathcal{E}(X) + b \cdot \mathcal{E}(Y) \quad \forall a, b \in \mathbb{R},$$

d.h. der Erwartungswert ist linear.

Remark 1.1.22 (Linearity of the expectation)

Let X and Y be random variables. Then it holds

i.e. the expectation is linear.

Bemerkung 1.1.23 (Erwartungswert einer Funktion von X)

Der Erwartungswert der Funktion $g: \mathbb{R} \rightarrow \mathbb{R}$, angewendet auf die Zufallsvariable X ist

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) d\mathcal{P}(x) = \int_{\mathbb{R}} g(x) dF(x) = \int_{\mathbb{R}} g(x) \cdot f(x) dx,$$

wobei in der letzten Gleichung von einer absolut-stetigen Zufallsvariable ausgegangen wurde.

Remark 1.1.23 (Expectation of a function of X)

The expectation of the function $g: \mathbb{R} \rightarrow \mathbb{R}$, applied to the random variable X is

where we assumed an absolutely continuous random variable in the last equality.

Definition 1.1.24 (Varianz und Standardabweichung)

Die Varianz einer Zufallsvariable X ist

$$\mathcal{V}\mathcal{A}\mathcal{R}(X) = \sigma^2(X) = \mathbb{E}\{[X - \mathbb{E}(X)]^2\}$$

und die Standardabweichung ist

$$\sigma(X) = \sqrt{\mathcal{V}\mathcal{A}\mathcal{R}(X)}.$$

Definition 1.1.24 (Variance and standard deviation)

The variance of the random variable X is

and the standard deviation is

Bemerkung 1.1.25

Für die Varianz der Zufallsvariable X gilt

$$\mathcal{V}\mathcal{A}\mathcal{R}(a \cdot X) = a^2 \cdot \mathcal{V}\mathcal{A}\mathcal{R}(X) \quad \forall a \in \mathbb{R}.$$

Um die lineare Abhängigkeit zwischen zwei Zufallsvariablen zu messen, führen wir die Kovarianz und Korrelation ein.

Remark 1.1.25

For the variance of the random variable X holds

We introduce the covariance and correlation to measure the linear dependence between two random variables.

Definition 1.1.26 (Kovarianz)

X und Y seien Zufallsvariablen.

Dann ist die Kovarianz von X und Y definiert als

$$COV(X, Y) = \mathbb{E}\{[X - \mathbb{E}(X)] \cdot [Y - \mathbb{E}(Y)]\}.$$

Definition 1.1.27 (Korrelation)

X und Y seien Zufallsvariablen.

Dann ist die Korrelation von X und Y definiert als

$$CORR(X, Y) = \frac{COV(X, Y)}{\sqrt{\mathcal{V}\mathcal{A}\mathcal{R}(X) \cdot \mathcal{V}\mathcal{A}\mathcal{R}(Y)}}.$$

Bemerkung 1.1.28

X und Y sind unkorreliert, falls

$$CORR(X, Y) = 0.$$

Bemerkung 1.1.29

Wir haben, dass $-1 \leq CORR(X, Y) \leq 1$.

$$CORR(X, Y) = \begin{cases} -1 \\ 1 \end{cases} \Leftrightarrow Y = a \cdot X + b; \quad b \in \mathbb{R}, \quad a \begin{cases} < 0 \\ > 0 \end{cases}$$

Definition 1.1.26 (Covariance)

Let X and Y be random variables.

Then the covariance of X and Y is defined as

Definition 1.1.27 (Correlation)

Let X and Y be random variables.

Then the correlation of X and Y is defined as

Remark 1.1.28

X and Y are uncorrelated if $CORR(X, Y) = 0$.

Remark 1.1.29

We have that $-1 \leq CORR(X, Y) \leq 1$.

Bemerkung 1.1.30

*X und Y sind unabhängig
 $\Rightarrow X$ und Y sind unkorreliert.
 Allerdings ist die Umkehrung im allgemeinen nicht wahr, wie in der nächsten Bemerkung erklärt wird.*

Remark 1.1.30

*X and Y are independent
 $\Rightarrow X$ and Y are uncorrelated.
 However, the converse is in general not true, as explained in the next remark.*

Bemerkung 1.1.31

*Korrelation ist ein Maß für die lineare Abhängigkeit von X und Y .
 Falls nun X und Y unkorreliert sind, so sind sie nicht notwendigerweise unabhängig, weil die Abhängigkeit nicht-linear sein kann.*

Remark 1.1.31

*Correlation is a measure of the linear dependence of X and Y .
 If X and Y are uncorrelated, then they are not necessarily independent, since the dependence can be non-linear.*

Bemerkung 1.1.32

Sind X und Y Zufallsvariablen, so gilt

$$\mathcal{V}\mathcal{A}\mathcal{R}(X+Y) = \mathcal{V}\mathcal{A}\mathcal{R}(X) + \mathcal{V}\mathcal{A}\mathcal{R}(Y) + COV(X, Y),$$

d.h. sind sie unkorreliert, so addiert sich lediglich die Varianz.

Remark 1.1.32

If X and Y are random variables, then

i.e. if they are uncorrelated, then the variance is just added.

1.1.3 Beschreibende Statistik Descriptive statistics

Wir führen einige Statistiken ein, die helfen werden, um die Daten zu beschreiben.
 Im Folgenden seien X_1, \dots, X_n Zufallsvariablen.

Zuerst betrachten wir einige Lokalisierungsmaße:

We introduce some statistics that will help in describing the data.
 In the following, we let X_1, \dots, X_n be random variables.

First, we consider some measures of location:

Definition 1.1.33 (Aritmetischer Mittelwert)

Der arithmetische Mittelwert ist definiert als

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

Definition 1.1.33 (Arithmetic mean)

The arithmetic mean is defined as

Definition 1.1.34 (Stichprobenmedian)

Der Stichprobenmedian wird auf die folgende Weise ermittelt:

$$\dot{X} = \begin{cases} X_{(k+1)}, & n = 2k + 1 \\ \frac{1}{2}(X_{(k)} + X_{(k+1)}), & n = 2k, \end{cases}$$

wobei wir die Orderstatistik benutzen, die wie folgt definiert ist:

Definition 1.1.35 (Ordnungsstatistik)

Seien die Zufallsvariablen X_1, \dots, X_n , gegeben, so ist die Ordnungsstatistik einfach die Zufallsvariablen der Größe nach geordnet, geschrieben als

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}.$$

Definition 1.1.36 (Modus)

Der Modus ist für diskrete Daten der am häufigsten beobachtete Wert.

Definition 1.1.34 (Sample median)

The sample median is determined in the following way:

where we use the order statistics defined as follows:

Definition 1.1.35 (Order statistic)

Given random variables X_1, \dots, X_n , then the order statistic is simply the ordered random variables, written as

Definition 1.1.36 (Mode)

The mode is for discrete data the most frequently observed value.

Jetzt führen wir einige Streuungsmaße ein:

Now, we introduce some measures of spread:

Definition 1.1.37 (Stichprobenvarianz)

Die Stichprobenvarianz ist gegeben durch

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

Definition 1.1.37 (Sample variance)

The sample variance is given by

Satz 1.1.40 (Tschebyscheffsche Ungleichung)

Sei X eine reelle Zufallsvariable, deren Varianz bzw. $\mathcal{E}(X^2)$ existiert.

Dann gilt

$$\mathcal{P}(|X - \mathcal{E}(X)| \geq \epsilon) \leq \frac{1}{\epsilon^2} \mathcal{V} \mathcal{A} \mathcal{R}(X) \quad \forall \epsilon > 0.$$

Theorem 1.1.40 (Chebyshev's inequality)

Let X be a real random variable and let the variance or $\mathcal{E}(X^2)$ respectively exist.

Then it holds

Definition 1.1.38 (Spannweite)

Die Spannweite ist

$$X_{\text{range}} = X_{(n)} - X_{(1)}.$$

Definition 1.1.38 (Range)

The range is

Satz 1.1.41 (Jensensche Ungleichung)

Sei X eine reelle Zufallsvariable mit existierendem Erwartungswert $\mathcal{E}(X)$.

Sei $g : \mathbb{R} \rightarrow \mathbb{R}$ konvex und $\mathcal{E}[g(X)]$ existiere ebenfalls.

Dann gilt

$$g[\mathcal{E}(X)] \leq \mathcal{E}[g(X)].$$

Theorem 1.1.41 (Jensen's inequality)

Let X be a real random variable with finite expectation $\mathcal{E}(X)$.

Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be convex and let the expectation $\mathcal{E}[g(X)]$ exist, too.

Then it holds

**1.1.4 Nützliche Ungleichungen
Useful inequalities**

Hier werden einige Ungleichungen vorgestellt:

Here some inequalities are presented:

Satz 1.1.39 (Markovsche Ungleichung)

Sei X eine reelle Zufallsvariable, deren Erwartungswert $\mathcal{E}(X)$ existiert.

Dann gilt

$$\mathcal{P}(|X| \geq \epsilon) \leq \frac{1}{\epsilon} \mathcal{E}(|X|) \quad \forall \epsilon > 0.$$

Theorem 1.1.39 (Markov's inequality)

Let X be a real random variable and let the expectation $\mathcal{E}(X)$ exist.

Then it holds

Satz 1.1.42 (Cauchy-Schwartzsche Ungleichung)

Seien X und Y reelle Zufallsvariablen mit endlichen Erwartungswerten $\mathcal{E}(X^2)$ und $\mathcal{E}(Y^2)$.

Dann gilt

$$[\mathcal{E}(XY)]^2 \leq \mathcal{E}(X^2) \cdot \mathcal{E}(Y^2).$$

Theorem 1.1.42 (Cauchy-Schwartz' inequality)

Let X and Y be real random variables and let the expectations $\mathcal{E}(X^2)$ and $\mathcal{E}(Y^2)$ exist.

Then it holds that

**1.1.5 Konvergenzarten
Concepts of convergence**

Hier führen wir fünf gewöhnliche Konvergenzarten ein.

Sei dazu $X_n, n \in \mathbb{N}$ eine Folge von Zufallsvariablen.

Here we introduce five common convergence concepts.

Suppose $X_n, n \in \mathbb{N}$, is a sequence of random variables.

Definition 1.1.43 (Konvergenz im p -ten Mittel)

X_n konvergiert gegen X im p -ten Mittel für $p \geq 1$, $X_n \xrightarrow{L^p} X$, wenn

$$\mathcal{E}\{(X_n - X)^p\} \rightarrow 0, n \rightarrow \infty.$$

Definition 1.1.43 (Convergence in p -th mean)

X_n converges to X in p -th mean where $p \geq 1$, $X_n \xrightarrow{L^p} X$, if

Definition 1.1.44 (Konvergenz in Wahrscheinlichkeit)

X_n konvergiert gegen X in Wahrscheinlichkeit, $X_n \xrightarrow{P} X$, wenn für alle $\varepsilon > 0$,

$$\mathcal{P}(|X_n - X| > \varepsilon) \rightarrow 0, n \rightarrow \infty.$$

Definition 1.1.44 (Convergence in probability)

X_n converges to X in probability, $X_n \xrightarrow{P} X$, if for all $\varepsilon > 0$,

Definition 1.1.45 (Fast sichere Konvergenz)

X_n konvergiert gegen X fast sicher, $X_n \xrightarrow{a.s.} X$, wenn

$$\mathcal{P}(\lim_{n \rightarrow \infty} X_n = X) = 1.$$

Definition 1.1.45 (Almost sure convergence)

X_n converges to X almost surely, $X_n \xrightarrow{a.s.} X$, if

Definition 1.1.46 (Konvergenz in Verteilung)

Angenommen, die Zufallsvariablen X_1, X_2, \dots haben die Verteilungsfunktionen F_1, F_2, \dots . Wir haben dann Konvergenz in Verteilung, $X_n \xrightarrow{L} X$, wenn

$$F_n(x) \rightarrow F(x), n \rightarrow \infty$$

Definition 1.1.46 (Convergence in distribution)

Suppose that the random variables X_1, X_2, \dots have the distribution functions F_1, F_2, \dots . We have then convergence in distribution, $X_n \xrightarrow{L} X$, if

für alle Stetigkeitspunkte von F .

for all points of continuity of F .

Bemerkung 1.1.47

Die folgende Implikationen gelten:

$$X_n \xrightarrow{L^p} X \Rightarrow X_n \xrightarrow{L^q} X \quad \forall p > q \geq 1$$

$$X_n \xrightarrow{L^p} X \Rightarrow X_n \xrightarrow{P} X \quad \forall p \geq 1$$

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{P} X$$

$$X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{L} X.$$

Remark 1.1.47

The following implications hold:

1.1.6 Gesetze der großen Zahlen und Grenzwertsätze

Laws of large numbers and limit theorems

Die Gesetze der großen Zahlen beschreiben das Konvergenzverhalten von Schätzern beim Anwachsen der Stichprobengröße.

Die zentralen Grenzwertsätze spezifizieren unter welchen Bedingungen eine Folge von Verteilungen gegen eine Normalverteilung konvergiert.

The laws of large numbers describe the convergence behavior of estimators as the sample size grows.

The central limit theorems specify under which conditions a sequence of distributions converges toward a normal distribution.

Satz 1.1.48 (Das schwache Gesetz der großen Zahlen)

Sei $X_i, i \in \mathbb{N}$, eine Folge von u.i.v. Zufallsvariablen mit Erwartungswert μ . Dann

Theorem 1.1.48 (The weak law of large numbers)

Let $X_i, i \in \mathbb{N}$, be a sequence of i.i.d. random variables with mean μ . Then

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mu.$$

Satz 1.1.49 (Das starke Gesetz der großen Zahlen)

Sei $X_i, i \in \mathbb{N}$, eine Folge von u.i.v. Zufallsvariablen mit Erwartungswert μ .
Dann

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{a.s.} \mu.$$

Theorem 1.1.49 (The strong law of large numbers)

Let $X_i, i \in \mathbb{N}$, be a sequence of i.i.d. random variables with mean μ .
Then

Satz 1.1.50 (Der zentrale Grenzwertsatz (ZGS))

Sei $X_i, i \in \mathbb{N}$, eine Folge von u.i.v. Zufallsvariablen mit Erwartungswert μ und Varianz σ^2 .
Dann gilt

$$\frac{\sum_{i=1}^n X_i - n \cdot \mu}{\sqrt{n \cdot \sigma^2}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{L}} Z, n \rightarrow \infty,$$

wobei $Z \sim \mathcal{N}(0, 1)$.

Theorem 1.1.50 (The central limit theorem (CLT))

Let $X_i, i \in \mathbb{N}$, be a sequence of i.i.d. random variables with mean μ and variance σ^2 .
Then it holds

where $Z \sim \mathcal{N}(0, 1)$.

Satz 1.1.51 (Zentraler Grenzwertsatz von Moivre – Laplace)

Für $n \in \mathbb{N}$ sei X_n eine binomialverteilte Zufallsvariable, $X_n \sim \mathcal{B}(n, p)$, mit $0 < p < 1$.
Dann gilt für die Folge X_1, X_2, \dots und jedes $x \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} \mathcal{P} \left(\frac{X_n - np}{\sqrt{np(1-p)}} \leq x \right) = \Phi(x),$$

wobei Φ die Verteilungsfunktion einer normalverteilten Zufallsvariable mit Mittelwert 0 und Varianz 1 ist.

**1.2 Beispiele von Verteilungen
Examples of distributions**

Wir beenden diese Einführung mit Beispielen verschiedener Verteilungen und deren Anwendung.

Beispiel 1.2.1 (Normalverteilung)

Messungen verschiedener Art werden oft als normalverteilt angenommen.
Zum Beispiel können wir eine normalverteilte Zufallsvariable X_i benutzen, um das Gewicht der Schnitzel in der Mensa in Beispiel 1.1.9 auf Seite 8 zu beschreiben.

Theorem 1.1.51 (Central limit theorem of Moivre – Laplace)

For $n \in \mathbb{N}$, let X_n be a binomially distributed random variable, $X_n \sim \mathcal{B}(n, p)$, with $0 < p < 1$.
Then it holds for the sequence X_1, X_2, \dots and every $x \in \mathbb{R}$

where Φ is the distribution function of a normally distributed random variable with mean 0 and variance 1.

We close this introduction with examples of different distributions and their applications.

Example 1.2.1 (Normal distribution)

Measurements of different kinds are often assumed to be normally distributed.
For example, we can use a normally distributed random variable X_i to describe the weight of the steaks in the student's refectory in Example 1.1.9 on page 8.

X_i ist normalverteilt mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$, bezeichnet mit $X \sim \mathcal{N}(\mu, \sigma^2)$.

Die Dichte von X ist

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

Die Verteilungsfunktion ist

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R},$$

wobei Φ die Verteilungsfunktion einer $\mathcal{N}(0, 1)$ -verteilten Zufallsvariablen ist, d.h.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}t^2} dt, \quad -\infty < x < \infty.$$

X hat Erwartungswert $\mathbb{E}(X) = \mu$ und Varianz $\mathcal{V}\mathcal{A}\mathcal{R}(X) = \sigma^2$.

Beispiel 1.2.2 (Poissonverteilung)

Die wandernden MathematikstudentInnen entscheiden sich zum Bau einer Elwedritschefalle und wollen sie irgendwo im Wald aufstellen. Sie nehmen an, dass die Anzahl X_i der gefangenen Elwedritschen am Tag i poissonverteilt mit Parameter λ ist.

Dann ist

$$\mathbb{P}(X = i) = \frac{\lambda^i}{i!} \cdot e^{-\lambda}, \quad i = 0, 1, \dots$$

X is normally distributed with the parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$.

The density function of X is

The distribution function is

where Φ is the distribution function of a $\mathcal{N}(0, 1)$ -distributed random variable, i.e.

X has expectation $\mathbb{E}(X) = \mu$ and variance $\mathcal{V}\mathcal{A}\mathcal{R}(X) = \sigma^2$.

Example 1.2.2 (Poisson distribution)

The hiking mathematics students decide to build an elwedritsche trap and set it out in the forest somewhere. They assume that the number of elwedritches X_i caught at day i in the trap, is poisson distributed with parameter λ .

Then is

X_i hat Erwartungswert $\mathbb{E}(X) = \lambda$ und Varianz $\mathcal{V}\mathcal{A}\mathcal{R}(X) = \lambda$.

Beispiel 1.2.3 (Exponentialverteilung)

Für die Falle wollen die StudentInnen auch einen elektronischen Sensor konstruieren, der ihnen ein Signal sendet, sobald eine Elwedritsche gefangen wurde. Die StudentInnen wissen, dass die Lebenszeit elektronischer Komponenten mit einer Zufallsvariable beschrieben werden können, die exponentialverteilt ist mit Parameter $\lambda > 0$, d.h. $X \sim \text{Exp}(\lambda)$ und die Dichte von X ist

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0, \\ 0 & t < 0. \end{cases}$$

Eine exponentialverteilte Zufallsvariable X hat den Erwartungswert $\mathbb{E}(X) = \frac{1}{\lambda}$ und Varianz $\mathcal{V}\mathcal{A}\mathcal{R}(X) = \frac{1}{\lambda^2}$.

Beispiel 1.2.4 (Multinomialverteilung)

Die StudentInnen haben keine Ahnung, wie eine Elwedritsche aussieht. Deshalb bitten sie eine Biologiestudentin, ihnen mehr über dieses Märchentier zu erzählen. Unter anderem, wird ihnen erklärt, dass die Elwedritsche kurzhaarig oder langhaarig sein kann, und dass der Mund klein oder gross sein kann. Die Eigenschaften „langes Fell“ und „kleiner Mund“ werden dominant und unabhängig voneinander vererbt.

X has expectation $\mathbb{E}(X) = \lambda$ and variance $\mathcal{V}\mathcal{A}\mathcal{R}(X) = \lambda$.

Example 1.2.3 (Exponential distribution)

For the trap, the students also want to construct an electronic sensor, emitting a signal to them as soon as an elwedritsche has been caught. The students know that the life time of electronic components can be described by a random variable, which is exponentially distributed with parameter $\lambda > 0$, i.e. $X \sim \text{Exp}(\lambda)$ and the density function of X is

An exponentially distributed random variable X has the expectation $\mathbb{E}(X) = \frac{1}{\lambda}$ and variance $\mathcal{V}\mathcal{A}\mathcal{R}(X) = \frac{1}{\lambda^2}$.

Example 1.2.4 (Multinomial distribution)

The students have no idea what an elwedritsche looks like. Therefore, they ask a biology student to tell them more about this fairy-tale animal. Among other things, she explains that the elwedritsche can be short-haired or long-haired and that the mouth can be small or big. The properties "long hair" and "small mouth" are dominantly inherited and they are independent of each other.

Dies würde in der zweiten Generation vier Phänotypen im Verhältnis 9:3:3:1 ergeben. Angenommen, dass Phänotyp i mit Wahrscheinlichkeit p_i auftritt, dass Z_i die Anzahl der Elwedritschen mit Phänotyp i , $i = 1, \dots, 4$ ist, und dass $Z_1 + \dots + Z_4 = n$, die Gesamtanzahl aller gefangenen Elwedritsche, ist.

Dann ist $\mathbf{Z} = (Z_1, \dots, Z_4)$ multinomialverteilt mit Parametern (n, p_1, \dots, p_4) .

Im allgemeinen, sei $\mathbf{p} = (p_1, \dots, p_r)$ mit $p_1 > 0, \dots, p_r > 0$ und $p_1 + \dots + p_r = 1$. Dann ist die r -dimensionale Zufallsvariable $\mathbf{Y} = (Y_1, \dots, Y_r)$ multinomialverteilt mit Parametern $n \in \mathbb{N}$ und $\mathbf{p} = (p_1, \dots, p_r)$, falls

$$\mathcal{P}(Y_1 = y_1, \dots, Y_r = y_r) = \begin{cases} \frac{n!}{y_1! \dots y_r!} p_1^{y_1} \dots p_r^{y_r} & y_1 + \dots + y_r = n, \\ 0 & y_1 + \dots + y_r \neq n. \end{cases}$$

Einer der MathematikstudentInnen wird von der Theorie der Genetik so sehr fasziniert dass er zurück bleibt, um mehr zu erfahren. Damit machen sich nur 9 StudentInnen zum Humberturm auf.

This would give four phenotypes with the proportion 9:3:3:1 in the second generation. Assume that phenotype i occurs with probability p_i and that Z_i is the number of elwedritches with phenotype i , $i = 1, \dots, 4$ and that $Z_1 + \dots + Z_4 = n$, the total number of caught elwedritches.

Then $\mathbf{Z} = (Z_1, \dots, Z_4)$ is multinomially distributed with parameters (n, p_1, \dots, p_4) .

In general, let $\mathbf{p} = (p_1, \dots, p_r)$ with $p_1 > 0, \dots, p_r > 0$ and $p_1 + \dots + p_r = 1$. Then the r -dimensional random variable $\mathbf{Y} = (Y_1, \dots, Y_r)$ is multinomially distributed with parameters $n \in \mathbb{N}$ and $\mathbf{p} = (p_1, \dots, p_r)$ if

One of the mathematics students becomes so fascinated by the theory of genetics, that he decides to stay and learn more. So, now they are only 9 students leaving for the Humbert tower.

Beispiel 1.2.5 (Gleichverteilung)

Bevor sie für die Humberturmwanderung aufbrechen, flanieren sie übers Universitätsgelände, weil es der Tag des traditionellen AStA-Sommerfests ist. Deswegen steht auch ein kleines Karussell hinter dem Mathematikgebäude.

Eine der StudentInnen möchte zeigen, was eine Gleichverteilung ist. Als sie das Karussell entdeckt, bekommt sie die Idee, wie man eine Gleichverteilung erhält.

Sie setzt sich ins Karussell, das sich mit konstanter Geschwindigkeit dreht, und isst Kirschen. Die Kerne spuckt sie ohne nachzudenken heraus. Der Winkel X zwischen einer Referenzrichtung und einem Kirschstein ist gleichverteilt auf $[0^\circ, 360^\circ]$.

Allgemein bezeichnet man X als gleichverteilt mit Parametern $-\infty < a < b < \infty$, $X \sim \mathcal{U}[a, b]$, und der Dichte

Example 1.2.5 (Uniform distribution)

Right before leaving for the Humbert tower trip, the students stroll over the campus, since it is the day of the traditional AStA summer party. Therefore, there is also a small merry-go-round behind the mathematics building.

One of the students wants to demonstrate what a uniform distribution is. As she discovers the roundabout, she gets an idea how to obtain a uniform distribution.

So, riding on the roundabout, which is moving with constant velocity, she eats cherries. She spits the cherry stones out without thinking. The angle X between a reference direction and a cherry stone is uniformly distributed on $[0^\circ; 360^\circ]$.

In general, X is uniformly distributed with parameters $-\infty < a < b < \infty$, $X \sim \mathcal{U}[a, b]$, which has the density function

$$f(t) = \begin{cases} \frac{1}{b-a} & a \leq t \leq b, \\ 0 & t \notin [a, b], \end{cases}$$

was zum Erwartungswert $\mathbb{E}(X) = \frac{a+b}{2}$ und der Varianz $\mathcal{V}\mathcal{A}\mathcal{R}(X) = \frac{(b-a)^2}{12}$ führt.

Die Studentin ist so vernarrt ins Karussellfahren, dass auch sie die Gruppe nicht weiter begleiten will, und so ziehen 8 StudentInnen weiter.

leading to the expectation $\mathbb{E}(X) = \frac{a+b}{2}$ and variance $\mathcal{V}\mathcal{A}\mathcal{R}(X) = \frac{(b-a)^2}{12}$.

The student is so crazy in riding the merry-go-round, so she does not want to accompany the group anymore, and 8 students are going on.

Beispiel 1.2.6 (Cauchy-Verteilung)

Neben dem Karussell steht ein Zelt. Auf dessen Außenwand sieht man nun die Spuren der dort aufgetroffenen Kirschsteine. Leider, so bemerkt die karussellfahrende Studentin, erstreckt sich diese Außenwand nicht „ins Unendliche“, denn dann wäre die Verteilung der Flecken Cauchy-verteilt, wenn sie ebenfalls unendlich viele Kirschen gegessen hätte.

Example 1.2.6 (Cauchy-distribution)

Next to the merry-go-round is a tent. On the outside, one can now see the result of the hitting stones. Unfortunately, remarks the merry-go-round-riding student, the outside does not extend “to infinity”, since in that case, the distribution of the spots would be Cauchy-distributed, if she also ate infinitely many cherries.

Bezeichnet X eine (standard)-Cauchy-verteilte Zufallsvariable, so hat diese die Dichte

$$f(x) = \frac{1}{\pi \cdot (1+x^2)},$$

aber der Erwartungswert existiert nicht (und damit auch keine höheren Momente).

Let X denote a (standard)-Cauchy-distributed random variable, which has the density

but the expectation does not exist (and therefore also no higher moments).

Beispiel 1.2.7 (Geometrische Verteilung)

Als die StudentInnen losgehen, kommt einem plötzlich, dass sein schwedischer Mitbewohner an diesem Abend seine Probleme haben wird, den richtigen Schlüssel für die gemeinsame Wohnung zu finden.

Wenn er betrunken ist, so probiert er immer wieder einen zufällig gewählten Schlüssel aus, denn er merkt sich nicht, welche er schon probiert hat, d.h. er könnte einige Schlüssel mehr als einmal probieren.

Sei X die Anzahl der Versuche, die der Student braucht, um die Tür zu öffnen.

X ist geometrisch verteilt mit Parameter p ($0 < p < 1$) und

$$\mathcal{P}(X = i) = p \cdot (1-p)^{i-1}, \quad i = 1, 2, \dots$$

X hat den Erwartungswert $\mathbb{E}(X) = \frac{1}{p}$ und Varianz $\mathcal{V}\mathcal{A}\mathcal{R}(X) = \frac{1-p}{p^2}$.

Weil es der Tag des AStA-Sommerfests ist, ist die Wahrscheinlichkeit hoch, dass der schwedische Student betrunken ist. Deswegen entscheidet sich der nette Mathematics International-Student die Wandergruppe zu verlassen und kehrt zurück zu seiner Wohnung um seinem Mitbewohner die Tür zu öffnen.

Example 1.2.7 (Geometric distribution)

As the students start their hiking, one of them suddenly recalls that his Swedish room mate probably will have trouble to find the right key for their common flat this evening.

When he is drunk, he tries one randomly chosen key, since he does not recognize which he has already tried, i.e. he might try some keys more than once.

Let X be the number of trials the student needs to open the door.

X follows a geometric distribution with the parameter p ($0 < p < 1$) and

X has the expectation $\mathbb{E}(X) = \frac{1}{p}$ and variance $\mathcal{V}\mathcal{A}\mathcal{R}(X) = \frac{1-p}{p^2}$.

Since it is the AStA-summer party day, the probability is high that the Swedish student is drunk. Therefore the kind mathematics international student decides to leave the hiking group and returns to his flat to open the door for his room mate.

1.3 Zufallsvektoren Random vectors

Ein Zufallsvektor, d.h. eine Zufallsvariable mit Werten in \mathbb{R}^d , $d \geq 2$, hat als Verteilung ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}^d, \mathcal{L}^d)$, wobei \mathcal{L}^d die Borel- σ -Algebra auf \mathbb{R}^d ist. Es gilt wie im eindimensionalen Fall

$$\mathcal{P}(X \in B) = P(B), \quad B \in \mathcal{L}^d.$$

Wenn P eine Dichte $f(x), x \in \mathbb{R}^d$, hat, so gilt wieder

$$\mathcal{P}(X \in B) = \int_B \cdots \int_B f(x_1, \dots, x_d) dx_1 \dots dx_d.$$

Der Erwartungswert des Zufallsvektors X ist wieder als Integral definiert, kann aber koordinatenweise berechnet werden.

Definition 1.3.1 (Erwartungswert) Sei $X = (X_1, \dots, X_d)^T$ eine Zufallsvariable in \mathbb{R}^d mit Verteilung P .

$$\mathcal{E}X = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} x dP(x) = (\mathcal{E}X_1, \dots, \mathcal{E}X_d)^T$$

Definition 1.3.2 (Kovarianzmatrix) Sei $X = (X_1, \dots, X_d)^T$ ein Zufallsvektor in \mathbb{R}^d . Die $x \times d$ -Kovarianzmatrix $\text{cov}(X)$ ist

$$\text{cov}(X) = (\text{cov}(X_i, X_j))_{i,j=1,\dots,d}.$$

A random vector, i.e. a random variable with values in \mathbb{R}^d , $d \geq 2$, has as distribution a probability measure P on $(\mathbb{R}^d, \mathcal{L}^d)$, where \mathcal{L}^d denotes the Borel- σ -Algebra on \mathbb{R}^d . As in the univariate case we have

If P has a density $f(x), x \in \mathbb{R}^d$, we have again

The expectation of a random vector X is defined as an integral again, but it can be evaluated coordinatewise.

Definition 1.3.1 (Expectation) Let $X = (X_1, \dots, X_d)^T$ be a random variable in \mathbb{R}^d with law P .

Definition 1.3.2 (Covariance matrix) Let $X = (X_1, \dots, X_d)^T$ be a random vector in \mathbb{R}^d . The $x \times d$ -covariance matrix $\text{cov}(X)$ is

Wir brauchen später oft, wie sich Kovarianzmatrizen unter affin-linearen Abbildungen ändern.

Bemerkung 1.3.3 Sei $X \in \mathbb{R}^d$ ein Zufallsvektor mit Kovarianzmatrix $\Sigma = \text{cov}(X)$

$$\Sigma = \text{cov}(X) = \mathcal{E}[(X - \mathcal{E}X)(X - \mathcal{E}X)^T]$$

Satz 1.3.4 Sei A eine $m \times d$ -Matrix, $b \in \mathbb{R}^m$. Betrachte den Zufallsvektor $Y \in \mathbb{R}^m$ gegeben durch

$$Y = AX + b.$$

Dann gilt

$$\begin{aligned} \mathcal{E}Y &= A\mathcal{E}X + b \\ \text{cov}(Y) &= A\Sigma A^T \end{aligned}$$

Bemerkung 1.3.5 (Definitheit von Kovarianzmatrizen) Sei $X \in \mathbb{R}^d$ ein Zufallsvektor mit Kovarianzmatrix $\Sigma = \text{cov}(X)$. Dann ist $\Sigma \geq 0$ (positiv semidefinit), d.h. $\alpha^T \Sigma \alpha \geq 0$ für alle $\alpha \in \mathbb{R}^d$.

Lateron, we need how covariance matrix change under affine-linear transformations.

Remark 1.3.3 Let $X \in \mathbb{R}^d$ be a random vector with covariance matrix $\Sigma = \text{cov}(X)$

Theorem 1.3.4 Let A be a $m \times d$ -matrix, $b \in \mathbb{R}^m$. Consider the random vector $Y \in \mathbb{R}^m$ given as

Remark 1.3.5 (Definiteness of covariance matrices) Let $X \in \mathbb{R}^d$ be a random vector with covariance matrix $\Sigma = \text{cov}(X)$. Then, $\Sigma \geq 0$ (positive semidefinite), i.e. $\alpha^T \Sigma \alpha \geq 0$ for all $\alpha \in \mathbb{R}^d$.

Definition 1.3.6 (Multivariate

Normalverteilung) Für Zufallsvektor $X \in \mathbb{R}^d$ heißt multivariat normalverteilt mit Mittelwert $\mu \in \mathbb{R}^d$ und Kovarianzmatrix $\Sigma > 0$ (positiv definit):

$$\mathcal{L}(X) = \mathcal{N}_d(\mu, \Sigma)$$

wenn es die Wahrscheinlichkeitsdichte hat:

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, x \in \mathbb{R}^d.$$

Es gilt

$$\mathbb{E}X = \mu, \text{cov}(X) = \Sigma.$$

Die Koordinaten X_1, \dots, X_d heißen **gemeinsam** normalverteilt.

Bemerkung 1.3.7 a) X_1, X_2 seien gemeinsam normalverteilt, d.h.

$$\mathcal{L}((X_1, X_2)^T) = \mathcal{N}_2(\mu, \Sigma).$$

Dann sind X_1, X_2 genau dann unabhängig, wenn sie unkorreliert sind.

b) Sei

$$\mathcal{L}(X) = \mathcal{N}_d(\mu, \Sigma), Y = AX + b \in \mathbb{R}^m.$$

Definition 1.3.6 (multivariate normal

distribution) A random vector $X \in \mathbb{R}^d$ has a multivariate normal distribution with mean $\mu \in \mathbb{R}^d$ and covarianz matrix $\Sigma > 0$ (positive definite):

if it has the probability density

We have

The coordinates X_1, \dots, X_d are called **jointly** normally distributed.

Remark 1.3.7 a) Let X_1, X_2 be jointly normally distributed, i.e.

Then, X_1, X_2 are independent iff they are uncorrelated.

b) Let

Dann ist

d.h. die Normalverteilung ist invariant unter affin-linearen Abbildungen.

c) Sei

Dann existieren

so daß

Then,

$$\mathcal{L}(Y) = \mathcal{N}_m(A\mu + b, A\Sigma A^T),$$

i.e. the normal law is invariant under affine-linear transformations.

c) Let

$$\mathcal{L}(X) = \mathcal{N}_d(\mu, \Sigma), \Sigma > 0.$$

Then, there are

$$Z_1, \dots, Z_d \text{ i.i.d.}, \mathcal{L}(Z_j) = \mathcal{N}(0, 1)$$

such that

$$X = \Sigma^{1/2}Z + \mu, Z = (Z_1, \dots, Z_d)^T.$$

2 Punktschätzung

Point estimation

In diesem Abschnitt werden wir parametrische Schätzverfahren, genauer Punktschätzungen untersuchen. Wir gehen von folgender Situation aus:

Sei \mathbf{X} eine Zufallsvariable mit Werten im Raum \mathbf{X} und die Verteilung

$$\mathcal{L}(\mathbf{X}) \in \{\mathcal{P}_\vartheta; \vartheta \in \Theta\}$$

soll von keinen anderen unbekanntem Parametern abhängen als ϑ . Unser Schätzproblem ist jetzt, eine Funktion g von ϑ , $g(\vartheta)$, aus einer Realisierung \mathbf{x} von \mathbf{X} , zu schätzen.

Im Folgenden sei $\Theta \subseteq \mathbb{R}^d$ für ein $d \geq 1$ und $\mathbf{X} = (X_1, \dots, X_n)$, wobei X_1, \dots, X_n unabhängige Zufallsvariablen sind, d.h. die Stichprobengröße n ist.

Zuerst führen wir einige Definitionen ein:

Definition 2.0.8 (Statistik)

Eine Statistik \mathbf{T} ist eine Zufallsvariable, die eine Funktion der Beobachtungen ist, aber nicht vom unbekanntem Parameter ϑ abhängt.

In this section, we are going to study parametric estimation, more precisely point estimation. We are faced with the following situation:

Let \mathbf{X} be a random variable with values in the space \mathbf{X} and assume that the distribution

does not depend on any other unknown parameters than ϑ . Our estimation problem is now to use a realization \mathbf{x} of \mathbf{X} to estimate a function of ϑ , $g(\vartheta)$.

In the following, let $\Theta \subseteq \mathbb{R}^d$ for a $d \geq 1$ and $\mathbf{X} = (X_1, \dots, X_n)$, where X_1, \dots, X_n are independent random variables, i.e. the sample size is n .

First, we introduce some definitions:

Definition 2.0.8 (Statistic)

A statistic \mathbf{T} is a random variable, which is a function of the observations, but does not depend on the unknown parameter ϑ .

Definition 2.0.9 (Schätzer)

Wenn eine Statistik $\mathbf{T}(X_1, \dots, X_n)$ zum Schätzen von $g(\vartheta)$ benutzt wird, so wird sie Schätzer genannt.

Zur Verdeutlichung, dass $g(\vartheta)$ aus n Daten geschätzt wird, führen wir die Notation

$$\widehat{g(\vartheta)}_n := \mathbf{T}(X_1, \dots, X_n)$$

bzw. $\hat{\vartheta}_n$, wenn wir nur ϑ schätzen, ein.

Definition 2.0.10 (Schätzwert)

Der spezifische Wert $\mathbf{T}(x_1, \dots, x_n)$, für die Beobachtungen x_1, \dots, x_n , wird Schätzwert genannt.

Beispiel 2.0.11 (Schätzer)

$\hat{\mu}_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ ist ein Schätzer (Zufallsvariable) für den Mittelwert.

Beispiel 2.0.12 (Schätzwert)

$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ ist ein Schätzwert, d.h. auf Daten basierender fester Wert.

Definition 2.0.9 (Estimator)

If a statistic $\mathbf{T}(X_1, \dots, X_n)$ is used for the estimation of $g(\vartheta)$, then it is called an estimator.

For clarification that $g(\vartheta)$ is estimated out of n data, we introduce the notation

$$\widehat{g(\vartheta)}_n := \mathbf{T}(X_1, \dots, X_n)$$

or $\hat{\vartheta}_n$ if we only estimate ϑ .

Definition 2.0.10 (Estimate)

The specific value $\mathbf{T}(x_1, \dots, x_n)$, taken for some observations x_1, \dots, x_n , is called an estimate.

Example 2.0.11 (Estimator)

$\hat{\mu}_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ is an estimator (random variable) for the mean.

Example 2.0.12 (Estimate)

$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ is an estimate, i.e. a fixed value based on data.

Bemerkung 2.0.13

Weil die Verteilung von \mathbf{X} vom unbekanntem Parameter ϑ abhängt, betonen wir diese Abhängigkeit und schreiben z.B. $\mathcal{P}_\vartheta(X \in B)$, $\mathcal{E}_\vartheta(h(X))$ für Wahrscheinlichkeiten und Erwartungswerte, wenn sie unter der Voraussetzung, dass ϑ der wahre Parameter der Verteilung von \mathbf{X} ist, berechnet werden.

Remark 2.0.13

Since the distribution of \mathbf{X} depends on the unknown parameter ϑ , we stress this dependence by writing e.g. $\mathcal{P}_\vartheta(X \in B)$, $\mathcal{E}_\vartheta(h(X))$ for probabilities and expectations, when they are calculated under the assumption that ϑ is the true parameter of the distribution of \mathbf{X} .

2.1 Wünschenswerte Eigenschaften von Schätzern
Desirable properties of estimators

Wie zuverlässig ist unser Schätzwert? Können wir ihm trauen?

Es gibt etliche Methoden, um unterschiedliche Schätzer zu vergleichen. Hier untersuchen wir die Qualität der Schätzer und stellen fest, welche die wünschenswerten Eigenschaften sind.

How reliable is our estimate? Can we trust it?

There are several methods to compare different estimators. Here we discuss the quality of estimators and find out which the desirable properties are.

2.1.1 Konsistenz, Verlust und Risiko
Consistency, loss and risk

Es ist vernünftig, dass die Schätzmethode einen Schätzwert geben soll, der sich verbessert, wenn die Stichprobengröße größer wird. Deswegen verlangen wir mindestens, dass der Schätzer gegen den zu schätzenden Wert konvergiert, falls die Stichprobengröße n gegen unendlich geht. Dies führt zum Begriff der Konsistenz.

It is reasonable that the estimation method shall give an estimate that improves as the sample size increases. Therefore, we at least require that the estimator converges to the value to be estimated as the sample size n tends to infinity. This leads to the concept of consistency.

Definition 2.1.1 (Konsistenz)

Eine Folge $T_n = T_n(X_1, \dots, X_n)$, $n \in \mathbb{N}$, von Schätzern für $g(\vartheta)$ ist konsistent, falls für alle $\vartheta \in \Theta$

$$\mathcal{P}_\vartheta \left(\|T_n(X_1, \dots, X_n) - g(\vartheta)\| > \varepsilon \right) \xrightarrow{n \rightarrow \infty} 0 \quad \forall \varepsilon > 0,$$

d.h. $T_n(X_1, \dots, X_n) \xrightarrow{P} g(\vartheta)$, falls $\mathcal{L}(X_1, \dots, X_n) = \mathcal{P}_{\vartheta, n}$ für alle $n \geq 1$.

Auch wenn wir verlangen, dass der Schätzer konsistent ist, gibt es immer noch viele Schätzer zur Auswahl.

Eine andere Möglichkeit ist die sogenannte Verlustfunktion $L(t, g(\vartheta))$ zu studieren. Die Verlustfunktion misst, wie groß die Abweichung zwischen dem Schätzwert t und dem wahren Wert $g(\vartheta)$ ist.

Einige Beispiele von Verlustfunktionen sind die Norm

$$L(t, g(\vartheta)) = \|t - g(\vartheta)\|$$

oder die quadratische Norm

$$L(t, g(\vartheta)) = \|t - g(\vartheta)\|^2$$

oder die ε -unempfindliche Verlustfunktion

$$L_\varepsilon(t, g(\vartheta)) = \begin{cases} 0, & \|t - g(\vartheta)\| \leq \varepsilon \\ \|t - g(\vartheta)\| - \varepsilon, & \|t - g(\vartheta)\| > \varepsilon, \end{cases}$$

Definition 2.1.1 (Consistency)

A sequence $T_n = T_n(X_1, \dots, X_n)$, $n \in \mathbb{N}$ of estimators for $g(\vartheta)$ is called consistent, if for all $\vartheta \in \Theta$

i.e. $T_n(X_1, \dots, X_n) \xrightarrow{P} g(\vartheta)$, if $\mathcal{L}(X_1, \dots, X_n) = \mathcal{P}_{\vartheta, n}$ for all $n \geq 1$.

However, even when we require that the estimator is consistent, there are still lots of estimators to choose between.

Another possibility is to study the so-called loss function $L(t, g(\vartheta))$. The loss function measures how big the difference between the estimated value t and the true value $g(\vartheta)$ is.

Some examples of loss functions are the norm

or the quadratic norm

or the ε -insensitive loss function

wobei für die letzte Verlustfunktion nur Abweichungen zwischen t und $g(\vartheta)$ größer als ϵ als signifikant angesehen werden. Wegen der Möglichkeit die quadratische Norm bzgl. t abzuleiten, wird diese Funktion sehr oft als Verlustfunktion herangezogen.

Bemerkung 2.1.2

Zu beachten ist, dass bei Verwendung des Schätzers T_n für t in der Verlustfunktion sich eine Zufallsvariable ergibt, d.h. man hat immer noch keine konkrete Zahl für die Güte des Schätzers T_n !

Ein Maß für die Qualität des Schätzers ist der Erwartungswert der Verlustfunktion:

Definition 2.1.3 (Risiko)

Das Risiko des Schätzers T_n für $g(\vartheta)$ ist

$$R(T_n, g(\vartheta)) = \mathcal{E}_{\vartheta} [L(T_n, g(\vartheta))].$$

Bemerkung 2.1.4

Das Risiko $R(T_n, \cdot)$ als Funktion von ϑ wird Risikofunktion oder Operationscharakteristik genannt.

Ist der Schätzer T_n Cauchy-verteilt, so existiert der Erwartungswert nicht (und somit nicht die Varianz) und damit kann das Risiko für keine der genannten Verlustfunktionen berechnet werden, d.h. wäre stets unendlich.

where for the last loss function deviations between t and $g(\vartheta)$ larger than ϵ are considered to be significant.

Because of the differentiability of the quadratic norm with respect to t , this function is often preferred as loss function.

Remark 2.1.2

Notice that by inserting the estimator T_n for t into the loss function, we get a random variable, i.e. we still do not have a fixed number for the quality of the estimator T_n !

One measure of the quality of the estimator is the expectation of the loss function:

Definition 2.1.3 (Risk)

The risk of the estimator T_n for $g(\vartheta)$ is

Remark 2.1.4

The risk $R(T_n, \cdot)$ as a function of ϑ is called the risk function or the operation characteristic.

If the estimator T_n is Cauchy-distributed, then the expectation does not exist (and therefore not the variance) and this implies that the risk cannot be calculated for any of the presented loss functions, i.e. would always be infinity.

Satz 2.1.5 ($R \rightarrow 0 \Rightarrow$ Konsistenz)

Sei $T_i, i \in \mathbb{N}$, eine Folge von Schätzern für $g(\vartheta)$. Wir betrachten als Verlustfunktion die Norm oder die quadratische Norm. Falls das Risiko $R(T_k, g(\vartheta))$ für jedes $\vartheta \in \Theta$ gegen Null konvergiert, dann ist die Folge von Schätzern konsistent.

Beweis:

Sei $p = 1$ im Falle der Norm und $p = 2$ im Falle der quadratischen Norm als Verlustfunktion.

$$R(T_k, g(\vartheta)) \xrightarrow{k \rightarrow \infty} 0 \Rightarrow T_k - g(\vartheta) \xrightarrow{L^p} 0 \Rightarrow T_k - g(\vartheta) \xrightarrow{p} 0,$$

was gleichbedeutend zur Konsistenz ist. ■

Bemerkung 2.1.6

Wichtig zu bemerken ist, dass das Risiko nicht existieren muss (z.B. falls T_k Cauchy-verteilt ist), aber die Folge der Schätzer trotzdem konsistent sein kann.

Bemerkung 2.1.7

Im Folgenden benutzen wir die Funktion $L(t, g(\vartheta)) = \|t - g(\vartheta)\|^2$, d.h. das Risiko

$$R(T_n, g(\vartheta)) = \mathcal{E}_{\vartheta} \|T_n - f(\vartheta)\|^2.$$

Theorem 2.1.5 ($R \rightarrow 0 \Rightarrow$ Consistency)

Let $T_i, i \in \mathbb{N}$, be a sequence of estimators of $g(\vartheta)$. Let us consider the norm or the quadratic norm for the loss function. If the quadratic risk $R(T_k, g(\vartheta))$ converges towards zero for all $\vartheta \in \Theta$, then the sequence of estimators is consistent.

Proof:

Let $p = 1$ in the case of the norm and $p = 2$ in the case of the quadratic norm for the loss function.

which is equivalent to consistency. ■

Remark 2.1.6

It is important to notice that the risk does not need to exist (e.g. if T_k is Cauchy-distributed), but the sequence of estimators can still be consistent.

Remark 2.1.7

In the following, we use the function $L(t, g(\vartheta)) = \|t - g(\vartheta)\|^2$, i.e. the risk

Bemerkung 2.1.8

Wir hätten natürlich gern einen Schätzer, der ein kleines Risiko für so viele ϑ wie möglich hat. Es ist allerdings nicht möglich den universell besten Schätzer T_n^* zu finden, d.h. der

Remark 2.1.8

We would of course like to have an estimator, which has a small risk for as many values of ϑ as possible. It is, however, not possible to find the uniformly best estimator T_n^* fulfilling

$$R(T_n^*, g(\vartheta)) = \min_{T_n} R(T_n, g(\vartheta)) \quad \forall \vartheta \in \Theta$$

erfüllt, weil der triviale Schätzer

because the trivial estimator

$$T_n = g(\vartheta_0)$$

für ein $\vartheta_0 \in \Theta$ zum Risiko Null bei ϑ_0 führen würde. Da ϑ_0 beliebig war, müsste T_n^* für alle $\vartheta \in \Theta$ zum Risiko Null führen.

for some $\vartheta_0 \in \Theta$, would lead to the risk zero at ϑ_0 . Since ϑ_0 was arbitrary, T_n^* would have risk zero for all $\vartheta \in \Theta$.

Allerdings gibt es diesen universell besten Schätzer schon, falls $\Theta = \{\vartheta_0\}$, d.h. nur ein Element hat, da wir dann mit $T_n^* := \vartheta_0$ solch einen Schätzer haben.

However, there exists the universally best estimator, if $\Theta = \{\vartheta_0\}$, i.e. consists of only one element, since we obtain with $T_n^* := \vartheta_0$ such an estimator.

Diese Situation erscheint aber vollkommen witzlos, da wir von Anfang an den Wert des Parameters wissen!

This situation seems to be totally reasonless, because we know from the beginning the value of the parameter!

2.1.2 Unverzerrtheit Unbiasedness

Eine andere Eigenschaft ist die Unverzerrtheit:

Another property is unbiasedness:

Definition 2.1.9 (Unverzerrtheit)

Der Schätzer T ist unverzerrt (biasfrei), falls

Definition 2.1.9 (Unbiasedness)

The estimator T is unbiased if

$$\mathcal{E}_\vartheta(T) = \vartheta \quad \forall \vartheta \in \Theta.$$

Bemerkung 2.1.10 (Bias)

$\mathcal{E}_\vartheta(T) - \vartheta$ wird der Bias des Schätzers T genannt. Im Mittel schätzt ein unverzerrter Schätzer den richtigen Parameterwert, d.h. der Schätzer ist richtig zentriert.

Remark 2.1.10 (Bias)

$\mathcal{E}_\vartheta(T) - \vartheta$ is called the bias of the estimator T . On the average, an unbiased estimator estimates the correct parameter value, i.e. the estimator is correctly centered.

Beispiel 2.1.11

Seien X_1, \dots, X_n Zufallsvariablen mit Erwartungswert $\mathcal{E}(X_i) = \mu$. Dann ist der arithmetische Mittelwert

Example 2.1.11

Let X_1, \dots, X_n be random variables with expectation $\mathcal{E}(X_i) = \mu$. Then the arithmetic mean

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

ein unverzerrter Schätzer für den Erwartungswert μ :

is an unbiased estimator for the expectation μ :

$$\mathcal{E}(\bar{X}_n) = \mathcal{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{n}{n} \mathcal{E}(X_1) = \mu.$$

Beispiel 2.1.12

Seien X_1, \dots, X_n unabhängige Zufallsvariablen und es existiere der Erwartungswert $\mathbb{E}(X_i) = \mu$ und die Varianz $\mathcal{V}\mathcal{A}\mathcal{R}(X_i) = \sigma^2 < \infty$.

Dann ist die Stichprobenvarianz

$$\hat{s}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

ein unverzerrter Schätzer für die Varianz von X_i .

Beweis:

$$\begin{aligned} \mathbb{E}[(X_i - \bar{X}_n)^2] &= \mathbb{E}[(X_i - \mu)^2] - 2 \cdot \mathbb{E}[(X_i - \mu) \cdot (\bar{X}_n - \mu)] + \mathbb{E}[(\bar{X}_n - \mu)^2] \\ &= \sigma^2 - 2 \cdot \frac{1}{n} \sum_{j=1}^n \mathbb{E}[(X_i - \mu) \cdot (X_j - \mu)] + \frac{1}{n^2} \sum_{j,l=1}^n \mathbb{E}[(X_j - \mu) \cdot (X_l - \mu)] \\ &= \sigma^2 - \frac{2}{n} \cdot \mathbb{E}[(X_i - \mu) \cdot (X_i - \mu)] + \frac{1}{n^2} \sum_{j=1}^n \mathbb{E}[(X_j - \mu) \cdot (X_j - \mu)] \\ &= \frac{n-1}{n} \cdot \sigma^2 \end{aligned}$$

weil wegen der Unabhängigkeit

$$\mathbb{E}[(X_j - \mu) \cdot (X_l - \mu)] = \begin{cases} 0, & j \neq l \\ \sigma^2, & j = l. \end{cases}$$

Example 2.1.12

Let X_1, \dots, X_n be independent random variables and let the expectation $\mathbb{E}(X_i) = \mu$ and the variance $\mathcal{V}\mathcal{A}\mathcal{R}(X_i) = \sigma^2 < \infty$ exist.

Then the sample variance

is an unbiased estimator for the variance of X_i .

Proof:

2.1.3 Quadratisches Risiko in Beziehung zu Bias und Varianz
Quadratic risk in connection to bias and variance

Obwohl wir sagen, dass es wünschenswert ist, einen unverzerrten Schätzer zu benutzen, ist dies nicht die ganze Wahrheit: Es gibt Fälle, wo es vorteilhaft ist, einen verzerrten Schätzer zu benutzen, wie unten erklärt.

Satz 2.1.13

Der mittlere quadratische Fehler des Schätzers T für $\vartheta \in \Theta \subseteq \mathbb{R}^1$, $R(T, \vartheta) = \mathbb{E}_\vartheta [(T - \vartheta)^2]$, kann geschrieben werden als

$$R(T, \vartheta) = [\text{Bias}_\vartheta(T)]^2 + \mathcal{V}\mathcal{A}\mathcal{R}_\vartheta(T).$$

Beweis:

$$\begin{aligned} R(T, \vartheta) &= \mathbb{E}_\vartheta [(T - \vartheta)^2] \\ &= \mathbb{E}_\vartheta \left\{ [T - \mathbb{E}_\vartheta(T)]^2 \right\} + \mathbb{E}_\vartheta \{ [T - \mathbb{E}_\vartheta(T)] \cdot [\mathbb{E}_\vartheta(T) - \vartheta] \} + [\mathbb{E}_\vartheta(T) - \vartheta]^2 \\ &= \mathcal{V}\mathcal{A}\mathcal{R}_\vartheta(T) + [\text{Bias}_\vartheta(T)]^2. \end{aligned}$$

Although we said that it is desirable to use an unbiased estimator, this is not the whole truth: There are cases where it is advantageous to use a biased estimator, which is explained below.

Theorem 2.1.13

The mean squared error of the estimator T for $\vartheta \in \Theta \subseteq \mathbb{R}^1$, $R(T, \vartheta) = \mathbb{E}_\vartheta [(T - \vartheta)^2]$, can be expressed as

Proof:

Unverzerrtheit zu verlangen bedeutet somit, dass das quadratische Risiko nur noch gleich der Varianz des Schätzers ist.

Um also die Konsistenz einer Folge von Schätzern zu zeigen, braucht nur noch die Konvergenz der Varianzen gegen Null gezeigt zu werden.

Allerdings kann die Varianz des Schätzers immer noch so groß sein, dass das Risiko für den unverzerrten Schätzer größer ist, als das Risiko eines verzerrten Schätzers mit kleiner Varianz.

Beispiel 2.1.14

Als nächstes betrachten wir die Verteilungen der beiden Schätzer in Abbildung 2.1.1. Wir nehmen an, dass ϑ_0 der zu schätzende Wert ist. Wir sehen, dass $\hat{\vartheta}_1$ verzerrt ist und $\hat{\vartheta}_2$ unverzerrt ist. Wegen der viel kleineren Varianz von $\hat{\vartheta}_1$, ist der mittlere quadratische Fehler kleiner als für $\hat{\vartheta}_2$, obwohl $\hat{\vartheta}_2$ unverzerrt ist.

Oft wird deswegen nicht verlangt, dass der Schätzer unverzerrt ist, sondern nur asymptotisch unverzerrt. Ansonsten würde das Risiko nicht gegen Null konvergieren und man müsste mit anderen Methoden die Konsistenz überprüfen.

Thus, to require unbiasedness means that the quadratic risk reduces to the variance of the estimator.

Thus, for showing the consistency of a sequence of estimators, it suffices to show the convergence of the variances towards zero.

However, the variance can still be so big that the risk is greater for the unbiased estimator, than it would be for a biased estimator with a smaller variance.

Example 2.1.14

Next we consider the distributions of the two estimators in Figure 2.1.1. Assume that ϑ_0 is the true value to be estimated. We note that $\hat{\vartheta}_1$ is biased and $\hat{\vartheta}_2$ is unbiased. Due to the much smaller variance of $\hat{\vartheta}_1$, the mean squared error is smaller than for $\hat{\vartheta}_2$, even though $\hat{\vartheta}_1$ is biased.

Frequently, it is therefore not required that the estimator is unbiased but asymptotically unbiased. Otherwise, the risk could not tend to zero and we would have to check the consistency by other methods.

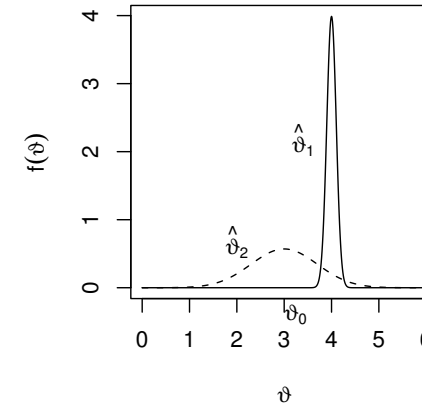


Figure 2.1.1:

Definition 2.1.15 (Asymptotische Unverzerrtheit)

Der Schätzer $T(X)$ ist asymptotisch unverzerrt, wenn

$$\lim_{k \rightarrow \infty} \mathcal{E}_{\vartheta}(T_k) = \vartheta \quad \forall \vartheta \in \Theta.$$

Definition 2.1.15 (Asymptotical unbiasedness)

The estimator $T(X)$ is asymptotically unbiased if

**2.1.4 Suffizienz und Vollständigkeit
Sufficiency and Completeness**

Zwei Eigenschaften, die den Informationsgehalt des Schätzers behandeln, sind Suffizienz und Vollständigkeit.

Two properties concerning the information content of the estimator are sufficiency and completeness.

Definition 2.1.16 (Suffizienz)

Eine Statistik S ist suffizient für ϑ , wenn

$$\mathcal{P}_\vartheta(X \in B | S(X) = t)$$

unabhängig von ϑ ist für alle Borel-Mengen

$B \subseteq \mathbb{R}^N$ und für alle $t \in S(\mathbb{R}^N)$.

Wenn $\mathcal{P}_\vartheta(S(X) = t) = 0$, setzen wir

$\mathcal{P}_\vartheta(X \in B | S(X) = t) = 0$.

Definition 2.1.16 (Sufficiency)

A statistic S is called sufficient for ϑ , if

is independent of ϑ for all Borel-sets $B \subseteq \mathbb{R}^N$

and for all $t \in S(\mathbb{R}^N)$.

If $\mathcal{P}_\vartheta(S(X) = t) = 0$, then we set

$\mathcal{P}_\vartheta(X \in B | S(X) = t) = 0$.

Bemerkung 2.1.17 (Bedeutung der Suffizienz; auch für Schätzer)

Wenn eine Statistik S suffizient ist, dann enthält die bedingte Verteilung von X , bedingt auf

$S(X) = t$, keine Information mehr bezüglich ϑ .

Alle vorhandene Information, die in X steckt, ist

somit in $S(X)$ enthalten. Somit genügt es,

Schätzer T für ϑ zu betrachten, die folgende

Form haben:

$$T(X) = g[S(X)].$$

Remark 2.1.17 (Meaning of sufficiency; also for estimators)

If a statistic S is sufficient, then the conditional distribution of X , conditioned on $S(X) = t$, does

not contain any further information about ϑ .

All available information, held in X , is already

contained in $S(X)$. Therefore, it suffices to

consider estimators T for ϑ of the following

form:

Beispiel 2.1.18

Seien X_1 und X_2 unabhängig $\mathcal{B}(n, \vartheta)$ -verteilt.

Wir betrachten die Statistik

$S(X_1, X_2) = X_1 + X_2$:

$$\begin{aligned} \mathcal{P}(X_1 = x | X_1 + X_2 = t) &= \frac{\mathcal{P}(X_1 = x, X_1 + X_2 = t)}{\mathcal{P}(X_1 + X_2 = t)} \\ &= \frac{\mathcal{P}(X_1 = x, X_2 = t - x)}{\mathcal{P}(X_1 + X_2 = t)} \\ &= \frac{\binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x} \cdot \binom{n}{t-x} \vartheta^{t-x} (1 - \vartheta)^{n-t+x}}{\binom{2n}{t} \vartheta^t (1 - \vartheta)^{2n-t}} \\ &= \frac{\binom{n}{x} \binom{n}{t-x}}{\binom{2n}{t}}. \end{aligned}$$

Dieses Ergebnis ist nicht von ϑ abhängig, d.h.

$X_1 + X_2$ ist eine suffiziente Statistik für ϑ .

Example 2.1.18

Suppose that X_1 and X_2 are independent

$\mathcal{B}(n, \vartheta)$ -distributed. Let us consider the statistic

$S(X_1, X_2) = X_1 + X_2$:

This result does not depend on ϑ , i.e. $X_1 + X_2$ is

a sufficient statistic for ϑ .

Definition 2.1.19 (Vollständigkeit)

Eine suffiziente Statistik T ist vollständig, wenn

wir für jede reellwertige Funktion $g : T(\mathbf{X}) \rightarrow \mathbb{R}$

of T haben:

$$\mathcal{E}_\vartheta \{g[T(X)]\} = 0 \quad \forall \vartheta \in \Theta \implies \mathcal{P}_\vartheta \{g[T(X)] = 0\} = 1 \quad \forall \vartheta \in \Theta.$$

Definition 2.1.19 (Completeness)

A sufficient statistic T is complete, if for every

real-valued function $g : T(X) \rightarrow \mathbb{R}$ of T we have:

Bemerkung 2.1.20

Von dem verschwindenden Erwartungswert, für alle möglichen Parameterwerte, folgt, dass

$g[T(X)]$ auch \mathcal{P}_ϑ -fast sicher gleich Null ist für

alle $\vartheta \in \Theta$.

Remark 2.1.20

From the vanishing expectation for all possible

parameter values, it follows that $g[T(X)]$ also

\mathcal{P}_ϑ -almost surely equals zero for all $\vartheta \in \Theta$.

Bemerkung 2.1.21

Vollständigkeit bedeutet, dass $T(X)$ alle Informationen über ϑ enthält, die in X sind (durch die Suffizienz), aber auch, dass keine unnötige Informationen in $T(X)$ sind. Die Statistik $T(X_1, \dots, X_n) = (X_1, \dots, X_n)$ ist offensichtlich suffizient, da keine Information verloren geht. Sie ist aber nicht vollständig.

Remark 2.1.21

Completeness means that $T(X)$ contains all the information about ϑ , which is in X (by sufficiency), but also that there is no unnecessary information in $T(X)$. The statistic $T(X_1, \dots, X_n) = (X_1, \dots, X_n)$ is obviously sufficient, because no information gets lost. However, it is not complete.

**2.2 Bayes-Schätzer
Bayes estimators**

Unser nächstes Thema ist die Bayesschätzung:

Hierbei ist der wichtigste Punkt, dass Bayesschätzung jede zusätzliche Information berücksichtigt, die wir haben könnten. Deswegen wird unsere Vermutung über den Parameter in die statistische Analyse mit einbezogen.

Die a-priori Verteilung π repräsentiert unsere ursprüngliche Meinung bezüglich des Parameters ϑ . Dies bedeutet, dass ϑ als Zufallsvariable betrachtet wird mit Dichte $\pi(\vartheta)$. Von Bayes Satz 1.1.5,

$$f(X|\vartheta) \cdot \pi(\vartheta) = \pi(\vartheta|X) \cdot f(X)$$

oder

$$\pi(\vartheta|X) = \frac{f(X|\vartheta)\pi(\vartheta)}{\int_{\Theta} f(X|\vartheta)\pi(\vartheta)d\vartheta},$$

Our next topic is Bayesian estimation:

The important point here is that Bayesian estimation takes into account any further information we may have. Thus, our belief concerning the parameter is introduced into the statistical analysis.

The prior distribution π represents our initial opinion about the parameter ϑ . This means that ϑ is regarded as a random variable with density function $\pi(\vartheta)$. From Bayes Theorem 1.1.5,

or

wobei $\pi(\vartheta|X)$ die posteriori Verteilung genannt wird. Sie kann als unsere Vermutung über ϑ , nach Berücksichtigung der Daten, angesehen werden.

Wir definieren das Bayessche Risiko auf folgende Weise:

Definition 2.2.1 (Bayessches Risiko)

Wenn T ein Schätzer von ϑ ist, dann ist das Bayessche Risiko definiert als

$$R_B(T) = \int_{\Theta} R(T, \vartheta) \cdot \pi(\vartheta) d\vartheta,$$

wobei $R(T, \vartheta)$ das normale Risiko ist.

Bemerkung 2.2.2

Auf analoge Weise können wir schreiben

$$\begin{aligned} R_B(T) &= \int_{\Theta} \int_{\mathbb{R}} L(T(x), \vartheta) \cdot f_{\vartheta}(x) dx \cdot \pi(\vartheta) d\vartheta \\ &= \int_{\mathbb{R}} \int_{\Theta} L(T(x), \vartheta) \cdot \pi(\vartheta|x) d\vartheta \cdot f(x) dx. \end{aligned}$$

Definition 2.2.3 (Bayes-Schätzer)

Wenn T_B der Schätzer von ϑ ist, der das Bayessche Risiko 2.2.1 minimiert, dann ist T_B der Bayes-Schätzer von ϑ .

where $\pi(\vartheta|X)$ is called the posterior distribution function. It can be regarded as our updated belief in ϑ , having taken account of the data.

We define the Bayesian risk as follows:

Definition 2.2.1 (Bayes risk)

If T is an estimator of ϑ , then the Bayesian risk is defined as

where $R(T, \vartheta)$ is the usual risk.

Remark 2.2.2

In an analogous manner, we can write

Definition 2.2.3 (Bayes estimator)

If T_B is the estimator of ϑ , which minimizes the Bayes risk 2.2.1, then T_B is the Bayes estimator of ϑ .

Bemerkung 2.2.4

Für X ist das Minimieren des Bayesschen Risikos äquivalent zur Minimierung von

$$\int_{\Theta} L(T(X), \vartheta) \cdot \pi(\vartheta|X) d\vartheta.$$

Remark 2.2.4

For X , minimizing the Bayesian risk is equivalent to minimizing

Bemerkung 2.2.5

Die Definition des Bayes-Schätzers heißt, dass wir den erwarteten posteriori Verlust minimieren.

Remark 2.2.5

The definition of the Bayes estimator means that we minimize the posterior expected loss.

Bemerkung 2.2.6

Um vergleichbar mit klassischer Schätztheorie zu sein, wird die quadratische Verlustfunktion oft auch in Bayesscher Theorie benutzt.

Remark 2.2.6

To suit comparison with classical estimation theory, the quadratic loss function is also often used in Bayesian theory.

Bemerkung 2.2.7

Das Bayessche Risiko ist bzgl. ϑ ein gewichtetes Mittel der Risikofunktion. Die Gewichtsfunktion ist groß für die ϑ , die wir genau schätzen wollen. Die Gewichtsfunktion π ist klein für die ϑ , wo wir einen größeren Schätzfehler akzeptieren können. Somit wählen wir durch π die Teile der Parametermenge, wo wir erwarten, dass der wahre Parameter liegt, bzw. wo wir ihn weniger erwarten.

Remark 2.2.7

The Bayesian risk is w.r.t. ϑ a weighted mean of the risk function. The weight function is big for those ϑ , which we want to estimate accurately. The weight function π is small for those ϑ , where we can accept a bigger estimation error. So, by choosing π , we choose the parts of the parameter space, where we expect the true parameter to be and at the same time the parts where we think that it is more unlikely for the parameter to be.

Wir können den Bayesschen Ansatz auf die folgende Weise interpretieren:

Der Statistiker hat schon im voraus irgendwelche Art von Kenntnissen oder Erwartungen bezüglich des wahren Parameters ϑ und benutzt diese um zu entscheiden wie die Gewichtsfunktion aussehen soll.

We can interpret the Bayesian estimation approach in the following way:

The statistician has already in advance some kind of knowledge or expectations about the true parameter ϑ and uses this to determine how the weight function should look like.

Bemerkung 2.2.8

Eine Verteilung, die oft als a -priori Verteilung für Bayes-Schätzer benutzt wird, ist die Betaverteilung mit Parametern $\alpha > 0$ und $\beta > 0$. Die Dichte der Betaverteilung ist gegeben durch

Remark 2.2.8

A distribution, which often is used as prior distribution for Bayes estimators is the beta distribution with parameters $\alpha > 0$ and $\beta > 0$. The density function of the beta distribution is given by

$$\pi_{\alpha, \beta}(\vartheta) = \frac{\vartheta^{\alpha-1} \cdot (1-\vartheta)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 < \vartheta < 1,$$

mit der Normierung

with the normation

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)},$$

die sicherstellt, dass $\pi_{\alpha, \beta}(\vartheta)$ eine Dichte ist. Abbildung 2.2.1 zeigt die Dichte für einige unterschiedliche Werte von α und β .

which ensures that $\pi_{\alpha, \beta}(\vartheta)$ is a density function. Figure 2.2.1 shows the density function for some different values of α and β .

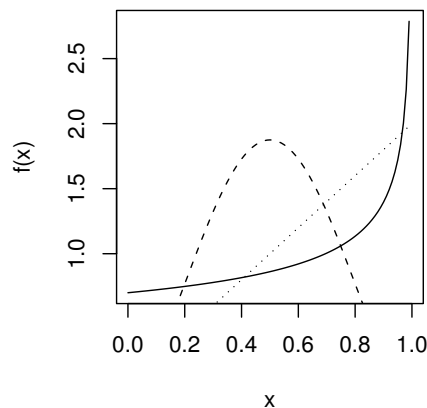


Figure 2.2.1: $\pi_{\alpha,\beta}(\vartheta)$

Bemerkung 2.2.9

Zum Beispiel ergibt sich für die Binomialverteilung, $X \sim \mathcal{B}(n, p_0)$, mit der Betaverteilung als a -priori Verteilung, dass die posteriori Verteilung die folgende Gestalt hat:

$$\pi(\vartheta|X) = \frac{\vartheta^{\alpha-1+X} \cdot (1-\vartheta)^{\beta-1+n-X}}{\int_0^1 u^{\alpha-1+X} \cdot (1-u)^{\beta-1+n-X} du}, \quad \vartheta \in [0, 1].$$

Remark 2.2.9

For example, we have for the binomial distribution, $X \sim \mathcal{B}(n, p_0)$, with prior distribution beta, that the posterior distribution has the following form:

Beispiel 2.2.10

Wir benutzen die posteriori Verteilung aus Bemerkung 2.2.9 und die quadratische Verlustfunktion. Unser Ziel ist es, den Bayes-Schätzer von ϑ zu bestimmen.

Wir minimieren

$$\int_{\Theta} (T(X) - \vartheta)^2 \cdot \pi(\vartheta|X) d\vartheta$$

durch Ableiten nach T:

$$-2 \int_{\Theta} (T(X) - \vartheta) \cdot \pi(\vartheta|X) d\vartheta = 0.$$

Daraus ergibt sich

$$T(X) = \int_{\Theta} \vartheta \cdot \pi(\vartheta|X) d\vartheta,$$

was ganz allgemein für die quadratische Verlustfunktion gilt. Es wurde bisher noch nicht die a -priori bzw. die sich daraus ergebende posteriori Verteilung ins Spiel gebracht.

Führen wir noch die posteriori Verteilung ein, so erhalten wir

Example 2.2.10

We use the posterior distribution from Remark 2.2.9 and the quadratic loss function. Our objective is to determine the Bayes estimator of ϑ .

We minimize

by differentiating w.r.t. T:

From this we receive

which in general is true for the quadratic loss function. Until now, we did not bring the a -priori and the resulting posterior distribution into the game.

If we now introduce the posterior distribution, then we arrive at

$$\begin{aligned}
T(X) &= \int_0^1 \vartheta \cdot \frac{\vartheta^{\alpha-1+X} \cdot (1-\vartheta)^{\beta-1+n-X}}{\int_0^1 u^{\alpha-1+X} \cdot (1-u)^{\beta-1+n-X} du} d\vartheta \\
&= \frac{\int_0^1 \vartheta^{\alpha+X} \cdot (1-\vartheta)^{\beta-1+n-X} d\vartheta}{\int_0^1 u^{\alpha-1+X} \cdot (1-u)^{\beta-1+n-X} du} \\
&= \frac{X+\alpha}{n+\alpha+\beta}.
\end{aligned} \tag{2.2.11}$$

Hier wurde folgende Umformung, die auf partieller Integration basiert, für den Nenner benutzt:

Here, we used the following result for the denominator, based on partial integration:

$$\begin{aligned}
\int_0^1 u^{\alpha-1+X} \cdot (1-u)^{\beta-1+n-X} du &= \int_0^1 [u + (1-u)] \cdot u^{\alpha-1+X} \cdot (1-u)^{\beta-1+n-X} du \\
&= \int_0^1 u^{\alpha+X} \cdot u^{\beta-1+n-X} du + \int_0^1 u^{\alpha-1+X} \cdot u^{\beta+n-X} du \\
&\stackrel{\text{part. Int.}}{=} \left(1 + \frac{\beta+n-X}{\alpha+X}\right) \cdot \int_0^1 u^{\alpha+X} \cdot u^{\beta-1+n-X} du.
\end{aligned}$$

Die Bayes-Schätzer sind mit dem Begriff der Zulässigkeit nahe verwandt:

The Bayes estimators are closely related to the concept of admissibility:

Definition 2.2.12 (Zulässigkeit)

Ein Schätzer T des Parameters ϑ ist zulässig, wenn für alle Schätzer S von ϑ mit

$$R(S, \vartheta) \leq R(T, \vartheta) \quad \forall \vartheta \in \Theta$$

folgt

$$R(S, \vartheta) = R(T, \vartheta) \quad \forall \vartheta \in \Theta.$$

Definition 2.2.12 (Admissibility)

An estimator T of the parameter ϑ is admissible when for all estimators S of ϑ with

follows

Proposition 2.2.13

Sei M der Abschluss der inneren Punkte von Θ , d.h. $M = \overline{\Theta \setminus \partial\Theta}$, und $M \cap \Theta = \emptyset$.

Sei die dem Risiko zugrundeliegende Verlustfunktion stetig in ϑ .

Wenn $\pi(\vartheta) > 0$ fast überall, dann ist der Bayes-Schätzer T_B von ϑ zur a-priori Dichte π ein zulässiger Schätzer.

Beweis:

$R_B(T)$ wird von T_B minimiert.

Wenn T_B nicht zulässig wäre, dann gäbe es einen Schätzer S und $\vartheta_0 \in \Theta$ mit

$$\begin{aligned}
R(S, \vartheta) &\leq R(T_B, \vartheta) \quad \forall \vartheta \in \Theta, \\
R(S, \vartheta_0) &< R(T_B, \vartheta_0).
\end{aligned}$$

Wir wissen, dass Θ keine isolierten Punkte hat und dass das Risiko R stetig in ϑ ist. Daraus folgt, dass $R(S, \vartheta) < R(T, \vartheta)$ für alle ϑ in einer Umgebung (in der Relativtopologie von Θ) von ϑ_0 , und deswegen ist

$$R_B(S) = \int_{\Theta} R(S, \vartheta) \cdot \pi(\vartheta) d\vartheta < R_B(T),$$

was ein Widerspruch zu der Definition von T_B ist. ■

Proposition 2.2.13

Let M be the closure of the inner points of Θ , i.e. $M = \overline{\Theta \setminus \partial\Theta}$, and $M \cap \Theta = \emptyset$.

We assume that the loss function, which is used for the risk, is continuous in ϑ .

If $\pi(\vartheta) > 0$ almost everywhere, then the Bayes estimator T_B of ϑ for the a-priori density π is an admissible estimator.

Proof:

$R_B(T)$ is minimized by T_B .

If T_B would not be admissible, then there would be an estimator S and $\vartheta_0 \in \Theta$ with

We know that Θ has no isolated points and that the risk R is continuous in ϑ . This gives us that $R(S, \vartheta) < R(T, \vartheta)$ for all ϑ in an environment of ϑ_0 (in the relative topology of Θ) and therefore

which is a contradiction to the definition of T_B . ■

Bemerkung 2.2.14

Es ist auch möglich zu zeigen, dass jeder zulässiger Schätzer ein Bayes-Schätzer ist, zumindest falls π gegen eine degenerierte Funktion geht.

Dies lässt sich am Beispiel 2.2.9 erkennen. Der Bayes-schätzer lautet $\frac{X+\alpha}{n+\alpha+\beta}$. Betrachtet man $\alpha = \beta = 0$, so erhält man als Schätzer die relative Häufigkeit $\frac{X}{n}$. Allerdings ist die zugehörige a-priori Betaverteilung für $\alpha = \beta = 0$ nicht definiert!

Bemerkung 2.2.15

Es scheint, dass Bayes-Schätzer das Konzept für die Konstruktion von Schätzern schlechthin sind.

Allerdings kann man sich lange und trefflich über die zu verwendende a-priori Verteilung streiten. Ferner kann man den Schätzer nicht immer so explizit bestimmen, wie in obigem Beispiel.

Es wird deshalb innerhalb der Bayesschen Statistik solch eine a-priori Verteilung gewählt, für die man den Schätzer explizit berechnen kann.

Remark 2.2.14

It is also possible to show that every admissible estimator is a Bayes estimator, at least if π tends to a degenerate function.

We can observe this in the example 2.2.9. The Bayes estimator is $\frac{X+\alpha}{n+\alpha+\beta}$. If we take $\alpha = \beta = 0$, then we receive the relative frequency $\frac{X}{n}$ as estimator. However, the corresponding a-priori beta distribution for $\alpha = \beta = 0$ is not defined!

Remark 2.2.15

It seems that Bayes estimators are the best concept for the construction of estimators.

However, there is always a long and intensive discussion about the a-priori distribution to be used. Furthermore, it is not always possible to get an explicit solution for the estimator as in the example above.

Very often, the Bayesian statistics chooses the a-priori distribution such that the estimator can be calculated explicitly.

2.3 Minimaxschätzer

Minimax estimators

In diesem Abschnitt studieren wir Minimaxschätzer. Diese Schätzer minimieren das maximale Risiko, d.h. sie sichern einen gegen die schlimmste Situation ab.

Definition 2.3.1 (Minimaxschätzer)

Ein Schätzer T_M , der

$$R_M(T) := \max_{\vartheta \in \Theta} R(T, \vartheta)$$

minimiert, wird der Minimaxschätzer von ϑ genannt.

Bemerkung 2.3.2

Minimaxschätzer werden oft als „zu pessimistische Schätzer“ bezeichnet. Sie sind optimal für die Kontrolle der schlimmsten möglichen Situation, d.h. sie begrenzen das maximale mögliche Risiko für alle möglichen Parameter ϑ . Allerdings ist das Risiko für viele $\vartheta \in \Theta$ wesentlich größer als für andere Schätzer.

In this section, we study minimax estimators. These estimators choose the smallest maximum risk, i.e. they are said to take precautions against the worst case situation.

Definition 2.3.1 (Minimax estimator)

An estimator T_M , which minimizes

is called the minimax estimator of ϑ .

Remark 2.3.2

Minimax estimators are often said to be “too pessimistic estimators”. They are optimal for the the control of the worst case situation, i.e. they limit the maximal risk for all possible parameters ϑ . However, the risk is for many $\vartheta \in \Theta$ much higher than for other estimators.

Bemerkung 2.3.3

Implizit wird angenommen, dass das Risiko $R(T, \vartheta)$ sein Maximum annimmt. Man kann aber die Definition auch über das Supremum machen. Im allgemeinen ist das Risiko stetig und sehr oft wird Θ als kompakte Menge gewählt, was die Existenz des Maximums sichert.

Remark 2.3.3

We assume implicitly that the risk $R(T, \vartheta)$ attains its maximum. However, we could formulate the definition also via the supremum. In general, the risk is continuous and very often, we have that Θ is a compact set, which together ensures the existence of the maximum.

Bemerkung 2.3.4

Es ist fast unmöglich den Minimaxschätzer direkt zu berechnen! Es gibt aber eine Beziehung zwischen Bayes- und Minimaxschätzern, die wir in folgendem Satz sehen und später in einem Beispiel ausnutzen werden.

Remark 2.3.4

It is nearly impossible to calculate the minimax estimator directly! However, there is a connection between Bayes and minimax estimators, which we will see in the next theorem and which we later will use in an example.

Satz 2.3.5

Sei T_B ein Bayes-Schätzer zu einer beliebigen a-priori Verteilung π mit

Theorem 2.3.5

Let T_B be a Bayes estimator for an arbitrary a-priori distribution π such that

$$R(T_B, \vartheta) \leq R_B(T_B) \quad \forall \vartheta \in \Theta.$$

Dann ist T_B auch der Minimaxschätzer T_M .

Then T_B is also the minimax estimator T_M .

Ist π stetig und $\pi(\vartheta) > 0$ für alle $\vartheta \in \Theta$, so gilt sogar

If π is continuous and $\pi(\vartheta) > 0$ for all $\vartheta \in \Theta$, then even

$$R(T_M, \vartheta) = R_B(T_M) \quad \forall \vartheta \in \Theta,$$

d.h. in diesem Fall ist das Risiko des Minimaxschätzers konstant.

i.e. in this case the risk of the minimax estimator is constant.

Beweis:

Im allgemeinen gilt (siehe Definition 2.2.1)

Proof:

We have in general (see Definition 2.2.1)

$$R_B(T) = \int_{\Theta} R(T, \vartheta) \cdot \pi(\vartheta) d\vartheta \leq \max_{\vartheta \in \Theta} R(T, \vartheta) \tag{2.3.6}$$

für jeden beliebigen Schätzer T.

for any estimator T.

Gilt nun $R_B(T_B) \geq R(T_B, \vartheta)$ für alle $\vartheta \in \Theta$, so folgt mit der allgemein gültigen Relation 2.3.6:

If it now holds that $R_B(T_B) \geq R(T_B, \vartheta)$ for all $\vartheta \in \Theta$, then with the general relation 2.3.6 follows:

$$\max_{\vartheta \in \Theta} R(T_B, \vartheta) \leq R_B(T_B) \leq R_B(T) \leq \max_{\vartheta \in \Theta} R(T, \vartheta)$$

für jeden beliebigen Schätzer T, da T_B der entsprechende Bayes-Schätzer ist. Somit ist gezeigt, dass $T_B = T_M$.

for any estimator T, because T_B is the corresponding Bayes estimator. Hence, we have shown that $T_B = T_M$.

Die zweite Aussage wird per Widerspruchsbeweis gezeigt: Wir nehmen an, dass

The second statement will be shown by contradiction: We assume that

$$r_{min} := \min_{\vartheta \in \Theta} R(T_B, \vartheta) < \max_{\vartheta \in \Theta} R(T_B, \vartheta) =: r_{max}.$$

Wir setzen

We define

$$\Theta_0 := \left\{ \vartheta \in \Theta \mid R(T_B, \vartheta) < \frac{1}{2}(r_{min} + r_{max}) \right\},$$

was wegen $r_{min} < r_{max}$ nicht leer ist. Wegen der Stetigkeit von π und $\pi(\vartheta) > 0$ folgt $\int_{\Theta_0} \pi(\vartheta) d\vartheta > 0$. Zusammen ergibt dies

which is not empty because of $r_{min} < r_{max}$. From the continuity of π and $\pi(\vartheta) > 0$ follows $\int_{\Theta_0} \pi(\vartheta) d\vartheta > 0$. Altogether, we arrive at

$$\begin{aligned} R_B(T_B) &= \int_{\Theta_0} R(T_B, \vartheta) \cdot \pi(\vartheta) d\vartheta + \int_{\Theta \setminus \Theta_0} R(T_B, \vartheta) \cdot \pi(\vartheta) d\vartheta \\ &\leq \frac{1}{2} \cdot (r_{min} + r_{max}) \cdot \int_{\Theta_0} \pi(\vartheta) d\vartheta + r_{max} \cdot \int_{\Theta \setminus \Theta_0} \pi(\vartheta) d\vartheta \\ &< r_{max}, \end{aligned}$$

was der Voraussetzung

$\max_{\vartheta \in \Theta} R(T_B, \vartheta) \leq R_B(T_B)$ widerspricht. ■

which contradicts the assumption

$\max_{\vartheta \in \Theta} R(T_B, \vartheta) \leq R_B(T_B)$. ■

Bemerkung 2.3.7

Hat man also einen Schätzer T , der konstantes Risiko hat, so braucht man „lediglich“ noch die geeignete a -priori Verteilung π , um nachzuweisen, dass es sich um den Minimaxschätzer handelt.

Allerdings kann es natürlich noch einen anderen Schätzer mit kleinerem konstanten Risiko geben!

Remark 2.3.7

If we have an estimator T , which exhibits a constant risk, then we “only” need an appropriate a -priori distribution π to prove that it is the minimax estimator.

However, there can be another estimator with constant risk which is smaller!

Beispiel 2.3.8

Sei $X \sim \mathcal{B}(n, p_0)$, d.h. binomialverteilt mit unbekanntem Parameter p_0 , wie im Beispiel 2.2.9.

Als Verlustfunktion wählen wir

$$L(p, p_0) = \frac{(p - p_0)^2}{p_0 \cdot (1 - p_0)},$$

Example 2.3.8

Let $X \sim \mathcal{B}(n, p_0)$, i.e. binomial distributed with unknown parameter p_0 as in example 2.2.9.

For the loss function, we choose

die klarerweise Abweichungen für p_0 nahe 0 oder 1 viel stärker gewichtet als für $p_0 = \frac{1}{2}$.
Wir untersuchen das Risiko für den Schätzer $\bar{X}_n = \frac{X}{n}$:

which clearly gives a higher weight to deviations for p_0 near 0 or 1 than for $p_0 = \frac{1}{2}$.
We investigate the risk for the estimator $\bar{X}_n = \frac{X}{n}$:

$$\begin{aligned} R(\bar{X}_n, p_0) &= \mathcal{E} \left[\frac{(\bar{X}_n - p_0)^2}{p_0 \cdot (1 - p_0)} \right] \\ &= \frac{1}{p_0 \cdot (1 - p_0)} \cdot \mathcal{V} \mathcal{R}(\bar{X}_n) \\ &= \frac{1}{n}, \end{aligned}$$

d.h. dies ist ein Kandidat für den Minimaxschätzer für dieses Risiko.
Wir brauchen nun noch eine a -priori Verteilung, für die dieser Schätzer der Bayes-Schätzer ist. Als Versuch nehmen wir die Gleichverteilung auf $[0; 1]$. Die posteriori Verteilung ist damit proportional zu $f(X|p_0) \cdot \pi(p_0) = p_0^X \cdot (1 - p_0)^{n-X}$.
Den Bayes-Schätzer ermitteln wir, indem wir das Bayesrisiko punktweise (d.h. nicht nochmals über X mitteln) minimieren. Wir minimieren

i.e. this is a candidate for the minimax estimator for this risk.
We just need an a -priori distribution such that this estimator is the Bayes estimator. Let us try with the uniform distribution on $[0; 1]$. The posterior distribution is now proportional to $f(X|p_0) \cdot \pi(p_0) = p_0^X \cdot (1 - p_0)^{n-X}$.
We receive the Bayes estimator by minimizing the Bayes risk pointwise (i.e. we do not average over the different values of X). We minimize

$$\int_0^1 \frac{(p - p_0)^2}{p_0 \cdot (1 - p_0)} \cdot p_0^X \cdot (1 - p_0)^{n-X} dp_0.$$

durch Ableiten nach dem (unbekannten) Schätzer p :

by differentiating with respect to the (unknown) estimator p :

$$p = \frac{\int_0^1 p_0^X \cdot p_0^{n-X-1} dp_0}{\int_0^1 p_0^{X-1} \cdot p_0^{n-X-1} dp_0}$$

Dies hatten wir bereits in Gleichung 2.2.11 mit $\alpha = \beta = 0$, d.h. wir haben hier die in Bemerkung 2.2.14 angesprochene Grenzsituation.

We had this equation already in 2.2.11 with $\alpha = \beta = 0$, i.e. we are here in the limit situation described in Remark 2.2.14.

2.4 Maximum-Likelihood-Schätzer Maximum likelihood estimators

2.4.1 Grundlegende Idee und Definitionen Basic idea and definitions

Die grundlegende Idee der Maximum-Likelihood-Schätzung beginnt mit der gemeinsamen Dichte von $\mathbf{X} = (X_1, \dots, X_n)$, die vom unbekanntem Parameter ϑ abhängt.

The basic idea of maximum likelihood estimation starts with the joint distribution of $\mathbf{X} = (X_1, \dots, X_n)$, depending upon the unknown parameter ϑ ,

$$f(\mathbf{X}, \vartheta) = f(X_1, \dots, X_n, \vartheta).$$

Für festes ϑ können wir Wahrscheinlichkeitsaussagen bezüglich \mathbf{X} machen. Vertauschen wir die Rolle von \mathbf{X} und ϑ (d.h. wir sehen \mathbf{X} als fest an), so kommen wir zur Likelihood-Funktion

For fixed ϑ , we can make probability statements about \mathbf{X} . If we interchange the role of \mathbf{X} and ϑ (i.e. we regard \mathbf{X} as fixed), then we arrive at the likelihood function

$$L(\vartheta|\mathbf{X}) = f(\mathbf{X}, \vartheta).$$

Der Wert von ϑ , der $L(\vartheta|\mathbf{X})$ maximiert, wird Maximum-Likelihood-Schätzwert von ϑ genannt.

The value of ϑ , which maximizes $L(\vartheta|\mathbf{X})$, is called the maximum likelihood estimate of ϑ .

Definition 2.4.1 (Likelihoodfunktion)

Sei \mathbf{x} eine Realisierung von \mathbf{X} mit Werten in \mathbf{X} und einer Verteilung aus $\{\mathcal{P}_\vartheta | \vartheta \in \Theta\}$.

- Falls $\mathcal{P}_\vartheta(X = x) > 0$, dann ist die Likelihoodfunktion definiert als

$$L(\vartheta|x) = \mathcal{P}_\vartheta(X = x) \quad x \in \mathbf{X}, \vartheta \in \Theta.$$

- Falls $\mathcal{P}_\vartheta(X = x) = 0$, dann ist die Likelihoodfunktion definiert als

$$L(\vartheta|x) = \lim_{h \rightarrow 0} \frac{\mathcal{P}_\vartheta(X \in [x-h, x+h])}{2h} \quad x \in \mathbf{X}, \vartheta \in \Theta.$$

Bemerkung 2.4.2

Wir setzen später auch X statt x ein.

Im zweiten Fall erhalten wir die Dichte an der Stelle x .

Wir möchten betonen, dass statt der Likelihoodfunktion es mathematisch oft einfacher ist, die Log-Likelihoodfunktion zu verwenden:

$$l(\vartheta|x) = \log L(\vartheta|x) \quad x \in \mathbf{X}, \vartheta \in \Theta.$$

Jetzt können wir die Definition des Maximum-Likelihood-Schätzers angeben:

Definition 2.4.1 (Likelihood function)

Let \mathbf{x} be a realization of \mathbf{X} with values in \mathbf{X} and a distribution in $\{\mathcal{P}_\vartheta, \vartheta \in \Theta\}$.

- If $\mathcal{P}_\vartheta(X = x) > 0$, then the likelihood function is defined as

- If $\mathcal{P}_\vartheta(X = x) = 0$, then the likelihood function is defined as

Remark 2.4.2

We will later also insert X instead of x .

In the second case, we receive the density at x .

We would like to stress that instead of using the likelihood function, it is often mathematically more convenient to take the log-likelihood function:

Definition 2.4.3
(Maximum-Likelihood-Schätzer)

$\hat{\vartheta}$ ist der Maximum-Likelihood-Schätzer (ML) von ϑ , wenn

$$L(\hat{\vartheta}(X)|X) = \max_{\vartheta \in \Theta} L(\vartheta|X).$$

Definition 2.4.3 (Maximum likelihood estimator)

$\hat{\vartheta}$ is the maximum likelihood estimator (ML) of ϑ , if

2.4.2 Beispiele
Examples

Beispiel 2.4.4

Seien X_1, \dots, X_n u.i. normalverteilte Zufallsvariablen, $\mathcal{N}(\mu, \sigma^2)$.
Dann ist die Likelihoodfunktion

$$\begin{aligned} L(\mu, \sigma^2 | X_1, \dots, X_n) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2\right) \end{aligned}$$

und die Log-Likelihoodfunktion ist

$$\begin{aligned} l(\mu, \sigma^2 | X_1, \dots, X_n) &= \log L(\mu, \sigma^2 | X_1, \dots, X_n) \\ &= -\frac{n}{2} \cdot \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^n (X_i - \mu)^2. \end{aligned}$$

Ableiten nach μ und σ^2 ergibt

Example 2.4.4

Suppose X_1, \dots, X_n are i.i. normally distributed random variables, $\mathcal{N}(\mu, \sigma^2)$.
Then the likelihood function is

and the log-likelihood function is

Differentiating with respect to μ and σ^2 results in

$$\begin{aligned} \frac{\partial l}{\partial \mu}(\mu, \sigma^2 | X_1, \dots, X_n) &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) \\ \frac{\partial l}{\partial \sigma^2}(\mu, \sigma^2 | X_1, \dots, X_n) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \cdot \sum_{i=1}^n (X_i - \mu)^2. \end{aligned}$$

Setzen wir diese Ableitungen gleich Null, so ergibt sich

Equating these derivatives to zero gives

$$\begin{aligned} \hat{\mu} &= \bar{X}_n \\ \hat{\sigma}^2 &= \frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X}_n)^2. \end{aligned}$$

Beispiel 2.4.5

Sei X eine binomialverteilte Zufallsvariable, $X \sim \mathcal{B}(n, p)$.
Dann ist die Likelihoodfunktion

$$L(p|X) = \binom{n}{X} \cdot p^X \cdot (1-p)^{n-X},$$

und die Log-Likelihoodfunktion

$$l(p|X) = X \cdot \log(p) + (n-X) \cdot \log(1-p) + \log\left(\binom{n}{X}\right).$$

Maximieren von l ergibt

Maximizing l results in

$$\frac{\partial l}{\partial p}(p|X) = \frac{X}{p} - \frac{n-X}{1-p} \stackrel{!}{=} 0,$$

und somit

which leads to

$$\hat{p} = \frac{X}{n}.$$

Der Test mit der zweiten Ableitung zeigt, dass dies der ML-Schätzer ist:

$$\frac{\partial^2 l}{\partial p^2}(p|X) = -\frac{X}{p^2} - \frac{n-X}{(1-p)^2} < 0.$$

The test by the second derivative shows that this is the ML estimator:

Beispiel 2.4.6

Wir nehmen an, dass die Zufallsvariablen X_1, \dots, X_n u.i. exponentialverteilt, mit Parameter $\lambda > 0$, sind. Die Likelihoodfunktion ergibt sich zu

$$L(\lambda|X_1, \dots, X_n) = \prod_{i=1}^n \lambda \cdot e^{-\lambda \cdot X_i}$$

und die Log-Likelihoodfunktion, samt Ableitung nach λ , ist

$$l(\lambda|X_1, \dots, X_n) = n \cdot \log(\lambda) - \lambda \cdot \sum_{i=1}^n X_i;$$

$$\frac{\partial l}{\partial \lambda}(\lambda|X_1, \dots, X_n) = \frac{n}{\lambda} - \sum_{i=1}^n X_i.$$

Wir sehen, dass $\frac{\partial l}{\partial \lambda}$ streng monoton fallend als Funktion von λ ist, d.h. die Lösung von $\frac{\partial l}{\partial \lambda} = 0$ ergibt den Maximum-Likelihood-Schätzer

$$\hat{\lambda} = \frac{1}{\bar{X}_n}.$$

Example 2.4.6

Suppose that the random variables X_1, \dots, X_n are i.i. exponentially distributed with parameter $\lambda > 0$. The likelihood function is given by

and the log-likelihood function, inclusive the derivative with respect to λ , is

We observe that $\frac{\partial l}{\partial \lambda}$ is strictly decreasing as a function of λ , i.e. the solution of $\frac{\partial l}{\partial \lambda} = 0$ leads to the maximum likelihood estimator

Beispiel 2.4.7

Seien die Zufallsvariablen X_1, \dots, X_n u.i. gleichverteilt auf $(0, \vartheta)$. Dann ist die Dichte

$$f(x_i, \vartheta) = \begin{cases} \frac{1}{\vartheta}, & 0 \leq x_i \leq \vartheta, \\ 0, & x_i \notin [0; \vartheta]. \end{cases}$$

Die Bestimmung des ML-Schätzers von ϑ ist nicht ganz einfach, weil der Träger der Dichte vom Parameter ϑ abhängt. Die Likelihoodfunktion ist

$$L(\vartheta|X_1, \dots, X_n) = \frac{1}{\vartheta^n}, \quad 0 \leq X_1, \dots, X_n \leq \vartheta.$$

Ableiten nach ϑ führt zur Erkenntnis, dass ϑ möglichst klein, d.h. nahe Null, sein muss. Man hat aber stets die Nebenbedingung $0 \leq X_1, \dots, X_n \leq \vartheta$. Somit ist der ML-Schätzer in diesem Fall die maximale Ordnungsstatistik

$$\hat{\vartheta} = \max_{i \in \{1, \dots, n\}} X_i = X_{(n)}.$$

Example 2.4.7

Let the random variables X_1, \dots, X_n be i.i. uniformly distributed on $(0, \vartheta)$. Then the density function is

To find the ML estimator of ϑ is not quite easy because the support of the density depends on the parameter. The likelihood function is

Differentiating with respect to ϑ leads to the fact that ϑ should be as small as possible, i.e. close to zero. However, we always have the constraint $0 \leq X_1, \dots, X_n \leq \vartheta$. Therefore, we obtain the ML estimator as the maximum order statistic

2.4.3 Eigenschaften Properties

Die erste Bemerkung geht auf die Nützlichkeit der Log-Likelihoodfunktion ein.

The first remark enlightens the usefulness of the log-likelihood function.

Bemerkung 2.4.8

Wie wir in den Beispielen 2.4.4 und 2.4.6 gesehen haben, wird bei unabhängigen Zufallsvariablen aus dem Produkt bei der Likelihoodfunktion eine Summe bei der Log-Likelihoodfunktion, die wesentlich einfacher abzuleiten ist, d.h. auch das Maximum und damit auch der ML-Schätzer ist wesentlich einfacher bestimmbar!

Der restliche Abschnitt behandelt einige theoretische Eigenschaften der Maximum-Likelihood-Schätzer.

Bemerkung 2.4.9

Maximum-Likelihood-Schätzer sind meistens konsistent, auch wenn einige pathologische Ausnahmen existieren.

Betrachten wir das Risiko, so können wir zeigen, dass unter gewissen Glattheitsannahmen, die Maximum-Likelihood-Schätzer das geringst mögliche Risiko haben.

Wir untersuchen wie klein der mittlere quadratische Fehler $R(T_n, \vartheta)$ für einen Schätzer $T_n = T_n(X_1, \dots, X_n)$ von ϑ werden kann. Wir beschränken uns auf den Fall mit eindimensionalem Parameter. Zunächst brauchen wir aber den Begriff der Fisher-Information:

Remark 2.4.8

As we have seen in the case of independent random variables in the Examples 2.4.4 and 2.4.6, we have a product for the likelihood function which becomes a sum in the case of the log-likelihood function. For the determination of the maximum, especially the differentiation, and therefore also the ML estimator, the sum is much more convenient than the product!

The rest of this section treats some theoretical properties of maximum likelihood estimators.

Remark 2.4.9

Maximum likelihood estimators are mostly consistent, even if some pathological exceptions exist.

Regarding the risk, we can show that under some smoothness conditions, the maximum likelihood estimators asymptotically have the smallest possible risk.

Let us investigate how small the mean squared error $R(T_n, \vartheta)$ for an estimator $T_n = T_n(X_1, \dots, X_n)$ of ϑ can be. We confine ourselves to the case with a one-dimensional parameter. First we need to introduce the concept of Fisher's information:

Definition 2.4.10 (Fisher-Information)

Sei $\mathcal{P}_{\vartheta_0, n}$ die Verteilung von $\mathbf{X} = (X_1, \dots, X_n)$ mit Werten in \mathbb{R}^n , wobei $\vartheta_0 \in \Theta \subseteq \mathbb{R}$. Die Log-Likelihoodfunktion sei partiell nach ϑ ableitbar. Dann ist die Fisher-Information von $\mathcal{P}_{\vartheta_0, n}$ definiert als

$$I(\mathcal{P}_{\vartheta_0, n}) := \mathcal{E}_{\vartheta_0} \left\{ \frac{\partial l}{\partial \vartheta} (\vartheta_0 | \mathbf{X}) \right\}^2.$$

Bemerkung 2.4.11**(Fisher-Informationsmatrix)**

Die Verallgemeinerung der Fisher-Information auf einen m -dimensionalen Parameterraum ist in der Literatur nicht ganz klar. Es gibt zwei Definitionen:

$$I(\mathcal{P}_{\vartheta_0, n}) := \left(\mathcal{E}_{\vartheta_0} \left\{ \frac{\partial l}{\partial \vartheta_i} (\vartheta_0 | \mathbf{X}) \cdot \frac{\partial l}{\partial \vartheta_j} (\vartheta_0 | \mathbf{X}) \right\} \right)_{i, j=1, \dots, m}$$

oder

$$I(\mathcal{P}_{\vartheta_0, n}) := \left(\mathcal{E}_{\vartheta_0} \left\{ - \frac{\partial^2 l}{\partial \vartheta_i \partial \vartheta_j} (\vartheta_0 | \mathbf{X}) \right\} \right)_{i, j=1, \dots, m}.$$

Definition 2.4.10 (Fisher's information)

Let $\mathcal{P}_{\vartheta_0, n}$ be the distribution of $\mathbf{X} = (X_1, \dots, X_n)$ with values in \mathbb{R}^n , where $\vartheta_0 \in \Theta \subseteq \mathbb{R}$. We assume that the log-likelihood function is partially differentiable w.r.t. ϑ . Then, the Fisher's information of $\mathcal{P}_{\vartheta_0, n}$ is defined as

Remark 2.4.11 (Fisher's information matrix)

The generalization of Fisher's information to an m -dimensional parameter space is not clear in the literature. There are two definitions:

Allerdings sind beide Definitionen identisch, sofern man die Integration bei der Erwartungswertbildung und die Ableitung nach ϑ_i vertauschen darf (einfache Übung). Anderst sieht die Situation aber aus, falls es sich um ein misspezifiziertes Modell handelt, d.h. die wahre Verteilung $\mathcal{P} \notin \{\mathcal{P}_{\vartheta,n} | \vartheta \in \Theta\}$. Dann nimmt man den Erwartungswert bzgl. der wahren Verteilung \mathcal{P} von \mathbf{X} und beide Definitionen stimmen nicht mehr überein! In diesem Fall wird die zweite Definition verwendet.

Wir betrachten hier allerdings nur richtig spezifizierte Modelle, d.h. $\mathcal{P} \in \{\mathcal{P}_{\vartheta,n} | \vartheta \in \Theta\}$ und beschränken uns auf die erste Definition.

Wenn die Daten X_1, \dots, X_n u.i.v. sind, dann ist die Fisher-Information besonders einfach:

However, both definitions are identical, whenever we are allowed to interchange integration and differentiation with respect to ϑ_i (simple exercise). The situation is totally different, if the model is misspecified, i.e. the true distribution $\mathcal{P} \notin \{\mathcal{P}_{\vartheta,n} | \vartheta \in \Theta\}$. Then we must take the expectation with respect to the true distribution \mathcal{P} of \mathbf{X} and both definitions do not coincide any more! In this case, the second definition is used.

However, we consider here only correctly specified models, i.e. $\mathcal{P} \in \{\mathcal{P}_{\vartheta,n} | \vartheta \in \Theta\}$ and restrict ourselves to the first definition.

In the case of i.i.d. data X_1, \dots, X_n , the Fisher's information is especially simple:

Korollar 2.4.12 (Fisher-Information bei u.i.v. Zufallsvariablen)

Seien X_1, \dots, X_n u.i.v. mit $\mathcal{L}(X_i) = \mathcal{P}_{\vartheta_0}$. Ferner sei $\{\mathcal{P}_{\vartheta} | \vartheta \in \Theta\}$ „glatt“, d.h. dass

1. alle \mathcal{P}_{ϑ} absolut-stetig bzgl. des Lebesgue-Maßes auf \mathbb{R} sind, d.h. es gibt eine Dichte f_{ϑ} ;
2. $f_{\vartheta}(x)$ als Funktion von ϑ stetig partiell ableitbar ist; die Schreibweise $\frac{\partial f_{\vartheta_0}}{\partial \vartheta}(x)$ bezeichnet die partielle Ableitung der Dichte nach ϑ an der Stelle (ϑ_0, x) ;
3. $\int_{\mathbb{R}} \left| \frac{\partial f_{\vartheta_0}}{\partial \vartheta}(x) \right| dx < \infty$.

Dann haben wir

$$I(\mathcal{P}_{\vartheta_0,n}) = n \cdot I(\mathcal{P}_{\vartheta_0}).$$

Beweis:

Die Log-Likelihoodfunktion $l(\vartheta | \mathbf{X})$ von $\mathbf{X} = (X_1, \dots, X_n)$ ist, wegen der Unabhängigkeit, $\sum_{i=1}^n l(\vartheta | X_i)$. Somit folgt

$$\begin{aligned} I(\mathcal{P}_{\vartheta_0,n}) &= \mathcal{E}_{\vartheta_0} \left\{ \left[\frac{\partial}{\partial \vartheta} \sum_{j=1}^n l(\vartheta_0 | X_j) \right]^2 \right\} \\ &= \sum_{j=1}^n \mathcal{E}_{\vartheta_0} \left\{ \left[\frac{\partial l}{\partial \vartheta}(\vartheta_0 | X_j) \right]^2 \right\} + \sum_{i \neq j} \mathcal{E}_{\vartheta_0} \left\{ \left[\frac{\partial l}{\partial \vartheta}(\vartheta | X_i) \right] \cdot \left[\frac{\partial l}{\partial \vartheta}(\vartheta | X_j) \right] \right\} \\ &= n \cdot I(\mathcal{P}_{\vartheta_0}), \end{aligned}$$

Corollary 2.4.12 (Fisher's information for i.i.d. random variables)

Let X_1, \dots, X_n be i.i.d. with $\mathcal{L}(X_i) = \mathcal{P}_{\vartheta_0}$. $\{\mathcal{P}_{\vartheta}; \vartheta \in \Theta\}$ should be "smooth", i.e.

1. all \mathcal{P}_{ϑ} are absolutely-continuous w.r.t. the Lebesgue-measure in \mathbb{R}^n , i.e. it exists a density f_{ϑ} ;
2. $f_{\vartheta}(x)$ is, as a function of ϑ , continuously partially differentiable; we will use the notation $\frac{\partial f_{\vartheta_0}}{\partial \vartheta}(x)$ for the partial derivative of the density w.r.t. ϑ at (ϑ_0, x) ;
3. $\int_{\mathbb{R}} \left| \frac{\partial f_{\vartheta_0}}{\partial \vartheta}(x) \right| dx < \infty$.

Then we have

Proof:

The log-likelihood function $l(\vartheta | \mathbf{X})$ of $\mathbf{X} = (X_1, \dots, X_n)$ is equal to $\sum_{i=1}^n l(\vartheta | X_i)$, because of the independence. This leads to

weil wir wegen der Unabhängigkeit der X_j für $i \neq j$ haben:

$$\mathcal{E}_{\vartheta_0} \left\{ \left[\frac{\partial}{\partial \vartheta} (\vartheta | X_i) \right] \cdot \left[\frac{\partial}{\partial \vartheta} (\vartheta | X_j) \right] \right\} = \left\{ \mathcal{E}_{\vartheta_0} \left[\frac{\partial}{\partial \vartheta} (\vartheta | X_i) \right] \right\} \cdot \left\{ \mathcal{E}_{\vartheta_0} \left[\frac{\partial}{\partial \vartheta} (\vartheta | X_j) \right] \right\} = 0.$$

Um die letzte Identität zu zeigen, benötigen wir die Glattheitsannahmen an \mathcal{P}_ϑ , insbesondere für die Vertauschung von Integration und Differentiation:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \vartheta} 1 = \frac{\partial}{\partial \vartheta} \int_{\mathbb{R}} f_{\vartheta_0}(x) dx = \int_{\mathbb{R}} \frac{\partial f_{\vartheta_0}}{\partial \vartheta}(x) dx = \int_{\mathbb{R}} \left\{ \frac{\partial}{\partial \vartheta} \ln[f_{\vartheta_0}(x)] \right\} \cdot f_{\vartheta_0}(x) dx \\ &= \mathcal{E}_{\vartheta_0} \left\{ \frac{\partial}{\partial \vartheta} (\vartheta | X_1) \right\}. \end{aligned}$$

since we have due to the independence of X_j for $i \neq j$:

In order to show the last identity, we must have the smoothness assumptions on \mathcal{P}_ϑ , especially for the interchange of integration and differentiation:

■

■

Bemerkung 2.4.13

Es gibt verschiedene Möglichkeiten die Glattheit von $\{\mathcal{P}_\vartheta | \vartheta \in \Theta\}$ zu definieren.

Die Existenz der Dichten ermöglicht die Darstellung der Fisher-Information mit dem Lebesgue-Maß als dominierendes Maß. Man kann hier aber auch annehmen, dass es ein beliebiges dominierendes Maß μ für alle \mathcal{P}_ϑ gibt. Die restlichen Bedingungen garantieren lediglich die Vertauschbarkeit von Integration und Differentiation. Bekanntermaßen gibt es hierfür auch andere Möglichkeiten, z.B.:

1. Es gibt eine kompakte Menge $A \subset \mathbb{R}^n$, so dass $\text{supp}(f_\vartheta) = \overline{\{x \in \mathbb{R} | f_\vartheta(x) > 0\}} \subseteq A$.
2. Die Dichten $f_\vartheta(x)$ sind stetig als Funktion von $(\vartheta, x) \in \Theta \times \mathbb{R}$ und bzgl. ϑ stetig partiell differenzierbar.

Unter ähnlichen Glattheitsannahmen, können wir die Fisher-Information benutzen um eine untere Schranke für das Risiko des Schätzers T_n von ϑ_0 zu finden.

Remark 2.4.13

There are different possibilities to define smoothness for $\{\mathcal{P}_\vartheta | \vartheta \in \Theta\}$.

The existence of the densities enables us to represent Fisher's information with the Lebesgue-measure as dominating measure. However, it suffices to have any dominating measure μ for all \mathcal{P}_ϑ . The other conditions guarantee the possibility to interchange the integration and differentiation. It is well known that there are also other possibilities, e.g.:

1. There exists a compact subset $A \subset \mathbb{R}^n$ such that $\text{supp}(f_\vartheta) = \overline{\{x \in \mathbb{R} | f_\vartheta(x) > 0\}} \subseteq A$.
2. The densities $f_\vartheta(x)$ are continuous as functions of $(\vartheta, x) \in \Theta \times \mathbb{R}$. The partial derivative w.r.t. ϑ exists and is continuous.

With similar smoothness assumptions, we can use the Fisher's information to find a lower bound of the risk for the estimator T_n of ϑ_0 .

Satz 2.4.14 (Die Cramér–Rao Ungleichung)

Sei T_n ein Schätzer von ϑ mit Bias $b_n(\vartheta)$ und die Ableitung $b'_n(\vartheta)$ nach ϑ existiere.

Ferner sei $\{\mathcal{P}_{\vartheta,n} | \vartheta \in \Theta\}$ „glatt“ im Sinne von

Bemerkung 2.4.13, oder Korollar 2.4.12 samt

$$\int_{\mathbb{R}^n} \left| T_n(\mathbf{x}) \cdot \frac{\partial f_{\vartheta_0,n}(\mathbf{x})}{\partial \vartheta} \right| d\mathbf{x} < \infty.$$

Dann gilt

$$R(T_n, \vartheta) = \mathcal{E}_{\vartheta} \left[(T_n(\mathbf{X}) - \vartheta)^2 \right] \geq \frac{(b'_n(\vartheta) + 1)^2}{I(\mathcal{P}_{\vartheta,n})} \quad \forall \vartheta \in \Theta.$$

Beweis:

$\vartheta_0 \in \Theta$ sei beliebig, aber fest.

Der Bias von T_n ist

$$b_n(\vartheta_0) = \mathcal{E}_{\vartheta_0}(T_n(\mathbf{X}) - \vartheta_0) = \int_{\mathbb{R}^n} \{T_n(\mathbf{x}) - \vartheta_0\} \cdot f_{\vartheta_0,n}(\mathbf{x}) d\mathbf{x}.$$

Wenn wir bzgl. ϑ ableiten und die Bedingungen

2. bis 4. benutzen, können wir die Ableitung und

die Integration vertauschen:

$$\begin{aligned} b'_n(\vartheta_0) &= \frac{\partial}{\partial \vartheta} \int_{\mathbb{R}^n} \{T_n(\mathbf{x}) - \vartheta_0\} \cdot f_{\vartheta_0,n}(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \frac{\partial}{\partial \vartheta} \left[\{T_n(\mathbf{x}) - \vartheta_0\} \cdot f_{\vartheta_0,n}(\mathbf{x}) \right] d\mathbf{x} \\ &= - \int_{\mathbb{R}^n} f_{\vartheta_0,n}(\mathbf{x}) d\mathbf{x} + \int_{\mathbb{R}^n} \{T_n(\mathbf{x}) - \vartheta_0\} \cdot \frac{\partial f_{\vartheta_0,n}(\mathbf{x})}{\partial \vartheta} d\mathbf{x}. \end{aligned}$$

Theorem 2.4.14 (The Cramér–Rao inequality)

Let T_n be an estimator of ϑ with bias $b_n(\vartheta)$

where the derivative $b'_n(\vartheta)$ w.r.t. ϑ exists.

$\{\mathcal{P}_{\vartheta,n}; \vartheta \in \Theta\}$ should be “smooth” in the sense

of Remark 2.4.13, or Corollary 2.4.12 together

$$\text{with } \int_{\mathbb{R}^n} \left| T_n(\mathbf{x}) \cdot \frac{\partial f_{\vartheta_0,n}(\mathbf{x})}{\partial \vartheta} \right| d\mathbf{x} < \infty.$$

Then we have

Proof:

Let $\vartheta_0 \in \Theta$ arbitrary but fixed.

The bias of T_n is

When we take the derivative w.r.t. ϑ and use the

conditions 2. till 4., we can change the order of

differentiation and integration:

Weil $f_{\vartheta_0,n}$ eine Dichte ist, folgt

$$\begin{aligned} b'_n(\vartheta_0) + 1 &= \int_{\mathbb{R}^n} \{T_n(\mathbf{x}) - \vartheta_0\} \cdot \frac{\partial}{\partial \vartheta} \left\{ \ln[f_{\vartheta_0,n}(\mathbf{x})] \right\} \cdot f_{\vartheta_0,n}(\mathbf{x}) d\mathbf{x} \\ &= \mathcal{E}_{\vartheta_0} \left\{ [T_n(\mathbf{X}) - \vartheta_0] \cdot \frac{\partial}{\partial \vartheta} \left[\ln(f_{\vartheta_0,n}(\mathbf{X})) \right] \right\} \\ &\leq \left\{ \mathcal{E}_{\vartheta_0} [T_n(\mathbf{X}) - \vartheta_0]^2 \cdot \mathcal{E}_{\vartheta_0} \left[\frac{\partial}{\partial \vartheta} \ln(f_{\vartheta_0,n}(\mathbf{X})) \right]^2 \right\}^{\frac{1}{2}}, \end{aligned}$$

wobei wir im letzten Schritt die

Cauchy–Schwartzsche Ungleichung benutzt

haben. Falls die Fisher–Information endlich ist,

dann folgt die Cramér–Rao Ungleichung sofort.

Wenn nicht, so ist die rechte Seite der

Cramér–Rao Ungleichung gleich Null, und die

Ungleichung ist automatisch erfüllt. ■

Since $f_{\vartheta_0,n}$ is a density, it follows

where we in the last step have used the

Cauchy–Schwarz’ inequality. If the Fisher’s

information is finite, then the Cramér–Rao

inequality follows immediately. If not, then the

right-hand side of the Cramér–Rao inequality

equals zero, and the inequality is automatically

fulfilled. ■

Bemerkung 2.4.15

Interessant ist, dass die Ableitung des Bias mit

eingeht, nicht aber der Bias selbst.

Hat man somit ein ϑ_0 wo die Ableitung

$b'(\vartheta_0) = -1$, so ist die

Cramér–Rao–Ungleichung nichtssagend.

Für einen unverzerrten Schätzer ist $b'(\vartheta) = 0$

für alle $\vartheta \in \Theta$.

Remark 2.4.15

It is interesting that we just need the derivative

of the bias, but not the bias itself.

If there is a ϑ_0 such that the derivative

$b'(\vartheta_0) = -1$, then the Cramér–Rao inequality

does not give any information.

If we have an unbiased estimator, then $b'(\vartheta) = 0$

for all $\vartheta \in \Theta$.

Definition 2.4.16 (Effizienz eines Schätzers)

Hat man die Folge T_n von Schätzern von ϑ und

gilt

Definition 2.4.16 (Efficiency of an estimator)

Let T_n be a series of estimators of ϑ with

$$R(T_n, \vartheta) \rightarrow \frac{1}{I(\mathcal{P}_{\vartheta,n})} \quad \forall \vartheta \in \Theta,$$

dann nennt man T_n effizient.

then we call T_n efficient.

Bemerkung 2.4.17

Man fordert für die Effizienz, dass asymptotisch der Bias verschwindet und die Varianz die inverse Fisher-Information erreicht. Es gibt Beispiele von Schätzern, die „super-effizient“ sind, d.h. deren Risiko schneller gegen Null geht. Im Abschnitt 2.4.4 sind zwei Beispiele ausgeführt. Allerdings sind dies Ausnahmen.

Remark 2.4.17

For efficiency, we must have that the bias vanishes and the variance behaves like the inverse Fisher's information. There are examples of estimators which are „super-efficient“, i.e. whose risk converges faster to zero. We will consider two examples in Section 2.4.4. However, these are exceptions.

Eine andere wichtige Eigenschaft ist die Invarianz der Maximum-Likelihood-Schätzer:

Another important property is the invariance of maximum likelihood estimators:

Satz 2.4.18

Seien X_1, \dots, X_n u.i. \mathcal{P}_ϑ mit $\vartheta \in \Theta$. Es gebe einen Maximum-Likelihood-Schätzer für ϑ , den wir mit $\hat{\vartheta}(X_1, \dots, X_n)$ bezeichnen.

Theorem 2.4.18

Let X_1, \dots, X_n be i.i. \mathcal{P}_ϑ with $\vartheta \in \Theta$. Assume that there exists a maximum likelihood estimator for ϑ , which we denote $\hat{\vartheta}(X_1, \dots, X_n)$.

Seien Y_1, \dots, Y_n gegeben durch $Y_i = g(X_i)$ für eine injektive Funktion g , d.h. die Verteilung der u.i.v. Zufallsvariablen Y_1, \dots, Y_n hat die Form $\mathcal{P}_\vartheta \circ g^{-1}$, mit $\vartheta \in \Theta$.

Let Y_1, \dots, Y_n be given as $Y_i = g(X_i)$ for an injective function g , i.e. the distribution of the i.i.d. random variables Y_1, \dots, Y_n has the form $\mathcal{P}_\vartheta \circ g^{-1}$, with $\vartheta \in \Theta$.

Dann ist die Likelihood-Funktion $L(\hat{\vartheta}(g^{-1}(Y_1), \dots, g^{-1}(Y_n)) | Y_1, \dots, Y_n)$ ebenfalls maximal, d.h. $\hat{\vartheta}(g^{-1}(Y_1), \dots, g^{-1}(Y_n))$ ist ein Maximum-Likelihood-Schätzer für den Parameter ϑ der Verteilung $\mathcal{P}_\vartheta \circ g^{-1}$ der Y_i .

Then the likelihood function $L(\hat{\vartheta}(g^{-1}(Y_1), \dots, g^{-1}(Y_n)) | Y_1, \dots, Y_n)$ is also at the maximum, i.e. $\hat{\vartheta}(g^{-1}(Y_1), \dots, g^{-1}(Y_n))$ is a maximum likelihood estimator for the parameter ϑ of the distribution $\mathcal{P}_\vartheta \circ g^{-1}$ of Y_i .

Bevor wir mit dem Beweis beginnen eine kurze Anmerkung zur Schreibweise $\mathcal{P}_\vartheta \circ g^{-1}$:

Before we start with the proof, a short remark to the notation $\mathcal{P}_\vartheta \circ g^{-1}$:

Bemerkung 2.4.19

Die Darstellung der Verteilung der Y_i läuft unter der Bezeichnung „induziertes Maß“. g^{-1} wird dabei im allgemeinen als Mengenfunktion angesehen, d.h. $g^{-1}(M)$ für eine Menge M ist die Menge \tilde{M} aller x mit $g(x) \in M$. In unserem Fall ist g aber injektiv, so dass das Urbild eines Punktes eindeutig ist oder nicht existiert.

Remark 2.4.19

The representation of the distribution of the Y_i is denoted by „induced measure“. g^{-1} is in general considered as a set function, i.e. $g^{-1}(M)$ for a set M is the set \tilde{M} of all x with $g(x) \in M$. However, in our case is g injective, so that the inverse image of a point is unique or does not exist.

Beweis:

Wir wollen nicht unterscheiden müssen, ob bei der Likelihood-Funktion die Verteilung selbst oder die entsprechende Dichte eingeht. Deshalb steht H_ϑ entweder für die Verteilung oder die Dichte, je nachdem ob die Verteilung diskret oder stetig ist. Das entsprechende für die Y_i ist $H_\vartheta \circ g^{-1}$. Für $\vartheta \in \Theta$ gilt

Proof:

We do not want to distinguish if we for the likelihood function are concerned with the distribution itself or the corresponding density. Therefore H_ϑ means either the distribution or the density, depending on whether the distribution is discrete or continuous. The corresponding for the Y_i is $H_\vartheta \circ g^{-1}$. For $\vartheta \in \Theta$ holds

$$L(\hat{\vartheta} | Y_1, \dots, Y_n) = H_\vartheta \circ g^{-1}(Y_1, \dots, Y_n) = H_\vartheta(g^{-1}(Y_1), \dots, g^{-1}(Y_n)) = H_\vartheta(X_1, \dots, X_n) = L(\hat{\vartheta} | X_1, \dots, X_n).$$

Wir wissen, dass $\hat{\vartheta}(X_1, \dots, X_n) = \hat{\vartheta}(g^{-1}(Y_1), \dots, g^{-1}(Y_n))$ die Likelihood-Funktion $L(\hat{\vartheta} | X_1, \dots, X_n)$ maximiert. Da diese aber identisch mit der Likelihood-Funktion $L(\hat{\vartheta} | Y_1, \dots, Y_n)$ ist, haben wir den Satz bewiesen.

We know that $\hat{\vartheta}(X_1, \dots, X_n) = \hat{\vartheta}(g^{-1}(Y_1), \dots, g^{-1}(Y_n))$ maximizes the likelihood function $L(\hat{\vartheta} | X_1, \dots, X_n)$. Since this is identical with the likelihood function $L(\hat{\vartheta} | Y_1, \dots, Y_n)$, we have proved the theorem.

2.4.4 Interessante Beispiele von Maximum-Likelihood-Schätzern
Interesting examples of maximum likelihood estimators

Beispiel 2.4.20 (Fortsetzung von Beispiel 2.4.7)

X_1, \dots, X_n u.i. uniform auf $[0; \vartheta_0]$ verteilt.
 Der Maximum-Likelihood-Schätzer ist $X_{(n)}$.
 Um das Risiko berechnen zu können, brauchen wir die Verteilung des Maximums:

$$\mathcal{P}(X_{(n)} \leq x) = \mathcal{P}(X_1 \leq x; \dots; X_n \leq x) = \prod_{i=1}^n \mathcal{P}(X_i \leq x) = \left(\frac{x}{\vartheta_0}\right)^n. \quad (2.4.21)$$

Damit können wir das Risiko dieses Schätzers bestimmen:

$$\begin{aligned} R(X_{(n)}, \vartheta_0) &= \mathcal{E}_{\vartheta_0} [(X_{(n)} - \vartheta_0)^2] = \int_0^{\vartheta_0} (x - \vartheta_0)^2 d \left(\frac{x}{\vartheta_0}\right)^n = \int_0^{\vartheta_0} (x - \vartheta_0)^2 \cdot \left[\frac{n}{\vartheta_0^n} \cdot x^{n-1}\right] dx \\ &= \vartheta_0^2 \cdot \frac{2}{n^2 + 3 \cdot n + 2}. \end{aligned}$$

Interessanterweise ist dieser Schätzer schon super-effizient, da er mit Ordnung n^2 gegen Null geht und die Fisher-Information nur von Ordnung n ist (siehe Korollar 2.4.12 auf Seite 70).

Allerdings geht es noch etwas besser: Wir betrachten den Schätzer $\frac{n+1}{n} \cdot X_{(n)}$ und dessen Risiko.

$$R\left(\frac{n+1}{n} \cdot X_{(n)}, \vartheta_0\right) = \int_0^{\vartheta_0} \left(\frac{n+1}{n} \cdot x - \vartheta_0\right)^2 \cdot \frac{n}{\vartheta_0^n} \cdot x^{n-1} dx = \vartheta_0^2 \cdot \frac{1}{n^2 + 2n}.$$

Example 2.4.20 (Continuation of Example 2.4.7)

X_1, \dots, X_n i.i. uniformly distributed on $[0; \vartheta_0]$.
 The maximum likelihood estimator is $X_{(n)}$.
 For the calculation of the risk, we need the distribution of the maximum:

With this knowledge, we can determine the risk of the estimator:

It is interesting that this estimator is already super-efficient, because it tends to zero of order n^2 whereas Fisher's information is only of order n (see Corollary 2.4.12 on page 70).

However, there is a slightly better estimator: We consider the estimator $\frac{n+1}{n} \cdot X_{(n)}$ and its risk.

Beispiel 2.4.22 (Maximum-Likelihood-Schätzer nicht eindeutig)

Betrachte die u.i.v. Zufallsvariablen X_1, \dots, X_n mit $X_i \sim \mathcal{U}(\vartheta_0 - \frac{1}{2}; \vartheta_0 + \frac{1}{2})$.
 Dies führt zu $\vartheta_0 - \frac{1}{2} \leq X_{(1)} < X_{(n)} \leq \vartheta_0 + \frac{1}{2}$,
 d.h. die Likelihoodfunktion ist

$$L(\vartheta | \mathbf{X}) = \begin{cases} 1, & \vartheta \in [X_{(n)} - \frac{1}{2}; X_{(1)} + \frac{1}{2}], \\ 0, & \vartheta \notin [X_{(n)} - \frac{1}{2}; X_{(1)} + \frac{1}{2}]. \end{cases}$$

Wir sehen, dass

$$\max_{\vartheta} L(\vartheta | \mathbf{X}) = 1 \quad \forall \vartheta \in \left[X_{(n)} - \frac{1}{2}; X_{(1)} + \frac{1}{2}\right].$$

Dies bedeutet, dass die Likelihoodfunktion für jede Statistik $\hat{\vartheta}(X_1, \dots, X_n)$ mit $\hat{\vartheta}(\mathbf{X}) \in [X_{(n)} - \frac{1}{2}; X_{(1)} + \frac{1}{2}]$ maximiert wird. Somit ist dieser Maximum-Likelihood-Schätzer im allgemeinen nicht eindeutig!

Wir wollen nun das Risiko für die beiden Maximum-Likelihood-Schätzer $X_{(n)} - \frac{1}{2}$ und $\frac{X_{(1)} + X_{(n)}}{2}$ miteinander vergleichen.

Example 2.4.22 (Maximum likelihood estimator not unique)

Consider the i.i.d. random variables X_1, \dots, X_n with $X_i \sim \mathcal{U}(\vartheta_0 - \frac{1}{2}; \vartheta_0 + \frac{1}{2})$.
 This leads to $\vartheta_0 - \frac{1}{2} \leq X_{(1)} < X_{(n)} \leq \vartheta_0 + \frac{1}{2}$, i.e. the likelihood function is

We note that

This yields that the likelihood function is maximized for every statistic $\hat{\vartheta}(X_1, \dots, X_n)$ such that $\hat{\vartheta}(\mathbf{X}) \in [X_{(n)} - \frac{1}{2}; X_{(1)} + \frac{1}{2}]$. Thus, this maximum likelihood estimator is in general not unique!

We now want to compare the risk for the two maximum likelihood estimators $X_{(n)} - \frac{1}{2}$ and $\frac{X_{(1)} + X_{(n)}}{2}$ with each other.

Die Verteilung des Maximums $X_{(n)}$ haben wir mehr oder weniger bereits in Beispiel 2.4.20 bestimmt. Somit können wir einfach das Risiko für den Maximum Likelihood Schätzer $X_{(n)}$ bestimmen:

$$R\left(X_{(n)} - \frac{1}{2}, \vartheta_0\right) = \int_{\vartheta_0 - \frac{1}{2}}^{\vartheta_0 + \frac{1}{2}} \left(x - \frac{1}{2} - \vartheta_0\right)^2 \cdot d\left[x - \left(\vartheta_0 - \frac{1}{2}\right)\right]^n = \frac{2}{(n+1) \cdot (n+2)}.$$

Die Bestimmung der Dichte von $(X_{(1)}; X_{(n)})$ basiert darauf, dass $\mathcal{P}(X_{(1)} \geq x; X_{(n)} \leq y) = \mathcal{P}(x \leq X_1, \dots, X_n \leq y) = (y-x)^n$, falls $x \leq y$ und Null sonst:

$$\begin{aligned} f(x,y) &= \lim_{dx \rightarrow 0+} \lim_{dy \rightarrow 0+} \frac{\mathcal{P}(X_{(1)} \geq x; X_{(n)} \leq y) - \mathcal{P}(X_{(1)} \geq x+dx; X_{(n)} \leq y-dy)}{dx \cdot dy} \\ &= \lim_{dx \rightarrow 0+} \lim_{dy \rightarrow 0+} \frac{(y-x)^n - (y-dy-x+dx)^n}{dx \cdot dy} \\ &= n \cdot (n-1) \cdot (y-x)^{n-2}, \quad \vartheta_0 - \frac{1}{2} \leq x \leq y \leq \vartheta_0 + \frac{1}{2}. \end{aligned}$$

Nun können wir endgültig das Risiko des Maximum-Likelihood-Schätzers $\frac{X_{(1)}+X_{(n)}}{2}$ berechnen:

$$R\left(\frac{X_{(1)}+X_{(n)}}{2}, \vartheta_0\right) = \int_{\vartheta_0 - \frac{1}{2}}^{\vartheta_0 + \frac{1}{2}} \int_x^{\vartheta_0 + \frac{1}{2}} \left(\frac{x+y}{2} - \vartheta_0\right)^2 \cdot f(x,y) dx dy = \frac{1}{2 \cdot (n+2) \cdot (n+1)}.$$

We have more or less determined the distribution of $X_{(n)}$ already in Example 2.4.20. Therefore, we can easily calculate the risk for the maximum likelihood estimator $X_{(n)}$:

The determination of the density of $(X_{(1)}; X_{(n)})$ is based on the fact that $\mathcal{P}(X_{(1)} \geq x; X_{(n)} \leq y) = \mathcal{P}(x \leq X_1, \dots, X_n \leq y) = (y-x)^n$, if $x \leq y$ and zero otherwise:

Now, we can finally calculate the risk of the maximum likelihood estimator $\frac{X_{(1)}+X_{(n)}}{2}$:

Auf den ersten Blick scheint es verwunderlich zu sein, dass das Risiko von $\frac{X_{(1)}+X_{(n)}}{2}$ nur ein Viertel des Risikos von $X_{(n)} - \frac{1}{2}$ ist, da doch beide Maximum-Likelihood-Schätzer sind. Auf den zweiten Blick wird aber klar, dass wir zwei unterschiedliche Dinge betrachten. Auf der einen Seite die Likelihood und auf der anderen Seite das quadratische Risiko!

At a first glance it seems mysterious that the risk of $\frac{X_{(1)}+X_{(n)}}{2}$ is a quarter of the risk of $X_{(n)} - \frac{1}{2}$, although both are maximum likelihood estimators. At a second glance, we notice that we consider two different things. On one hand, the likelihood and on the other hand the quadratic loss!

2.5 Konsistenz und asymptotische Normalität von M-Schätzern Consistency and asymptotic normality of M-estimators

Zunächst definieren wir, was M-Schätzer sind.

First, we define what M-estimators are.

Definition 2.5.1 (M-Schätzer)

Sei $Q_n : (\mathbf{X}_n; \vartheta) \rightarrow \mathbb{R}$ ein Funktional des Zufallsvektors $\mathbf{X}_n = (X_1, \dots, X_n)$ und des unbekanntem Parametervektors $\vartheta \in \Theta$.

Ein M-Schätzer $\hat{\vartheta}_n(\mathbf{X}_n)$ von ϑ erfüllt

$$Q_n(\mathbf{X}_n; \hat{\vartheta}_n(\mathbf{X}_n)) \geq Q_n(\mathbf{X}_n; \vartheta) \quad \forall \vartheta \in \Theta.$$

Definition 2.5.1 (M-estimator)

Let $Q_n : (\mathbf{X}_n; \vartheta) \rightarrow \mathbb{R}$ be a functional of the random vector $\mathbf{X}_n = (X_1, \dots, X_n)$ and of the unknown parameter vector $\vartheta \in \Theta$.

An M-estimator $\hat{\vartheta}_n(\mathbf{X}_n)$ of ϑ fulfills

Bemerkung 2.5.2

Der Name M-Schätzer ist von Maximum-Likelihood-Schätzern abgeleitet. Das M steht hier für „Maximum“, da ein Funktional maximiert wird.

Remark 2.5.2

The name M-estimator is derived from the maximum likelihood estimators. M stands for „maximum“, since a functional is maximized.

Hat man die Situation, dass man ein Funktional minimieren will, wie z.B. bei Bayes-Schätzern, so will man nach einem Vorzeichenwechsel des Funktional das neue maximieren, d.h. auch diese Situation fällt unter M-Schätzern.

Somit ist die Definition eines M-Schätzers sehr allgemein gehalten. Insbesondere sind eben Maximum-Likelihood-, Bayes- und Minimax-Schätzer auch M-Schätzer.

Wir gehen davon aus, dass wir irgendeinen Algorithmus haben, der uns einen M-Schätzer $\hat{\vartheta}_n(\mathbf{X}_n)$ liefert, d.h. es muss sich nicht um einen expliziten Schätzer, wie z.B. \bar{X}_n , handeln.

Wir verwenden „einen“ M-Schätzer, da dieser im allgemeinen nicht eindeutig sein muss.

In der Statistik ist für Punktschätzer das Ziel, die Konsistenz und asymptotische Normalität zu zeigen.

Wie können wir dies beweisen, wenn wir lediglich das zu maximierende Funktional Q_n kennen, aber keinen expliziten Schätzer haben?

If you have the situation that a functional shall be minimized, as e.g. for the Bayes estimators, then you want to maximize the new functional, obtained after a change of sign, i.e. also this would be a case of M-estimators.

Thus, the definition of an M-estimator is very general. Especially, also maximum likelihood, Bayes and minimax estimators are M-estimators.

We assume that we have some algorithm to receive an M-estimator $\hat{\vartheta}_n(\mathbf{X}_n)$. However, it does not have to be an explicit estimator as for example \bar{X}_n .

We use the notion “an” M-estimator, because in general it does not have to be unique.

In statistics, the aim is for point estimators to show consistency and asymptotic normality.

How can we show that, if we only know the functional Q_n to be maximized, but have no explicit estimators?

Satz 2.5.3 (Konsistenz von M-Schätzern)

Falls

1. die Parametermenge $\Theta \subset \mathbb{R}^m$ kompakt ist,
2. $Q_n(\mathbf{X}_n; \vartheta)$ stetig bzgl. ϑ und messbar bzgl. \mathbf{X}_n ist,
3. $\frac{1}{n} \cdot Q_n(\mathbf{X}_n; \vartheta)$, gleichmäßig bzgl. ϑ , in Wahrscheinlichkeit gegen eine nicht-stochastische Funktion $Q(\vartheta)$ konvergiert,
4. $Q(\vartheta)$ ein eindeutiges globales Maximum in $\vartheta_0 \in \Theta$ hat,

dann gilt für den M-Schätzer $\hat{\vartheta}_n(\mathbf{X}_n)$, dass

Theorem 2.5.3 (Consistency of M-estimators)

If

1. the parameter space $\Theta \subset \mathbb{R}^m$ is compact,
2. $Q_n(\mathbf{X}_n; \vartheta)$ is continuous in ϑ and measurable with respect to \mathbf{X}_n ,
3. $\frac{1}{n} \cdot Q_n(\mathbf{X}_n; \vartheta)$ converges in probability and uniformly in ϑ to a nonstochastic function $Q(\vartheta)$,
4. $Q(\vartheta)$ attains a unique global maximum in $\vartheta_0 \in \Theta$,

then holds for the M-estimator $\hat{\vartheta}_n(\mathbf{X}_n)$ that

$$\hat{\vartheta}_n(\mathbf{X}_n) \xrightarrow[n \rightarrow \infty]{P} \vartheta_0.$$

Beweis:

Wir müssen zeigen, dass

Proof:

We must show that

$$P(\hat{\vartheta}_n \in N) \rightarrow 1$$

für jede offene Umgebung N von ϑ_0 .

Sei nun N beliebig, aber fest.

Mit N^c bezeichnen wir das abgeschlossene

Komplement von N in Θ

for any open neighborhood N of ϑ_0 .

So, let N be arbitrary but fixed.

We denote with N^c the closed complement of N

in Θ

$$\varepsilon := Q(\vartheta_0) - \max_{\vartheta \in N^c} Q(\vartheta) > 0,$$

weil Q das eindeutige, globale Maximum ϑ_0 hat.

because Q has the unique global maximum ϑ_0 .

Sei

Let

$$A_n := \left\{ \xi \in \Omega : \left| \frac{1}{n} \cdot Q_n(\mathbf{X}_n; \vartheta) - Q(\vartheta) \right| < \frac{\varepsilon}{2} \quad \forall \vartheta \in \Theta \right\}.$$

A_n beinhaltet alle Elementarereignisse ξ unseres Wahrscheinlichkeitsraums Ω , für die $\frac{1}{n} \cdot Q_n$ nahe bei Q ist.
Da $\frac{1}{n} \cdot Q_n \xrightarrow{P} Q$ uniform in ϑ , gilt

A_n collects all elementary events ξ from our probability space Ω where $\frac{1}{n} \cdot Q_n$ and Q are close.
Since $\frac{1}{n} \cdot Q_n \xrightarrow{P} Q$ uniformly in ϑ , we obtain

$$\mathcal{P}(A_n) \xrightarrow{n \rightarrow \infty} 1.$$

Für $\xi \in A_n$ gilt

For $\xi \in A_n$ holds

1. $Q(\hat{\vartheta}_n(X_1(\xi), \dots, X_n(\xi))) > \frac{1}{n} \cdot Q_n(X_1(\xi), \dots, X_n(\xi); \hat{\vartheta}_n(X_1(\xi), \dots, X_n(\xi))) - \frac{\varepsilon}{2};$
2. $\frac{1}{n} \cdot Q_n(X_1(\xi), \dots, X_n(\xi); \vartheta_0) > Q(\vartheta_0) - \frac{\varepsilon}{2}.$

Daraus folgern wir, dass für $\xi \in A_n$ gilt

From this we deduce that for $\xi \in A_n$ holds

$$\begin{aligned} Q(\hat{\vartheta}_n(X_1(\xi), \dots, X_n(\xi))) &> \frac{1}{n} \cdot Q_n(X_1(\xi), \dots, X_n(\xi); \hat{\vartheta}_n(X_1(\xi), \dots, X_n(\xi))) - \frac{\varepsilon}{2} \\ &\geq \frac{1}{n} Q_n(X_1(\xi), \dots, X_n(\xi); \vartheta_0) - \frac{\varepsilon}{2} \\ &> Q(\vartheta_0) - \varepsilon \end{aligned}$$

wobei wir die Optimalität von $\hat{\vartheta}_n$ für Q_n , im Gegensatz zu ϑ_0 , ausgenutzt haben.

where we used the optimality of $\hat{\vartheta}_n$ for Q_n compared with ϑ_0 .

Aus der Definition von ε folgt nun offenbar, dass

It is now obvious from the definition of ε that

$$\hat{\vartheta}_n(X_1(\xi), \dots, X_n(\xi)) \in N \quad \forall \xi \in A_n.$$

Dies impliziert

This implies

$$\mathcal{P}(\hat{\vartheta}_n(X_1(\xi), \dots, X_n(\xi)) \in N) \geq \mathcal{P}(A_n) \xrightarrow{n \rightarrow \infty} 1.$$

■

■

Bemerkung 2.5.4

1. Es gibt viele ähnliche Sätze.
Einige fordern nicht die Kompaktheit von Θ , allerdings wird im Beweis dann eine kompakte Menge $K \subset \Theta$ konstruiert und gezeigt, dass $\hat{\vartheta}_n \xrightarrow{P} K$.
2. Die gleichmäßige Konvergenz bzgl. ϑ von $\frac{1}{n} \cdot Q_n$ ist am schwierigsten nachzuweisen.
3. Falls Q kein eindeutiges globales Maximum hat, sondern es eine Menge $B \subset \Theta$, mit $Q(\tilde{\vartheta}) > Q(\vartheta)$ für alle $\tilde{\vartheta} \in B$ und $\vartheta \in \Theta \setminus B$, gibt, dann erhält man $\hat{\vartheta}_n \xrightarrow{P} B$, d.h. $\inf_{\vartheta \in B} \|\hat{\vartheta}_n - \vartheta\|$ konvergiert in Wahrscheinlichkeit gegen Null.
4. Falls das Modell misspezifiziert ist, dann ist $\vartheta_0 \in \Theta$ optimal im Sinne der Kullback–Leibler Diskrepanz.

Nach dem Satz über die Konsistenz, kommen wir nun zur asymptotischen Normalität.

Remark 2.5.4

1. There are many theorems of that kind.
Some relax the condition about the compactness of Θ , however in the proofs, a compact set $C \subset \Theta$ is constructed and it is shown that $\hat{\vartheta}_n \xrightarrow{P} C$.
2. The critical point to be shown is the uniform convergence in ϑ of $\frac{1}{n} \cdot Q_n$.
3. If we have not a unique maximum of Q but a whole set $B \subset \Theta$ with $Q(\tilde{\vartheta}) > Q(\vartheta)$ where $\tilde{\vartheta} \in B$ and $\vartheta \in \Theta \setminus B$, then $\hat{\vartheta}_n \xrightarrow{P} B$, i.e. $\inf_{\vartheta \in B} \|\hat{\vartheta}_n - \vartheta\|$ converges to zero in probability.
4. If the model is misspecified, then $\vartheta_0 \in \Theta$ is optimal in the sense of Kullback–Leibler discrepancy.

After having shown consistency, we can talk about asymptotic normality.

Satz 2.5.5 (Asymptotische Normalität)

Zusätzlich zu den Annahmen im Satz 2.5.3 benötigen wir, dass

1. die Hesse-Matrix $\frac{\partial^2 Q_n}{\partial \vartheta \partial \vartheta^T}$ existiert und stetig in einer Umgebung von ϑ_0 ist;
2. $\frac{1}{n} \cdot \left(\frac{\partial^2 Q_n}{\partial \vartheta \partial \vartheta^T} \right) \Big|_{\vartheta_n^*} \xrightarrow{p} \mathcal{A}(\vartheta_0)$ für $\vartheta_n^* \xrightarrow{p} \vartheta_0$ und $\mathcal{A}(\vartheta_0)$ eine deterministische, invertierbare $n \times n$ Matrix ist;
3. $\frac{1}{\sqrt{n}} \left(\frac{\partial Q_n}{\partial \vartheta} \Big|_{\vartheta_0} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{B}(\vartheta_0))$ und $\mathcal{B}(\vartheta_0) = \lim_{n \rightarrow \infty} \frac{1}{n} \cdot \mathcal{E} \left\{ \frac{\partial Q_n}{\partial \vartheta} \Big|_{\vartheta_0} \cdot \left(\frac{\partial Q_n}{\partial \vartheta} \Big|_{\vartheta_0} \right)^T \right\}$.

Dann folgt

$$\sqrt{n} \cdot (\hat{\vartheta}_n(\mathbf{X}_n) - \vartheta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{A}(\vartheta_0)^{-1} \cdot \mathcal{B}(\vartheta_0) \cdot \mathcal{A}(\vartheta_0)^{-1}).$$

Beweis:

Der einzige Beweistrick ist die Anwendung der Taylorreihenentwicklung von $\frac{\partial Q_n}{\partial \vartheta} \Big|_{\hat{\vartheta}_n}$ um ϑ_0 :

$$\frac{\partial Q_n}{\partial \vartheta} \Big|_{\hat{\vartheta}_n} = \frac{\partial Q_n}{\partial \vartheta} \Big|_{\vartheta_0} + \frac{\partial^2 Q_n}{\partial \vartheta \partial \vartheta^T} \Big|_{\vartheta^*} \cdot (\hat{\vartheta}_n - \vartheta_0)$$

wobei ϑ^* zwischen ϑ_0 und $\hat{\vartheta}_n$ liegt.

Theorem 2.5.5 (Asymptotic normality)

Additionally to the assumptions of Theorem 2.5.3, let us assume that

1. the Hesse-matrix $\frac{\partial^2 Q_n}{\partial \vartheta \partial \vartheta^T}$ exists and is continuous in an open neighbourhood of ϑ_0 ;
2. $\frac{1}{n} \cdot \left(\frac{\partial^2 Q_n}{\partial \vartheta \partial \vartheta^T} \right) \Big|_{\vartheta_n^*} \xrightarrow{p} \mathcal{A}(\vartheta_0)$ for $\vartheta_n^* \xrightarrow{p} \vartheta_0$ where $\mathcal{A}(\vartheta_0)$ is a deterministic and invertible $n \times n$ matrix;
3. $\frac{1}{\sqrt{n}} \left(\frac{\partial Q_n}{\partial \vartheta} \Big|_{\vartheta_0} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{B}(\vartheta_0))$ and $\mathcal{B}(\vartheta_0) = \lim_{n \rightarrow \infty} \frac{1}{n} \cdot \mathcal{E} \left\{ \frac{\partial Q_n}{\partial \vartheta} \Big|_{\vartheta_0} \cdot \left(\frac{\partial Q_n}{\partial \vartheta} \Big|_{\vartheta_0} \right)^T \right\}$.

Then follows

Wir wissen, dass die linke Seite $\frac{\partial Q_n}{\partial \vartheta} \Big|_{\hat{\vartheta}_n} = 0$, da $\hat{\vartheta}_n$ das Funktional Q_n maximiert. Dies liefert uns

We know that the left-hand-side $\frac{\partial Q_n}{\partial \vartheta} \Big|_{\hat{\vartheta}_n} = 0$, because $\hat{\vartheta}_n$ maximizes Q_n . This gives us

$$\begin{aligned} \sqrt{n} \cdot (\hat{\vartheta}_n - \vartheta_0) &= - \underbrace{\left(\frac{1}{n} \cdot \frac{\partial^2 Q_n}{\partial \vartheta \partial \vartheta^T} \Big|_{\vartheta^*} \right)^+}_{\rightarrow [\mathcal{A}(\vartheta_0)]^{-1}} \cdot \underbrace{\left(\frac{1}{\sqrt{n}} \cdot \frac{\partial Q_n}{\partial \vartheta} \Big|_{\vartheta_0} \right)}_{\xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{B}(\vartheta_0))} \\ &\xrightarrow{\mathcal{L}} -[\mathcal{A}(\vartheta_0)]^{-1} \cdot Z \end{aligned}$$

da / since $\vartheta^* \xrightarrow{p} \vartheta_0$

wobei $Z \sim \mathcal{N}(0, \mathcal{B}(\vartheta_0))$ und $+$ bei der Hesse-Matrix die sogenannte Pseudoinverse bezeichnet.

Der Nachweis, dass $-[\mathcal{A}(\vartheta_0)]^{-1} \cdot Z$ die angegebene Verteilung hat, ist relativ einfach und deshalb nicht ausgeführt. ■

where $Z \sim \mathcal{N}(0, \mathcal{B}(\vartheta_0))$ and $+$ at the Hesse-matrix denotes the so-called pseudo-inverse.

To check that $-[\mathcal{A}(\vartheta_0)]^{-1} \cdot Z$ has the desired distribution is rather easy and therefore omitted. ■

Bemerkung 2.5.6

1. Für Maximum-Likelihood-Schätzer in einem korrekt spezifizierten Modell gilt $-\mathcal{A}(\vartheta_0) = \mathcal{B}(\vartheta_0) = \mathbb{I}(\mathcal{P}_{\vartheta_0})$ (siehe auch Bemerkung 2.4.11).
Damit erhält man dann als Kovarianzmatrix $-\left[\mathcal{A}(\vartheta_0)\right]^{-1}$, d.h. die Inverse der Fisher-Informationsmatrix, und somit ist die Effizienz des Maximum-Likelihood-Schätzers laut Cramér-Rao Ungleichung 2.4.14 gegeben.
2. In einem misspezifiziertem Modell erhalten wir immer noch einen Schätzer $\hat{\vartheta}_n$, der zu einem „vernünftigen“ ϑ_0 konvergiert.
Zudem ist $\sqrt{n} \cdot (\hat{\vartheta}_n - \vartheta_0)$ immer noch asymptotisch normalverteilt. Allerdings erhalten wir nicht mehr die optimale Kovarianzmatrix.

Beispielhaft wollen wir nun diese Sätze auf Maximum-Likelihood-Schätzer anwenden.

Remark 2.5.6

1. For maximum likelihood estimators and a correctly specified model, we receive $-\mathcal{A}(\vartheta_0) = \mathcal{B}(\vartheta_0) = \mathbb{I}(\mathcal{P}_{\vartheta_0})$ (see also Remark 2.4.11).
Therefore, we obtain for the covariance matrix $-\left[\mathcal{A}(\vartheta_0)\right]^{-1}$, i.e. the inverse of Fisher's information matrix, and this shows the efficiency of the maximum likelihood estimator by the Cramér-Rao inequality 2.4.14.
2. If the model is misspecified, then we still obtain an estimate $\hat{\vartheta}_n$ which converges to a “reasonable” ϑ_0 .
So, $\sqrt{n} \cdot (\hat{\vartheta}_n - \vartheta_0)$ is still asymptotically normally distributed. However, we do not get the optimal covariance matrix.

As an example, we want to apply these theorems to maximum likelihood estimators.

Satz 2.5.7 (Konsistenz von Maximum-Likelihood-Schätzern)

Seien X_1, \dots, X_n u.i. $\mathcal{P}_{\vartheta_0}$ -verteilt.
 $\mathcal{P}_{\vartheta_0} \in \{\mathcal{P}_{\vartheta} \mid \vartheta \in \Theta\}$, wobei $\Theta \subset \mathbb{R}^m$ kompakt sei und $\{\mathcal{P}_{\vartheta} \mid \vartheta \in \Theta\}$ die Glattheitsannahmen von Korollar 2.4.12 oder Bemerkung 2.4.13 erfülle.
Zusätzlich sei $\frac{\partial f_{\vartheta}}{\partial \vartheta}(x)$ als Funktion von $(\tilde{\vartheta}, x)$ beschränkt, wobei f_{ϑ} die zugehörige Dichte zu \mathcal{P}_{ϑ} bezeichnet, und $f_{\vartheta_0} \neq f_{\vartheta}$ für alle $\vartheta \in \Theta$ mit $\vartheta \neq \vartheta_0$.
Es gebe einen Maximum-Likelihood-Schätzer $\hat{\vartheta}_n$.
Dann gilt, dass $\hat{\vartheta}_n$ konsistent ist, d.h.

$$\hat{\vartheta}_n \xrightarrow[n \rightarrow \infty]{p} \vartheta_0.$$

Beweis:

Für Q_n nehmen wir die Log-Likelihood-Funktion:

$$Q_n(\mathbf{X}_n; \vartheta) = \sum_{i=1}^n f_{\vartheta}(X_i).$$

Q definieren wir als den normierten Erwartungswert von Q_n :

$$Q(\vartheta) = \mathcal{E}_{\vartheta_0} [f_{\vartheta}(X_1)] = \int_{\mathbb{R}} f_{\vartheta}(x) \cdot f_{\vartheta_0}(x) dx.$$

Es genügt nun, die Bedingungen von Satz 2.5.3 nachzuprüfen.

1. $\Theta \subset \mathbb{R}^m$ ist kompakt laut Annahme.

Theorem 2.5.7 (Consistency of maximum likelihood estimators)

Let X_1, \dots, X_n be i.i. $\mathcal{P}_{\vartheta_0}$ -distributed.
 $\mathcal{P}_{\vartheta_0} \in \{\mathcal{P}_{\vartheta} \mid \vartheta \in \Theta\}$, where $\Theta \subset \mathbb{R}^m$ is compact and $\{\mathcal{P}_{\vartheta} \mid \vartheta \in \Theta\}$ the smoothness assumptions of Corollary 2.4.12 or Remark 2.4.13 fulfills.
Additionally, let $\frac{\partial f_{\vartheta}}{\partial \vartheta}(x)$ as a function of $(\tilde{\vartheta}, x)$ be bounded, where f_{ϑ} denotes the density of \mathcal{P}_{ϑ} , and $f_{\vartheta_0} \neq f_{\vartheta}$ for all $\vartheta \in \Theta$ with $\vartheta \neq \vartheta_0$.
Let the maximum likelihood estimator $\hat{\vartheta}_n$ exist.
Then it holds that $\hat{\vartheta}_n$ is consistent, i.e.

Proof:

As Q_n , we take the log-likelihood function:

We define Q as the normed expectation of Q_n :

It is now sufficient to prove the assumptions of Theorem 2.5.3.

1. $\Theta \subset \mathbb{R}^m$ is compact according to the assumptions.

2. $Q_n(\mathbf{X}_n; \vartheta)$ ist stetig bzgl. ϑ und messbar bzgl. \mathbf{X}_n , da dies aus den Glattheitsannahmen an \mathcal{P}_ϑ bzw. f_ϑ direkt hervorgeht.

3. Das Gesetz der großen Zahlen zeigt, dass $\frac{1}{n} \cdot Q_n(\mathbf{X}_n; \vartheta) \xrightarrow[n \rightarrow \infty]{f.s.} Q(\vartheta)$ für jedes $\vartheta \in \Theta$. Allerdings benötigen wir diese Konvergenz uniform in ϑ !

Sei $k := \max_{(\vartheta, x) \in \Theta \times \mathbb{R}} \left| \frac{\partial f_\vartheta}{\partial \vartheta}(x) \right|$.
Sei $\varepsilon > 0$ und $0 < p < 1$ gegeben.

Für jedes $\vartheta \in \Theta$ sei $U(\vartheta)$ die offene Kugel mit Radius $\frac{\varepsilon}{2k}$ in Θ .

Zu dieser offenen Überdeckung von Θ gibt es wegen der Kompaktheit von Θ eine endliche Teilüberdeckung. Die Mittelpunkte dieser Teilüberdeckung bezeichnen wir mit $\tilde{\vartheta}_j$ und $j \in J$, wobei J die endliche Indexmenge ist.

$\#J$ soll die Kardinalität von J angeben.

Sei $\tilde{p} = 1 - \frac{1-p}{\#J}$.

Für jedes $\tilde{\vartheta}_j$ gibt es nun ein $n_{0,j}$, so dass für alle $n \geq n_{0,j}$ gilt

$$\mathcal{P} \left(\left| \frac{1}{n} \cdot Q_n(\mathbf{X}_n; \tilde{\vartheta}_j) - Q(\tilde{\vartheta}_j) \right| < \frac{\varepsilon}{2} \right) \geq \tilde{p}.$$

Die zugehörige Menge der $\omega \in \Omega$ nennen wir $\Omega_{j,n}$, d.h. $\mathcal{P}(\Omega_{j,n}) \geq \tilde{p}$.

Wir betrachten nun $\vartheta \in \Theta$ beliebig, aber fest und $\tilde{\vartheta}$ sei eines der $\tilde{\vartheta}_j$ mit $\|\vartheta - \tilde{\vartheta}\| < \frac{\varepsilon}{2k}$.

Für $n \geq \max\{n_{0,j} \mid j \in J\}$ und $\omega \in \cap_{j \in J} \Omega_{j,n}$ gilt nun

2. $Q_n(\mathbf{X}_n; \vartheta)$ is continuous w.r.t. ϑ and measurable w.r.t. \mathbf{X}_n , since this follows directly from the smoothness assumptions on \mathcal{P}_ϑ and f_ϑ , respectively.

3. The law of large numbers shows that $\frac{1}{n} \cdot Q_n(\mathbf{X}_n; \vartheta) \xrightarrow[n \rightarrow \infty]{a.s.} Q(\vartheta)$ for every $\vartheta \in \Theta$. However, we need this convergence uniform in ϑ !

Let $k := \max_{(\vartheta, x) \in \Theta \times \mathbb{R}} \left| \frac{\partial f_\vartheta}{\partial \vartheta}(x) \right|$.
Let $\varepsilon > 0$ and $0 < p < 1$ be given.

For every $\vartheta \in \Theta$, let $U(\vartheta)$ be the open ball with radius $\frac{\varepsilon}{2k}$ in Θ .

To this open cover of Θ , there is a finite sub-cover, due to the compactness of Θ . We denote the mid points of this sub-cover by $\tilde{\vartheta}_j$ and $j \in J$, where J is the finite index set. $\#J$ denotes the cardinality of J .

Let $\tilde{p} = 1 - \frac{1-p}{\#J}$.

For every $\tilde{\vartheta}_j$ there is now an $n_{0,j}$, such that for all $n \geq n_{0,j}$ it holds

The corresponding set of $\omega \in \Omega$ is called $\Omega_{j,n}$, i.e. $\mathcal{P}(\Omega_{j,n}) \geq \tilde{p}$.

Now we consider $\vartheta \in \Theta$ arbitrary, but fixed and $\tilde{\vartheta}$ is one of the $\tilde{\vartheta}_j$ with $\|\vartheta - \tilde{\vartheta}\| < \frac{\varepsilon}{2k}$.

For $n \geq \max\{n_{0,j} \mid j \in J\}$ and $\omega \in \cap_{j \in J} \Omega_{j,n}$ it now holds that

$$\begin{aligned} \left| \frac{1}{n} \cdot Q_n(\mathbf{X}_n(\omega); \vartheta) - Q(\vartheta) \right| &\leq \left| \frac{1}{n} \cdot Q_n(\mathbf{X}_n(\omega); \vartheta) - \frac{1}{n} \cdot Q_n(\mathbf{X}_n(\omega); \tilde{\vartheta}) \right| + \left| \frac{1}{n} \cdot Q_n(\mathbf{X}_n; \tilde{\vartheta}) - Q(\tilde{\vartheta}) \right| \\ &\leq \left[\frac{1}{n} \cdot \sum_{i=1}^n |f_\vartheta(X_i) - f_{\tilde{\vartheta}}(X_i)| \right] + \frac{\varepsilon}{2} \\ &< \left[\frac{1}{n} \sum_{i=1}^n k \cdot \frac{\varepsilon}{2 \cdot k} \right] + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Daraus folgt

$$\mathcal{P} \left(\left| \frac{1}{n} \cdot Q_n(\mathbf{X}_n; \vartheta) - Q(\vartheta) \right| < \varepsilon \right) \geq \mathcal{P}(\cap_{j \in J} \Omega_{j,n}) \geq 1 - \sum_{j \in J} (1 - \tilde{p}) = p.$$

Wir haben somit die uniforme Konvergenz in Wahrscheinlichkeit, da wir ein n_0 und eine Menge $\cap_{j \in J} \Omega_{j,n}$ haben, so dass für $n \geq n_0$ der Abstand von $\frac{1}{n} \cdot Q_n$ zu Q für alle ϑ klein ist.

Der Beweis basierte im Wesentlichen auf der Kompaktheit von Θ , die uns den Rückzug auf eine endliche Menge $\tilde{\vartheta}_j$ ermöglichte, und die Beschränktheit der Ableitung von f_ϑ , die uns erlaubte, von einem beliebigen ϑ zu einem $\tilde{\vartheta}_j$ ohne größere Abweichung übergehen zu können.

Der Rest war lediglich technisch.

4. Die Cauchy-Schwartzsche-Ungleichung ergibt

$$Q(\vartheta) < \left\{ \int_{\mathbb{R}} [f_\vartheta(x)]^2 dx \right\}^{\frac{1}{2}} \cdot \left\{ \int_{\mathbb{R}} [f_{\vartheta_0}(x)]^2 dx \right\}^{\frac{1}{2}}$$

für $\vartheta \neq \vartheta_0$, da $f_\vartheta \neq f_{\vartheta_0}$, und

From this follows

We have thus the uniform convergence in probability, since we have an n_0 and a set $\cap_{j \in J} \Omega_{j,n}$, such that for $n \geq n_0$ the distance from $\frac{1}{n} \cdot Q_n$ to Q is small for all ϑ .

The proof was mainly based on the compactness of Θ , which allowed us to restrict ourselves to a finite set $\tilde{\vartheta}_j$, and the boundedness of the derivative of f_ϑ , which made it possible for us to go from an arbitrary ϑ to $\tilde{\vartheta}_j$ with a small error only.

The rest was just technical.

4. The Cauchy Schwartz' inequality gives

$$Q(\vartheta_0) = \int_{\mathbb{R}} [f_{\vartheta_0}(x)]^2 dx.$$

■

Satz 2.5.8 (Asymptotische Normalität von Maximum-Likelihood-Schätzern)

Zusätzlich zu den Annahmen des Satzes 2.5.7 benötigen wir, dass f_{ϑ} dreimal stetig differenzierbar nach ϑ ist und die dritte Ableitung als Funktion von (ϑ, x) beschränkt ist. Dann gilt, dass

$$\sqrt{n} \cdot (\hat{\vartheta}_n - \vartheta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{I}(\mathcal{P}_{\vartheta_0})).$$

Beweis:

Da wir von einem richtig spezifizierten Modell ausgehen, ist nach Bemerkung 2.5.6

$$\mathcal{A}(\vartheta_0) = \mathcal{B}(\vartheta_0) = \mathbf{I}(\mathcal{P}_{\vartheta_0}).$$

Wiederum genügt es, die Voraussetzungen von Satz 2.5.5 nachzuprüfen:

1. Die Hesse-Matrix existiert, da f_{ϑ} zweimal stetig differenzierbar ist.

2. Sei $k := \max_{(\vartheta, x) \in \Theta \times \mathbb{R}} \left\| \frac{\partial^3 f_{\vartheta}}{\partial \vartheta^3} \right\|$, wobei $\frac{\partial^3 f_{\vartheta}}{\partial \vartheta^3}$ der Tensor dritter Stufe ist.

Da $\vartheta_n^* \xrightarrow{p} \vartheta_0$ und mit Hilfe des Gesetzes der großen Zahlen folgt nun

$$\begin{aligned} \frac{1}{n} \cdot \left(\frac{\partial^2 Q_n}{\partial \vartheta \partial \vartheta^T} \right) \Big|_{\vartheta_n^*} &= \frac{1}{n} \cdot \sum_{i=1}^n \frac{\partial^2 f_{\vartheta_n^*}(X_i)}{\partial \vartheta \partial \vartheta^T} = \frac{1}{n} \cdot \sum_{i=1}^n \left\{ \frac{\partial^2 f_{\vartheta_0}(X_i)}{\partial \vartheta \partial \vartheta^T} + k \cdot \|\vartheta_n^* - \vartheta_0\| \right\} \\ &\xrightarrow{p} \frac{1}{n} \cdot \sum_{i=1}^n \frac{\partial^2 f_{\vartheta_0}(X_i)}{\partial \vartheta \partial \vartheta^T} \xrightarrow{p} \mathcal{E} \left(\frac{\partial^2 f_{\vartheta_0}(X_1)}{\partial \vartheta \partial \vartheta^T} \right) \\ &= -\mathbf{I}(\mathcal{P}_{\vartheta_0}). \end{aligned}$$

■

Theorem 2.5.8 (Asymptotic normality of maximum likelihood estimators)

In addition to the assumptions of Theorem 2.5.7, we need that f_{ϑ} is three times continuously differentiable w.r.t. ϑ and that the third derivative is bounded as a function of (ϑ, x) .

Then it holds

Proof:

Since our starting point is a correct specified model, we have from Remark 2.5.6

$$\mathcal{A}(\vartheta_0) = \mathcal{B}(\vartheta_0) = \mathbf{I}(\mathcal{P}_{\vartheta_0}).$$

Once more, it is sufficient to prove the assumptions of Theorem 2.5.5:

1. The Hessian exists, since f_{ϑ} is twice continuously differentiable.

2. Let $k := \max_{(\vartheta, x) \in \Theta \times \mathbb{R}} \left\| \frac{\partial^3 f_{\vartheta}}{\partial \vartheta^3} \right\|$, where $\frac{\partial^3 f_{\vartheta}}{\partial \vartheta^3}$ is a tensor of the third order.

Since $\vartheta_n^* \xrightarrow{p} \vartheta_0$ and using the law of large numbers, it follows

3. Hier ist entscheidend, dass die Zufallsvektoren $\mathbf{Y}_i = \frac{\partial f_{\vartheta_0}}{\partial \vartheta}(X_i)$ ebenfalls u.i.v. mit Mittelwert Null und Kovarianzmatrix $\mathbf{I}(\mathcal{P}_{\vartheta_0})$ sind:

$$\mathcal{E}(\mathbf{Y}_1) = \mathcal{E} \left(\frac{\partial f_{\vartheta_0}}{\partial \vartheta}(X_1) \right) = \frac{\partial}{\partial \vartheta} \mathcal{E}(f_{\vartheta_0}(X_1)) = \frac{\partial}{\partial \vartheta} Q(\vartheta_0) = 0,$$

da ϑ_0 das eindeutige Maximum von Q ist.

3. Here it is crucial, that the random vectors $\mathbf{Y}_i = \frac{\partial f_{\vartheta_0}}{\partial \vartheta}(X_i)$ also are i.i.d. with mean zero and covariance matrix $\mathbf{I}(\mathcal{P}_{\vartheta_0})$:

where ϑ_0 is the unique maximum of Q .

$$\mathcal{V} \mathcal{A} \mathcal{R}(\mathbf{Y}_1) = \mathcal{E} \left\{ \left[\frac{\partial f_{\vartheta_0}}{\partial \vartheta}(X_1) - 0 \right] \cdot \left[\frac{\partial f_{\vartheta_0}}{\partial \vartheta}(X_1) - 0 \right]^T \right\} = \mathbf{I}(\mathcal{P}_{\vartheta_0}).$$

Nun können wir den zentralen Grenzwertsatz anwenden und erhalten

Now, we can use the central limit theorem and obtain

$$\frac{1}{\sqrt{n}} \cdot \left(\frac{\partial Q_n}{\partial \vartheta} \Big|_{\vartheta_0} \right) = \sqrt{n} \cdot \left(\frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{I}(\mathcal{P}_{\vartheta_0})).$$

■

■

3 Intervallschätzung

Interval estimation

Oft sind wir nicht nur an einer Punktschätzung des unbekanntes Parameters interessiert, sondern würden auch ein Intervall haben wollen, worin der wahre Parameter mit einer gewissen Wahrscheinlichkeit liegt. Dies führt zum Begriff des „Konfidenzintervalls“.

Oftentimes, we are not only interested in a point estimate of the unknown parameter, but would also like to have an interval, where the true parameter is located with a certain probability. This leads to the concept of a “confidence interval”.

Definition 3.0.9 (Konfidenzintervall)

Sei $\Theta \subseteq \mathbb{R}$ der Parameterraum und $0 < \gamma < 1$ eine vorgegebene Wahrscheinlichkeit.

Ein γ -Konfidenzintervall für den Parameter $\vartheta \in \Theta$ ist ein zufälliges Intervall

$[g(\mathbf{X}), h(\mathbf{X})] \subseteq \Theta$ aus dem Zufallsvektor

$\mathbf{X} = (X_1, \dots, X_n)$, mit der Eigenschaft

$$\mathcal{P}_{\vartheta} \{g(\mathbf{X}) \leq \vartheta \leq h(\mathbf{X})\} \geq \gamma \quad \forall \vartheta \in \Theta.$$

Definition 3.0.9 (Confidence interval)

Let $\Theta \subseteq \mathbb{R}$ be the parameter space and $0 < \gamma < 1$ a prespecified probability.

A γ -confidence interval for the parameter $\vartheta \in \Theta$ is a random interval $[g(\mathbf{X}), h(\mathbf{X})] \subseteq \Theta$,

determined by the random vector

$\mathbf{X} = (X_1, \dots, X_n)$ and with the property

Definition 3.0.10 (Konfidenzbereich)

Sei $\Theta \subseteq \mathbb{R}^d$ ($d > 1$) der Parameterraum und $0 < \gamma < 1$ eine vorgegebene Wahrscheinlichkeit.

Ein γ -Konfidenzbereich für $\vartheta \in \Theta$ ist eine zufällige Teilmenge $A(\mathbf{X}) \subseteq \Theta$ aus dem

Datenvektor \mathbf{X} mit der Eigenschaft

$$\mathcal{P}_{\vartheta} \{\vartheta \in A(\mathbf{X})\} \geq \gamma \quad \forall \vartheta \in \Theta.$$

Definition 3.0.10 (Confidence region)

Let $\Theta \subseteq \mathbb{R}^d$ ($d > 1$) be the parameter space and $0 < \gamma < 1$ a prespecified probability.

A γ -confidence region for $\vartheta \in \Theta$ is a random set $A(\mathbf{X}) \subseteq \Theta$, determined by the data vector \mathbf{X} and

with the property

Bemerkung 3.0.11

Wie das Konfidenzintervall konstruiert wird, hängt von der Situation ab. Manchmal ist ein einseitiges Intervall (z.B.

$\mathcal{P}_{\vartheta}(\vartheta \geq g(\mathbf{X})) \geq \gamma \quad \forall \vartheta \in \Theta$) passender als ein zweiseitiges Intervall.

Remark 3.0.11

How the confidence interval is constructed depends on the situation. Sometimes a one-sided confidence interval (e.g.

$\mathcal{P}_{\vartheta}(\vartheta \geq g(\mathbf{X})) \geq \gamma \quad \forall \vartheta \in \Theta$) is more appropriate than a two-sided interval.

Bemerkung 3.0.12

Tatsächlich ist die Wahrscheinlichkeit, dass ein zufälliges γ -Konfidenzintervall den wahren Parameter ϑ überdeckt mindestens γ .

Remark 3.0.12

In fact, the probability that a random γ -confidence interval covers the true parameter ϑ is at least γ .

Definition 3.0.13 (Quantil)

Sei $F(x)$, $-\infty < x < \infty$, eine Verteilungsfunktion.

Definition 3.0.13 (Quantile)

Let $F(x)$, $-\infty < x < \infty$, be a distribution function.

$$c := \inf\{s \mid F(s) \geq \gamma\}$$

ist das γ -Quantil von F .

is called the γ -quantile of F .

Bemerkung 3.0.14 (Median)

Der Median von F ist das 0,5-Quantil.

Remark 3.0.14 (Median)

The median of F is the 0.5-quantile.

3.1 Einige Verteilungen und nützliche Eigenschaften

Some distributions and properties needed

Wir führen einige Verteilungen und Eigenschaften ein, die wir im nächsten Abschnitt brauchen:

We introduce some distributions and properties, which we need in the next section:

Definition 3.1.1 (χ^2 -Verteilung)

Seien X_1, X_2, \dots u.i.v. und $\mathcal{N}(0, 1)$ -verteilt.
Dann ist die Zufallsvariable $Z = X_1^2 + \dots + X_n^2$ χ^2 -verteilt mit n Freiheitsgraden (mit χ_n^2 -Verteilung bezeichnet) und hat die Dichte

$$f_n(z) = 2^{-\frac{n}{2}} \cdot \frac{1}{\Gamma(\frac{n}{2})} \cdot z^{\frac{n}{2}-1} \cdot e^{-\frac{z}{2}}, \quad z \geq 0,$$

wo

where

$$\Gamma(x) = \int_0^{\infty} t^{x-1} \cdot e^{-t} dt$$

die Γ -Funktion ist und insbesondere $\Gamma(n+1) = n!$ für $n \in \mathbb{N}$.

is the Γ -function and especially $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$.

Einige χ^2 -Verteilungen werden in Abbildung 3.1.1 gezeigt.

Some χ^2 -distributions are shown in Figure 3.1.1.

Definition 3.1.2 (t-Verteilung)

Seien X_0, X_1, \dots, X_n u.i. $\mathcal{N}(0, 1)$ -verteilt.
Die Verteilung von

$$V = \frac{\sqrt{n} \cdot X_0}{\sqrt{X_1^2 + \dots + X_n^2}}$$

heißt Student- oder t-Verteilung mit n Freiheitsgraden (t_n -Verteilung).

Definition 3.1.2 (t-distribution)

Let X_0, X_1, \dots, X_n be i.i. $\mathcal{N}(0, 1)$ -distributed.
The distribution of

is called student's or t-distribution with n degrees of freedom (t_n -distribution).

Definition 3.1.1 (χ^2 -distribution)

Let X_1, X_2, \dots be i.i.d. and $\mathcal{N}(0, 1)$ -distributed.
Then the random variable $Z = X_1^2 + \dots + X_n^2$ is χ^2 -distributed with n degrees of freedom (denoted χ_n^2 -distribution) and has the density function

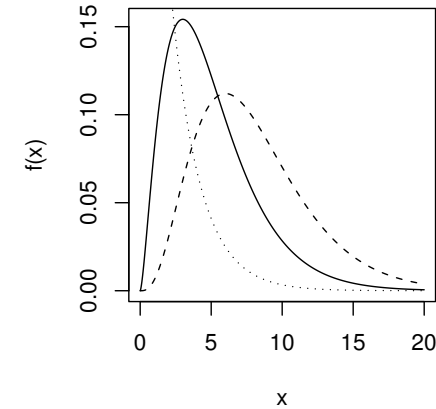


Figure 3.1.1: χ_n^2

Die Dichte der t_n -Verteilung wird in Abbildung 3.1.2 für verschiedene Anzahl von Freiheitsgraden gezeigt.

The density of the t_n -distribution is shown in Figure 3.1.2 for different degrees of freedom.

Definition 3.1.3 (F-Verteilung)

Seien X_1, \dots, X_n und Y_1, \dots, Y_m u.i. $\mathcal{N}(0, 1)$ -verteilt.
Die Verteilung von

$$U = \frac{m}{n} \cdot \frac{(X_1^2 + \dots + X_n^2)}{(Y_1^2 + \dots + Y_m^2)}$$

heißt F-Verteilung mit n und m Freiheitsgraden ($F_{n,m}$ -Verteilung).

Definition 3.1.3 (F-distribution)

Let X_1, \dots, X_n and Y_1, \dots, Y_m be i.i. $\mathcal{N}(0, 1)$ -distributed.
The distribution of

is called F-distribution with n and m degrees of freedom ($F_{n,m}$ -distribution).

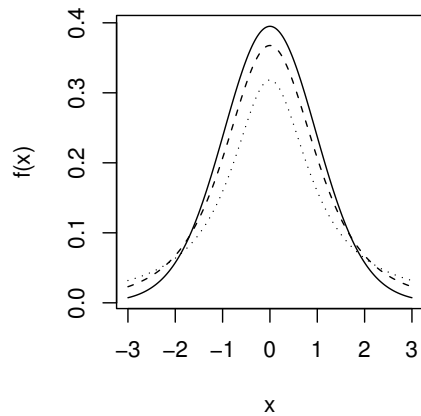


Figure 3.1.2: t_n

Die F-Verteilung wird in Abbildung 3.1.3 illustriert.

The F-distribution is illustrated in Figure 3.1.3.

Bemerkung 3.1.4

Wegen der Definition der F-Verteilung in 3.1.3, haben wir für die Quantile

Remark 3.1.4

Due to the definition of the F-distribution in 3.1.3, we have for the quantiles

$$F_{n,m}^{-1}(\alpha) = \frac{1}{F_{m,n}^{-1}(1-\alpha)}.$$

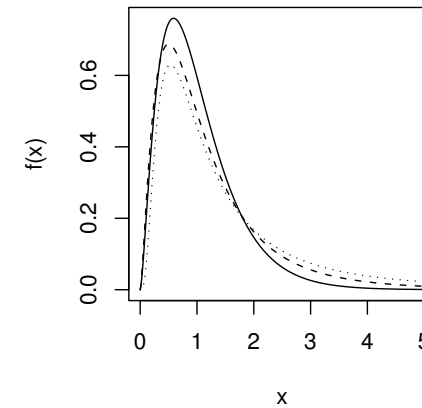


Figure 3.1.3: $F_{n,m}$

3.1.1 Die Dichten der χ^2 , t und F-Verteilung
The densities of the χ^2 , t and F-distribution

Bemerkung 3.1.5

In den folgenden Propositionen bezeichnet $\Gamma(x)$ die Γ -Funktion und $n, m \in \mathbb{N} \setminus \{0\}$.

Remark 3.1.5

In the following propositions, $\Gamma(x)$ denotes the Γ -function and $n, m \in \mathbb{N} \setminus \{0\}$.

Proposition 3.1.6

Die χ_n^2 -Verteilung hat die Dichte

Proposition 3.1.6

The χ_n^2 -distribution has the density function

$$f_n(z) = \frac{2^{-\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} \cdot z^{\frac{n}{2}-1} \cdot e^{-\frac{z}{2}}, \quad z > 0.$$

Beweis:

Zuerst zeigen wir, dass X_1^2 die Dichte $\frac{1}{\sqrt{2\pi z}} \cdot e^{-\frac{z}{2}}$ hat für $z > 0$, wobei $\mathcal{L}(X_1) = \mathcal{N}(0, 1)$:

$$\begin{aligned} \mathcal{P}(X_1^2 \leq t) &= \mathcal{P}(-\sqrt{t} \leq X_1 \leq \sqrt{t}) \\ &= 2 \cdot \int_0^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{s^2}{2}} ds \\ &= \frac{1}{\sqrt{2\pi}} \cdot \int_0^t \frac{1}{\sqrt{z}} \cdot e^{-\frac{z}{2}} dz. \end{aligned}$$

Wir wissen, dass die Summe $X + Y$ von unabhängigen Zufallsvariablen die Dichte $f * g(y) = \int_{-\infty}^{\infty} f(x) \cdot g(y-x) dx$ hat, wobei $*$ die Faltung bezeichnet und f und g die Dichten von X bzw. Y sind.

Per Induktion erhalten wir das gewünschte Ergebnis. ■

Proposition 3.1.7

Die t_n -Verteilung hat die Dichte

$$f_n(v) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi \cdot n} \cdot \Gamma(\frac{n}{2})} \cdot \left(1 + \frac{v^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < v < \infty.$$

Beweis:

Wenn V gleich t_n -verteilt ist, dann hat V die Form

$$V = \frac{\sqrt{n} \cdot X_0}{\sqrt{Z}},$$

Proof:

First we show that X_1^2 has the density function $\frac{1}{\sqrt{2\pi z}} \cdot e^{-\frac{z}{2}}$ for $z > 0$ where $\mathcal{L}(X_1) = \mathcal{N}(0, 1)$:

Now, we know that the sum $X + Y$ of independent random variables has the density function $f * g(y) = \int_{-\infty}^{\infty} f(x) \cdot g(y-x) dx$ where $*$ denotes the convolution and f and g are the densities of X and Y , respectively.

Induction now gives the desired result. ■

Proposition 3.1.7

The t_n -distribution has the density function

mit $\mathcal{L}(X_0) = \mathcal{N}(0, 1)$, $\mathcal{L}(Z) = \chi_n^2$ und X_0 und Z sind unabhängig.

Die gemeinsame Dichte von (X_0, Z) hat deswegen die Produktform:

$$g(x, z) = \text{const.} \cdot e^{-\frac{x^2}{2}} \cdot z^{\frac{n}{2}-1} \cdot e^{-\frac{z}{2}}, \quad -\infty < x < \infty, z > 0.$$

Es folgt

$$\begin{aligned} \mathcal{P}\left(\frac{\sqrt{n} \cdot X_0}{\sqrt{Z}} \leq t\right) &= \int \int_{\{(x,z) \in \mathbb{R}^2 \mid \sqrt{n} \cdot x \leq t \cdot \sqrt{z}\}} g(x, z) dx dz \\ &= \text{const.} \cdot \int_0^{\infty} z^{\frac{n}{2}-1} \cdot e^{-\frac{z}{2}} \cdot \left(\int_{-\infty}^{t \cdot \sqrt{\frac{z}{n}}} e^{-\frac{x^2}{2}} dx\right) dz \\ &= \text{const.} \cdot \int_0^{\infty} z^{\frac{n-1}{2}} \cdot e^{-\frac{z}{2}} \cdot \left(\int_{-\infty}^{\frac{t}{\sqrt{n}} \cdot s^2} e^{-\frac{1}{2} \cdot s^2 \cdot z} ds\right) dz \\ &= \text{const.} \cdot \int_{-\infty}^{\frac{t}{\sqrt{n}}} \int_0^{\infty} z^{\frac{n-1}{2}} \cdot e^{-\frac{1}{2} \cdot z \cdot (1+s^2)} dz ds \\ &= \text{const.} \cdot \int_{-\infty}^{\frac{t}{\sqrt{n}}} (1+s^2)^{-\frac{n+1}{2}} \cdot \left(\int_0^{\infty} u^{\frac{n-1}{2}} \cdot e^{-u} du\right) ds \\ &= \text{const.} \cdot \int_{-\infty}^t \left(1 + \frac{v^2}{n}\right)^{-\frac{n+1}{2}} dv, \end{aligned}$$

wobei const. verschiedene, geeignete Konstanten bezeichnet.

Die dritte Zeile erhält man durch die Substitution $s = x/\sqrt{z}$, die fünfte Zeile durch $u = z \cdot (1 + s^2)$ und die letzte Zeile durch $v = \sqrt{n} \cdot s$. ■

where $\mathcal{L}(X_0) = \mathcal{N}(0, 1)$, $\mathcal{L}(Z) = \chi_n^2$ and X_0 and Z are independent.

The joint density of (X_0, Z) therefore has the product form:

It follows:

where const. denotes different appropriate constants.

The third row is obtained by the substitution $s = x/\sqrt{z}$, the fifth row by $u = z \cdot (1 + s^2)$ and the last row by $v = \sqrt{n} \cdot s$. ■

Proposition 3.1.8

Die $F_{n,m}$ -Verteilung hat die Dichte

$$f_{n,m}(u) = \frac{\Gamma\left(\frac{n+m}{2}\right) \cdot \left(\frac{n}{m}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right) \cdot \Gamma\left(\frac{m}{2}\right)} \cdot \frac{u^{\frac{n}{2}-1}}{\left(1 + \frac{n}{m}u\right)^{\frac{n+m}{2}}}, \quad u > 0.$$

Beweis:

Wenn V gleich $F_{n,m}$ -verteilt ist, dann hat V die Form

$$V = \frac{m \cdot X}{n \cdot Z},$$

wobei $\mathcal{L}(X) = \chi_n^2$, $\mathcal{L}(Z) = \chi_m^2$ und X_0 und Z sind unabhängig.

Der Rest des Beweises ist ähnlich des Beweises von Proposition 3.1.7. ■

Proposition 3.1.8

The $F_{n,m}$ -distribution has the density function

Proof:

If V is $F_{n,m}$ -distributed, then V has the form

where $\mathcal{L}(X) = \chi_n^2$, $\mathcal{L}(Z) = \chi_m^2$ and X_0 and Z are independent.

The rest of the proof is now similar to the proof of Proposition 3.1.7. ■

3.1.2 Schätzer der Parameter einer Normalverteilung und ihre Verteilungen

Estimators of the parameters of a normal distribution and their distributions

Satz 3.1.9

Seien X_1, \dots, X_n u.i. $\mathcal{N}(\mu, \sigma^2)$ -verteilt
Dann gilt, dass

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

Theorem 3.1.9

Let X_1, \dots, X_n be i.i. $\mathcal{N}(\mu, \sigma^2)$ -distributed
Then it holds that

Beweis:

\bar{X}_n ist eine Linearkombination von normalverteilten Zufallsvariablen und demnach auch normalverteilt. Die anderen Terme sind lediglich Konstanten, d.h. ändern nichts an der Normalverteilungseigenschaft.

$$\begin{aligned} \mathcal{E}\left(\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma}\right) &= \frac{\sqrt{n}}{\sigma} \cdot \left\{ \mathcal{E}\left(\frac{1}{n} \cdot \sum_{i=1}^n X_i\right) - \mu \right\} = \frac{\sqrt{n}}{\sigma} \cdot \left\{ \frac{1}{n} \cdot \sum_{i=1}^n [\mathcal{E}(X_i) - \mu] \right\} = 0, \\ \mathcal{V}\mathcal{A}\mathcal{R}\left(\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma}\right) &= \frac{n}{\sigma^2} \cdot \mathcal{V}\mathcal{A}\mathcal{R}\left(\frac{1}{n} \cdot \sum_{i=1}^n X_i\right) \stackrel{COV(X_i, X_j) = \delta_{ij} \cdot \sigma^2}{=} 1. \end{aligned}$$

■

■

Proof:

\bar{X}_n is a linear combination of normally distributed random variables and thus also normally distributed. The other terms are just constants, i.e. they do not change the property to be normally distributed.

Satz 3.1.10

Seien X_1, \dots, X_n u.i. $\mathcal{N}(\mu, \sigma^2)$ -verteilt und
 $\hat{s}_n^2 = \frac{1}{n-1} \cdot \sum_{j=1}^n (X_j - \bar{X}_n)^2$ ein Schätzer für die Varianz σ^2 .
Es gilt nun

Theorem 3.1.10

Let X_1, \dots, X_n be i.i. $\mathcal{N}(\mu, \sigma^2)$ -distributed and
let $\hat{s}_n^2 = \frac{1}{n-1} \cdot \sum_{j=1}^n (X_j - \bar{X}_n)^2$ be an estimator of the variance σ^2 .
Then it follows that

$$\frac{(n-1) \cdot \hat{s}_n^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Beweis:

Sei $Y_i = X_i - \mu$.
Dann gilt

Proof:

Let $Y_i = X_i - \mu$.
Then it holds

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

Die gemeinsame Dichte von Y ist

The joint density function of Y is

$$f(y_1, \dots, y_n) = \frac{1}{(2 \cdot \pi \cdot \sigma^2)^{\frac{n}{2}}} \cdot \exp\left(-\frac{1}{2 \cdot \sigma^2} \cdot \sum_{i=1}^n y_i^2\right),$$

d.h. Y_1, \dots, Y_n sind ebenfalls unabhängig laut dem Faktorisierungssatz für Dichten.

Wir benutzen die Substitution $\mathbf{Z} = \mathbf{Q} \cdot \mathbf{Y}$, wobei Q eine $n \times n$ Drehmatrix ist, d.h. die transponierte Matrix Q^T ist die Inverse zu Q . Die erste Zeile von Q sei gleich $\left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$.

Dies ergibt

$$Z_1 = \frac{1}{\sqrt{n}} \cdot \sum_{i=1}^n Y_i = \sqrt{n} \cdot \bar{Y}_n, \quad (3.1.11)$$

$$\begin{aligned} \sum_{i=2}^n Z_i^2 &= \left(\sum_{i=1}^n Z_i^2 \right) - Z_1^2 = \mathbf{Z}^T \cdot \mathbf{Z} - Z_1^2 \\ &= \mathbf{Y}^T \cdot \mathbf{Q}^T \cdot \mathbf{Q} \cdot \mathbf{Y} - n \cdot \bar{Y}_n^2 = \left(\sum_{i=1}^n Y_i^2 \right) - n \cdot \bar{Y}_n^2 = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \\ &= (n-1) \cdot \hat{s}_n^2, \end{aligned} \quad (3.1.12)$$

$$\mathcal{E}(Z_i) = [\mathbf{Q} \cdot \mathcal{E}(\mathbf{Y})]_i = 0, \quad (3.1.13)$$

$$\mathcal{E}(\mathbf{Z} \cdot \mathbf{Z}^T) = \mathcal{E}(\mathbf{Q} \cdot \mathbf{Y} \cdot \mathbf{Y} \cdot \mathbf{Q}^T) = \mathbf{Q} \cdot [\mathcal{E}(\mathbf{Y} \cdot \mathbf{Y})] \cdot \mathbf{Q}^T = \mathbf{Q} \cdot \sigma^2 \cdot \text{Id}_{n \times n} \cdot \mathbf{Q}^T = \sigma^2 \cdot \text{Id}_{n \times n}. \quad (3.1.14)$$

Aus den vorherigen Relationen folgt für die gemeinsame Dichte der Z_i :

$$f_{\mathbf{Z}}(z_1, \dots, z_n) = \frac{1}{(2 \cdot \pi \cdot \sigma^2)^{\frac{n}{2}}} \cdot \exp\left(-\frac{1}{2 \cdot \sigma^2} \cdot \sum_{i=1}^n z_i^2\right). \quad (3.1.15)$$

Aus den Gleichungen 3.1.12 und 3.1.15 folgt nun unmittelbar

$$\frac{(n-1)\hat{s}_n^2}{\sigma^2} \sim \chi_{n-1}^2,$$

wie behauptet. ■

i.e. Y_1, \dots, Y_n are also independent due to the factorization theorem for densities.

We use the substitution $\mathbf{Z} = \mathbf{Q} \cdot \mathbf{Y}$, where Q is an $n \times n$ rotation matrix, i.e. the transposed matrix Q^T is the inverse of Q . The first row of Q should be $\left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$.

This leads to

The former considerations give us for the common density of the Z_i :

We receive automatically from equations 3.1.12 and 3.1.15 that

as claimed. ■

Korollar 3.1.16

Seien X_1, \dots, X_n u.i. $\mathcal{N}(\mu, \sigma^2)$ -verteilt. Dann folgt, dass \bar{X}_n und \hat{s}_n^2 unabhängig sind.

Beweis:

Das Korollar folgt sofort aus den Gleichungen 3.1.11, 3.1.12 und 3.1.15. ■

Satz 3.1.17

Seien X_1, \dots, X_n u.i. $\mathcal{N}(\mu, \sigma^2)$ -verteilt.

ist t_{n-1} -verteilt.

Beweis:

Der Satz folgt aus Satz 3.1.9, Satz 3.1.10 und Korollar 3.1.16 und der Definition der t -Verteilung. ■

**3.2 Konfidenzintervalle für die Parameter einiger gewöhnlichen Verteilungen
Confidence intervals for the parameters of some common distributions**

Wir sind nun in der Lage, approximative und exakte Konfidenzintervalle für die Parameter einiger gewöhnlichen Verteilungen zu berechnen.

Corollary 3.1.16

Let X_1, \dots, X_n be i.i. $\mathcal{N}(\mu, \sigma^2)$ -distributed. Then it follows that \bar{X}_n and \hat{s}_n^2 are independent.

Proof:

The corollary follows immediately from equations 3.1.11, 3.1.12 and 3.1.15. ■

Theorem 3.1.17

Let X_1, \dots, X_n be i.i. $\mathcal{N}(\mu, \sigma^2)$ -distributed.

$$\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\hat{s}_n}$$

is t_{n-1} -distributed.

Proof:

The theorem follows from Theorem 3.1.9, Theorem 3.1.10 and Corollary 3.1.16 and the definition of the t -distribution. ■

Satz 3.2.1

Seien X_1, \dots, X_n u.i. exponentialverteilt mit Parameter $\lambda \in \mathbb{R}^+$.
Dann ist $2 \cdot \lambda \cdot n \cdot \bar{X}_n$ gleich $\chi_{2,n}^2$ -verteilt.

Theorem 3.2.1

Let X_1, \dots, X_n be i.i. exponentially distributed with parameter $\lambda \in \mathbb{R}^+$.
Then $2 \cdot \lambda \cdot n \cdot \bar{X}_n$ is $\chi_{2,n}^2$ -distributed.

Beweis:

Aus

$$\mathcal{P}(2 \cdot \lambda \cdot X_j > t) = \mathcal{P}\left(X_j > \frac{t}{2 \cdot \lambda}\right) = \exp\left(-\frac{1}{2} \cdot t\right)$$

folgt, dass

$$2 \cdot \lambda \cdot X_j \sim \text{Exp}\left(\frac{1}{2}\right) = \chi_2^2$$

mit Hilfe von Proposition 3.1.6.

Aus der Unabhängigkeit der X_i folgt somit

$$\mathcal{L}(2 \cdot \lambda \cdot n \cdot \bar{X}_n) = \mathcal{L}\left(\sum_{j=1}^n 2 \cdot \lambda \cdot X_j\right) = \chi_{2,n}^2.$$

Proof:

From

follows that

with help of Proposition 3.1.6.

From the independence of the X_i follows

Beispiel 3.2.2 (Exaktes Konfidenzintervall für den Parameter λ der Exponentialverteilung)

Wir nehmen an, dass X_1, \dots, X_n u.i. exponentialverteilt mit Parameter λ sind.
Als Schätzer für λ verwenden wir den Maximum-Likelihood-Schätzer $\hat{\lambda} = \frac{1}{\bar{X}_n}$.

Example 3.2.2 (Exact confidence interval for the parameter λ of the exponential distribution)

Let us assume that X_1, \dots, X_n are i.i. exponentially distributed with parameter λ .
We use the maximum likelihood estimator $\hat{\lambda} = \frac{1}{\bar{X}_n}$ for the parameter λ .

Unser Ziel ist es, ein exaktes γ -Konfidenzintervall für λ zu konstruieren.

Von Satz 3.2.1 wissen wir, dass $2 \cdot \lambda \cdot n \cdot \bar{X}_n$ gleich $\chi_{2,n}^2$ -verteilt ist.
Damit erhalten wir

$$\mathcal{P}_\lambda\left(\chi_{2,n, \frac{1-\gamma}{2}}^2 \leq 2 \cdot \lambda \cdot n \cdot \bar{X}_n \leq \chi_{2,n, 1-\frac{1-\gamma}{2}}^2\right) = \mathcal{P}_\lambda\left(\frac{\chi_{2,n, \frac{1-\gamma}{2}}^2}{2 \cdot n \cdot \bar{X}_n} \leq \lambda \leq \frac{\chi_{2,n, 1-\frac{1-\gamma}{2}}^2}{2 \cdot n \cdot \bar{X}_n}\right) = \gamma,$$

wobei $\chi_{2,n, \frac{1-\gamma}{2}}^2$ das $\frac{1-\gamma}{2}$ -Quantil und $\chi_{2,n, 1-\frac{1-\gamma}{2}}^2$ das $(1 - \frac{1-\gamma}{2})$ -Quantil der $\chi_{2,n}^2$ -Verteilung sind.
Somit bekommen wir das Konfidenzintervall

$$\left[\frac{\chi_{2,n, \frac{1-\gamma}{2}}^2}{2 \cdot n \bar{X}_n}; \frac{\chi_{2,n, 1-\frac{1-\gamma}{2}}^2}{2 \cdot n \bar{X}_n}\right].$$

Unser nächstes Ziel ist es, ein exaktes Konfidenzintervall für die Parameter einer Normalverteilung zu bestimmen. Wir untersuchen vier verschiedene Fälle, die sich darin unterscheiden, welche von den Parametern μ und σ^2 wir als bekannt betrachten. Wir beginnen mit der Aufgabe, Konfidenzintervalle für μ zu finden:

Our aim is to construct an exact γ -confidence interval for λ .

Due to Theorem 3.2.1 we know that $2 \cdot \lambda \cdot n \cdot \bar{X}_n$ is $\chi_{2,n}^2$ -distributed.
Thus, we arrive at

where $\chi_{2,n, \frac{1-\gamma}{2}}^2$ is the $\frac{1-\gamma}{2}$ -quantile and $\chi_{2,n, 1-\frac{1-\gamma}{2}}^2$ the $(1 - \frac{1-\gamma}{2})$ -quantile of the $\chi_{2,n}^2$ -distribution.
So, we obtain the confidence interval

Our next objective is to determine exact confidence intervals for the parameters of the normal distribution. We arrive at four different cases, depending on which of the parameters μ and σ^2 we regard as known. We start with the task to find confidence intervals for μ :

Beispiel 3.2.3 (Exaktes Konfidenzintervall für den Mittelwert einer Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ falls σ bekannt ist)

Seien X_1, \dots, X_n u.i. $\mathcal{N}(\mu, \sigma^2)$ -verteilte Zufallsvariablen. Wir nehmen an, dass σ^2 bekannt ist und dass wir ein γ -Konfidenzintervall für μ finden wollen. Von Satz 3.1.9 haben wir, dass

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

Somit haben wir

$$\begin{aligned} \mathcal{P}_\mu \left(\bar{X}_n - u_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + u_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) &= \mathcal{P}_\mu \left(-u_{\frac{1+\gamma}{2}} \leq \frac{\sqrt{n}}{\sigma} \cdot (\bar{X}_n - \mu) \leq u_{\frac{1+\gamma}{2}} \right) \\ &= \Phi(u_{\frac{1+\gamma}{2}}) - \Phi(-u_{\frac{1+\gamma}{2}}) = \gamma, \end{aligned}$$

wobei $u_{\frac{1+\gamma}{2}}$ das $\frac{1+\gamma}{2}$ -Quantil einer $\mathcal{N}(0, 1)$ -verteilten Zufallsvariable ist. Also haben wir das Konfidenzintervall

$$I_\mu = \left[\bar{X}_n - u_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} ; \bar{X}_n + u_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

für μ mit Niveau γ .

Example 3.2.3 (Exact confidence interval for the mean of a normal distribution $\mathcal{N}(\mu, \sigma^2)$ if σ is known)

Let X_1, \dots, X_n be i.i. $\mathcal{N}(\mu, \sigma^2)$ -distributed random variables. We assume that σ^2 is known and that we want to find a γ -confidence interval for μ . From Theorem 3.1.9, we have that

So, we have

where $u_{\frac{1+\gamma}{2}}$ is the $\frac{1+\gamma}{2}$ -quantile of a $\mathcal{N}(0, 1)$ -distributed random variable. Thus, we have the confidence interval

for μ with confidence level γ .

Beispiel 3.2.4 (Exaktes Konfidenzintervall für den Mittelwert einer Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ falls σ unbekannt ist)

Seien X_1, \dots, X_n u.i. $\mathcal{N}(\mu, \sigma^2)$ -verteilte Zufallsvariablen. In diesem Beispiel nehmen wir an, dass σ^2 unbekannt ist und dass wir ein γ -Konfidenzintervall für μ finden wollen. Wir schätzen σ^2 durch

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

Von Satz 3.1.17 haben wir, dass

$$\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\hat{\sigma}_n} \sim t_{n-1}.$$

Damit haben wir

$$\begin{aligned} \mathcal{P}_\mu \left(\bar{X}_n - t_{n-1, \frac{1+\gamma}{2}} \cdot \frac{\hat{\sigma}_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{n-1, \frac{1+\gamma}{2}} \cdot \frac{\hat{\sigma}_n}{\sqrt{n}} \right) &= \mathcal{P}_\mu \left(-t_{n-1, \frac{1+\gamma}{2}} \leq \frac{\sqrt{n}}{\hat{\sigma}_n} \cdot (\bar{X}_n - \mu) \leq t_{n-1, \frac{1+\gamma}{2}} \right) \\ &= \frac{1+\gamma}{2} - \left(1 - \frac{1+\gamma}{2} \right) = \gamma, \end{aligned}$$

wobei $t_{n-1, \frac{1+\gamma}{2}}$ das $\frac{1+\gamma}{2}$ -Quantil einer t_{n-1} -verteilten Zufallsvariable bezeichnet. Somit ist das exakte Konfidenzintervall

$$I_\mu = \left[\bar{X}_n - t_{n-1, \frac{1+\gamma}{2}} \cdot \frac{\hat{\sigma}_n}{\sqrt{n}} ; \bar{X}_n + t_{n-1, \frac{1+\gamma}{2}} \cdot \frac{\hat{\sigma}_n}{\sqrt{n}} \right].$$

Example 3.2.4 (Exact confidence interval for the mean of a normal distribution $\mathcal{N}(\mu, \sigma^2)$ if σ is unknown)

Let X_1, \dots, X_n be i.i. $\mathcal{N}(\mu, \sigma^2)$ -distributed random variables. In this example, we assume that σ^2 is unknown and that we want to find a γ -confidence interval for μ . We estimate σ^2 by

From Theorem 3.1.17, we have that

Then we have

where $t_{n-1, \frac{1+\gamma}{2}}$ denotes the $\frac{1+\gamma}{2}$ -quantile of a t_{n-1} -distributed random variable. Thus, the exact confidence interval is

Bemerkung 3.2.5

In dem Beispiel mit unbekanntem σ , haben wir, im Vergleich mit dem Fall bekannter Varianz, σ mit der Schätzung \hat{s}_n und die Normalverteilung mit einer t_{n-1} -Verteilung ersetzt.

Jetzt betrachten wir in zwei Beispielen Konfidenzintervalle für σ^2 :

Beispiel 3.2.6 (Exaktes Konfidenzintervall für die Varianz einer Normalverteilung)

$\mathcal{N}(\mu, \sigma^2)$ falls μ bekannt ist)

Seien X_1, \dots, X_n u.i. $\mathcal{N}(\mu, \sigma^2)$ -verteilte Zufallsvariablen. Wir nehmen an, dass μ bekannt ist und dass wir ein γ -Konfidenzintervall für σ^2 finden wollen.

Sei $\hat{\sigma}_n = \sum_{i=1}^n (X_i - \mu)^2$. Die Zufallsvariable

$\frac{1}{\sigma^2} \cdot \hat{\sigma}_n$ ist die Summe von n unabhängigen, quadrierten $\mathcal{N}(0, 1)$ -verteilten Zufallsvariablen und ist damit χ_n^2 -verteilt, d.h.

$$\begin{aligned} \mathcal{P}_{\sigma^2} \left(\frac{\hat{\sigma}_n}{\chi_{n, \frac{1+\gamma}{2}}^2} \leq \sigma^2 \leq \frac{\hat{\sigma}_n}{\chi_{n, \frac{1-\gamma}{2}}^2} \right) &= \mathcal{P}_{\sigma^2} \left(\chi_{n, \frac{1-\gamma}{2}}^2 \leq \frac{\hat{\sigma}_n}{\sigma^2} \leq \chi_{n, \frac{1+\gamma}{2}}^2 \right) \\ &= \frac{1+\gamma}{2} - \frac{1-\gamma}{2} = \gamma, \end{aligned}$$

wobei $\chi_{n, \frac{1+\gamma}{2}}^2$ und $\chi_{n, \frac{1-\gamma}{2}}^2$ die $\frac{1+\gamma}{2}$ - bzw.

$\frac{1-\gamma}{2}$ -Quantile einer χ_n^2 -verteilten Zufallsvariable sind.

$$\Rightarrow I_{\sigma^2} = \left[\frac{\hat{\sigma}_n}{\chi_{n, \frac{1+\gamma}{2}}^2}; \frac{\hat{\sigma}_n}{\chi_{n, \frac{1-\gamma}{2}}^2} \right]$$

Remark 3.2.5

In the example with σ unknown, we replaced σ with the estimation \hat{s}_n and the normal distribution with a t_{n-1} -distribution, in comparison with the case where σ is known.

Now, we consider confidence intervals for σ^2 in two examples:

Example 3.2.6 (Exact confidence interval for the variance of a normal distribution)

$\mathcal{N}(\mu, \sigma^2)$ if μ is known)

Let X_1, \dots, X_n be i.i. $\mathcal{N}(\mu, \sigma^2)$ -distributed random variables. We assume that μ is known and that we want to find a γ -confidence interval for σ^2 .

Let $\hat{\sigma}_n = \sum_{i=1}^n (X_i - \mu)^2$. The random variable

$\frac{1}{\sigma^2} \cdot \hat{\sigma}_n$ is the sum of n independent quadrates of $\mathcal{N}(0, 1)$ -distributed random variables and hence χ_n^2 -distributed, i.e.

where $\chi_{n, \frac{1+\gamma}{2}}^2$ and $\chi_{n, \frac{1-\gamma}{2}}^2$ are the $\frac{1+\gamma}{2}$ - and the

$\frac{1-\gamma}{2}$ -quantiles, respectively, of a χ_n^2 -distributed random variable.

ist das Konfidenzintervall für σ^2 mit Niveau γ .

Beispiel 3.2.7 (Exaktes Konfidenzintervall für die Varianz einer Normalverteilung)

$\mathcal{N}(\mu, \sigma^2)$ wenn μ unbekannt ist)

Seien X_1, \dots, X_n u.i. $\mathcal{N}(\mu, \sigma^2)$ -verteilte Zufallsvariablen. Wir nehmen an, dass μ unbekannt ist. Das Ziel ist es, ein

γ -Konfidenzintervall für σ^2 zu konstruieren.

Sei

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

Gemäß Satz 3.1.10 ist $\frac{n-1}{\sigma^2} \cdot \hat{s}_n^2$ dann

χ_{n-1}^2 -verteilt.

Dies ergibt

$$\begin{aligned} \mathcal{P}_{\sigma^2} \left(\frac{(n-1) \cdot \hat{s}_n^2}{\chi_{n-1, \frac{1+\gamma}{2}}^2} \leq \sigma^2 \leq \frac{(n-1) \cdot \hat{s}_n^2}{\chi_{n-1, \frac{1-\gamma}{2}}^2} \right) &= \mathcal{P}_{\sigma^2} \left(\chi_{n-1, \frac{1-\gamma}{2}}^2 \leq \frac{(n-1) \cdot \hat{s}_n^2}{\sigma^2} \leq \chi_{n-1, \frac{1+\gamma}{2}}^2 \right) \\ &= \frac{1+\gamma}{2} - \left(1 - \frac{1-\gamma}{2} \right) = \gamma, \end{aligned}$$

wobei $\chi_{n-1, \frac{1+\gamma}{2}}^2$ und $\chi_{n-1, \frac{1-\gamma}{2}}^2$ die $\frac{1+\gamma}{2}$ - bzw.

$\frac{1-\gamma}{2}$ -Quantile einer χ_{n-1}^2 -verteilten Zufallsvariablen sind.

$$\Rightarrow I_{\sigma^2} = \left[\frac{(n-1) \cdot \hat{s}_n^2}{\chi_{n-1, \frac{1+\gamma}{2}}^2}; \frac{(n-1) \cdot \hat{s}_n^2}{\chi_{n-1, \frac{1-\gamma}{2}}^2} \right]$$

ist das Konfidenzintervall für σ^2 mit Niveau γ .

is the confidence interval for σ^2 with level γ .

Example 3.2.7 (Exact confidence interval for the variance of a normal distribution)

$\mathcal{N}(\mu, \sigma^2)$ if μ is unknown)

Let X_1, \dots, X_n be i.i. $\mathcal{N}(\mu, \sigma^2)$ -distributed random variables. We assume that μ is unknown. Now we are going to determine a γ -confidence interval for σ^2 .

Let

According to Theorem 3.1.10 is $\frac{n-1}{\sigma^2} \cdot \hat{s}_n^2$ then

χ_{n-1}^2 -distributed.

This yields

where $\chi_{n-1, \frac{1+\gamma}{2}}^2$ and $\chi_{n-1, \frac{1-\gamma}{2}}^2$ are the $\frac{1+\gamma}{2}$ - and

the $\frac{1-\gamma}{2}$ -quantiles of a χ_{n-1}^2 -distributed random variable.

Nach all diesen exakten Konfidenzintervallen kommen wir nun zu approximativen Konfidenzintervallen.

After all these exact confidence intervals, we now come to approximate confidence intervals.

Beispiel 3.2.8 (Approximatives Konfidenzintervall für den Parameter p der Binomialverteilung)

Seien X_1, \dots, X_n u.i. binomialverteilt $\mathcal{B}(1, p)$ mit Parameter p , $0 < p < 1$. Dann ist $Y = X_1 + \dots + X_n$ nun $\mathcal{B}(n, p)$ -verteilt. Wenn n groß ist, können wir Satz 1.1.51 anwenden, d.h. die Zufallsvariable

$$\frac{Y - n \cdot p}{\sqrt{n \cdot p \cdot (1 - p)}}$$

ist approximativ $\mathcal{N}(0, 1)$ -verteilt. Demnach haben wir für $c > 0$ approximativ:

$$\mathcal{P}_p \left(-c \leq \frac{Y - n \cdot p}{\sqrt{n \cdot p \cdot (1 - p)}} \leq c \right) \approx \Phi(c) - \Phi(-c) = 2 \cdot \Phi(c) - 1,$$

wobei Φ die Verteilungsfunktion einer Standardnormalverteilung ist. Wenn wir ein Konfidenzintervall mit Niveau γ wollen, bestimmen wir c durch

$$2 \cdot \Phi(c) - 1 \stackrel{!}{=} \gamma.$$

Wenn $c = u_{\frac{1+\gamma}{2}}$ das $\frac{1+\gamma}{2}$ -Quantil einer $\mathcal{N}(0, 1)$ -verteilten Zufallsvariable ist, dann bekommen wir approximativ

$$\mathcal{P}_p \left(-u_{\frac{1+\gamma}{2}} \leq \frac{Y - n \cdot p}{\sqrt{n \cdot p \cdot (1 - p)}} \leq u_{\frac{1+\gamma}{2}} \right) \approx \gamma$$

Example 3.2.8 (Approximate confidence interval for the parameter p of the binomial distribution)

Let X_1, \dots, X_n be i.i. binomially distributed $\mathcal{B}(1, p)$ with parameter p , $0 < p < 1$. Then $Y = X_1 + \dots + X_n$ is $\mathcal{B}(n, p)$ -distributed. If n is big, then we can use Theorem 1.1.51 and therefore the random variable

is approximately $\mathcal{N}(0, 1)$ -distributed. Thus, for $c > 0$ we have approximately:

where Φ is the distribution function of the standard normal distribution. If we want the confidence level γ , we determine c as

If $c = u_{\frac{1+\gamma}{2}}$ is the $\frac{1+\gamma}{2}$ -quantile of a $\mathcal{N}(0, 1)$ -distributed random variable, then we approximately obtain

für jedes p mit $0 < p < 1$.

Wir benutzen $\bar{p}_n = \frac{Y}{n}$ als Schätzer für p . Dann bekommen wir folgendes approximative Konfidenzintervall für p :

$$\mathbb{I}_p = \left[\bar{p}_n - u_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\bar{p}_n \cdot (1 - \bar{p}_n)}{n}}; \bar{p}_n + u_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\bar{p}_n \cdot (1 - \bar{p}_n)}{n}} \right].$$

Beispiel 3.2.9

Wir erinnern uns an Beispiel 1.0.4 auf Seite 2. Angenommen, dass die wandernden StudentInnen insgesamt 300 wilde Tiere während ihrer Wanderung zum Humberturm und wieder zurück zur Universität beobachten. Außerdem nehmen sie an, dass die Anzahl der Elwedritschen unter den beobachteten Tieren binomialverteilt, $\mathcal{B}(300, p)$, ist. Jetzt wollen die StudentInnen ein 95% Konfidenzintervall für p finden, wobei p die Wahrscheinlichkeit ist, dass ein beobachtetes Tier wirklich eine Elwedritsche ist. Sie denken, dass sie 4 Elwedritsche gesehen haben. Also ist ein approximatives Konfidenzintervall für p mit Niveau 0,95:

$$\mathbb{I}_p = \left[\frac{4}{300} - 1,96 \cdot \sqrt{\frac{\frac{4}{300} \cdot (1 - \frac{4}{300})}{300}}; \frac{4}{300} + 1,96 \cdot \sqrt{\frac{\frac{4}{300} \cdot (1 - \frac{4}{300})}{300}} \right] = [0,0004, 0,0263].$$

Die StudentInnen können einen Wert für p zwischen 0,04% und 2,63% erwarten. (Unglücklicherweise stellt sich heraus, dass diese vier Tiere Wildschweine waren und keine Elwedritsche.)

for every p with $0 < p < 1$.

We use $\bar{p}_n = \frac{Y}{n}$ as an estimator for p . Then we obtain the following approximate confidence interval for p :

Example 3.2.9

We recall Example 1.0.4 on page 2. Assume that the hiking students totally observe 300 wild animals during their hike from the university to the Humbert tower and back again. Furthermore they suppose that the number of elwedritsche among the observed animals is binomially distributed, $\mathcal{B}(300, p)$. Now, the students want to find a 95% confidence interval for p , where p is the probability that an observed animal really is an elwedritsche. They think that they saw 4 elwedritsche. Thus, an approximate confidence interval with level 0.95 is:

The students can expect a value for p between 0.04% and 2.63%. (Unfortunately, it turns out that these four animals were wild boars and not elwedritsche.)

Wir beenden diesen Abschnitt mit einem approximativen Konfidenzintervall in einer relativ allgemeinen Situation:

We conclude this section by determining an approximate confidence interval for a rather general situation:

Beispiel 3.2.10 (Approximatives Konfidenzintervall für ϑ , falls $\hat{\vartheta}$ asymptotisch normalverteilt ist)

Example 3.2.10 (Approximate confidence interval for ϑ if $\hat{\vartheta}$ is asymptotically normally distributed)

Im allgemeinen können wir ein approximatives Konfidenzintervall für einen Parameter ϑ bestimmen, wenn wir einen asymptotisch normalverteilten Schätzer $\hat{\vartheta}$ Schätzer haben, für den die asymptotische Varianz bestimmt werden kann. Zum Beispiel haben wir einen solchen Schätzer, wenn der Maximum-Likelihood-Schätzer asymptotisch normalverteilt ist.

In general, we can determine an approximate confidence interval for a parameter ϑ , if we have an asymptotically normally distributed estimator $\hat{\vartheta}$ whose asymptotic variance can be determined. For example, we have such an estimator if the maximum likelihood estimator is asymptotically normally distributed.

Wir gehen also von der Situation in Satz 2.5.8 aus, d.h.

We therefore start with the situation of Theorem 2.5.8, i.e.

$$\sqrt{n \cdot \mathbb{I}(\mathcal{P}_{\vartheta})} \cdot (\hat{\vartheta}_n - \vartheta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Deswegen erhalten wir das folgende approximative Konfidenzintervall mit Niveau γ :

Therefore we obtain the following approximate confidence interval with level γ :

$$\mathbb{I}_{\vartheta} = \left[\hat{\vartheta}_n - \frac{u_{\frac{1+\gamma}{2}}}{\sqrt{n \cdot \mathbb{I}(\hat{\mathcal{P}})}}; \hat{\vartheta}_n + \frac{u_{\frac{1+\gamma}{2}}}{\sqrt{n \cdot \mathbb{I}(\hat{\mathcal{P}})}} \right],$$

wobei $u_{\frac{1+\gamma}{2}}$ das $\frac{1+\gamma}{2}$ -Quantil einer $\mathcal{N}(0, 1)$ -verteilten Zufallsvariable und $\mathbb{I}(\hat{\mathcal{P}})$ die geschätzte Fisher-Information ist.

where $u_{\frac{1+\gamma}{2}}$ is the $\frac{1+\gamma}{2}$ -quantile of a $\mathcal{N}(0, 1)$ -distributed random variable and $\mathbb{I}(\hat{\mathcal{P}})$ is the estimated Fisher's information.

4 Testtheorie

Test theory

4.1 Eine allgemeine Beschreibung von Hypothesentests

A general description of hypothesis testing

In der Testtheorie ist unser Ziel zu erkennen, ob eine gegebene Hypothese wahr oder falsch ist. Im allgemeinen haben wir ein Modell für unsere Daten, in dem ein unbekannter Parameter ϑ aus der Parametermenge Θ vorkommt. Die Hypothese besagt nun im allgemeinen, dass ϑ ein Element einer Teilmenge Θ_0 von Θ ist. Dies bedeutet, dass die Hypothese wahr ist, wenn $\vartheta \in \Theta_0$, aber falsch wenn $\vartheta \notin \Theta_0$.

In test theory, our aim is to conclude about the truth or falsehood of a given hypothesis. In general, we have a model for our data including an unknown parameter ϑ from the parameter space Θ . The hypothesis states in general that ϑ is an element of some subset Θ_0 of Θ . This means that the hypothesis is true if $\vartheta \in \Theta_0$, but false if $\vartheta \notin \Theta_0$.

Beispiel 4.1.1

Die wandernden MathematikstudentInnen kommen zu einer Kreuzung im Wald und sind nicht sicher welchen Pfad sie nehmen sollen. Sie werfen eine Münze und fangen sofort an über Hypothesentests zu diskutieren: Die Hypothese ist, dass es eine faire Münze ist, d.h. dass beide Seiten mit gleicher Wahrscheinlichkeit erscheinen. Somit ist $\Theta_0 = \{\frac{1}{2}\}$, das nur aus einem Element der Parametermenge $\Theta = [0, 1]$ besteht. Das Komplement von Θ_0 in Θ wird mit Θ_1 bezeichnet. Als Konvention nennen wir die ursprüngliche Hypothese ($\vartheta \in \Theta_0$) die Nullhypothese H_0 . Daneben gibt es noch die Alternative $\vartheta \in \Theta_1$, die mit H_1 bezeichnet wird.

Example 4.1.1

The hiking mathematics students come to a cross-road in the forest and are not sure which path they shall take. They toss a coin and immediately start to discuss about hypothesis testing: The hypothesis is that it is a fair coin, i.e. both sides show up with equal probability. Hence, $\Theta_0 = \{\frac{1}{2}\}$ contains just one element of the parameter space $\Theta = [0, 1]$. The complement of Θ_0 in Θ is denoted by Θ_1 . As a convention, we call the original hypothesis ($\vartheta \in \Theta_0$) the null hypothesis H_0 . Additionally, there is the alternative that $\vartheta \in \Theta_1$ which is denoted by H_1 .

Definition 4.1.2 (Test)

Ein Test wird durchgeführt, indem eine Teststatistik, mit bekannter Verteilung unter der Nullhypothese H_0 , betrachtet wird. Diese Teststatistik wird benutzt, um die Wahrheit der Nullhypothese, im Gegensatz zur Alternative H_1 , zu untersuchen.

Beispiel 4.1.3

In einem Experiment werden Daten gesammelt um zwei Messgeräte zu vergleichen (altes gegen neues).

Die Nullhypothese sei:

H_0 : kein Unterschied zwischen den Messgeräten.

Die Alternativhypothesen könnten sein:

- a) H_1 : sie sind unterschiedlich (zweiseitig);*
oder
- b) H_1 : das Neue ist besser (einseitig);*
oder
- c) H_1 : das Alte ist besser (einseitig).*

Definition 4.1.2 (Test)

A test is conducted using a test statistic whose distribution is known under the null hypothesis H_0 . This test statistic is used to consider the truth of the null hypothesis in contrast to the alternative H_1 .

Example 4.1.3

In an experiment, data are collected to compare two measuring instruments (old versus new).

We take for the null hypothesis:

H_0 : no difference between the measuring instruments.

The alternative hypothesis might be:

- a) H_1 : they are different (two-sided);*
or
- b) H_1 : the new one is better (one-sided);*
or
- c) H_1 : the old one is better (one-sided).*

Angenommen, wir haben zufällige Daten $\mathbf{X} = (X_1, \dots, X_n)$ und die Teststatistik $T(\mathbf{X})$.

Im Voraus bestimmen wir den Annahmebereich C_0 und den Ablehn- oder kritischen Bereich C_1 , d.h. gilt $T(\mathbf{X}) \in C_0$, so akzeptieren wir H_0 , und für $T(\mathbf{X}) \in C_1$ lehnen wir H_0 ab.

Klarerweise muss $C_0 \cup C_1$ gleich der Bildmenge der Teststatistik und $C_0 \cap C_1 = \emptyset$.

Es zwei mögliche Fehlerarten in der Testtheorie:

Definition 4.1.4 (Fehler erster/zweiter Art)

Der Fehler erster Art ist die Ablehnung der Nullhypothese H_0 , obwohl sie wahr ist.

Der Fehler zweiter Art ist die Ablehnung der Alternative H_1 , obwohl sie wahr ist, bzw. fälschliche Annahme der Nullhypothese.

	H_0 wahr	H_0 falsch
H_0 akzeptiert	Richtig	Fehler 2.Art
H_0 abgelehnt	Fehler 1.Art	Richtig

Bemerkung 4.1.5

Wir können bei Hypothesentests wie bei einem Gerichtsfall denken:

H_0 : Der Angeklagte ist unschuldig.

Der Fehler erster Art entspricht, mit anderen Worten, einen unschuldigen Angeklagten für schuldig zu erklären. Andererseits, bedeutet der Fehler zweiter Art, dass wir einen Schuldigen freisprechen.

Let us assume that we have the random data $\mathbf{X} = (X_1, \dots, X_n)$ and the test statistic $T(\mathbf{X})$.

We decide in advance on the acceptance region C_0 and the critical or rejection region C_1 , i.e. if $T(\mathbf{X}) \in C_0$, then we accept H_0 , and if $T(\mathbf{X}) \in C_1$, then we reject H_0 .

Obviously, we must have that $C_0 \cup C_1$ is the whole image space of the test statistic and that $C_0 \cap C_1 = \emptyset$.

There are two types of possible errors in test theory:

Definition 4.1.4 (Type I/II error)

The error of type I is the rejection of the null hypothesis H_0 , although it is true.

The error of type II is the rejection of the alternative H_1 , although it is true, or similarly the false acceptance of the null hypothesis.

	H_0 true	H_0 false
H_0 accepted	Correct	Type II error
H_0 rejected	Type I error	Correct

Remark 4.1.5

We can think of hypothesis testing as a jury trial:

H_0 : The defendant is innocent.

In other words, the type I error corresponds to convicting an innocent defendant. On the other hand, the type II error means that we acquit a guilty defendant.

Mit den beiden Fehlern sind entsprechenden Wahrscheinlichkeiten assoziiert:

$$\begin{aligned} \mathcal{P}(\text{Fehler 1.Art}) &= \mathcal{P}(\mathbf{X} \in C_1 | H_0); \\ \mathcal{P}(\text{Fehler 2.Art}) &= \mathcal{P}(\mathbf{X} \in C_0 | H_1). \end{aligned}$$

Wir können lediglich eine der beiden Wahrscheinlichkeiten kontrollieren.

Definition 4.1.6 (Signifikanzniveau)

Das Signifikanzniveau α ist die Wahrscheinlichkeit für den Fehler erster Art.

Um ein Test zu konstruieren, gehen wir wie folgt voran:

Bemerkung 4.1.7 (Konstruktion eines Tests)

1. Formuliere ein Modell mit noch freien Parametern $\vartheta \in \Theta$.
2. Formuliere die Nullhypothese H_0 und die Alternative H_1 .
3. Wähle das Signifikanzniveau α .
4. Wähle eine Teststatistik T und bestimme die Verteilung von T unter der Nullhypothese.
5. Bestimme aus der Verteilung von T , unter der Nullhypothese, den Annahmebereich C_0 bzw. Ablehnbereich C_1 . Erst danach berechnet man die Teststatistik für die gegebenen Daten und entscheidet, ob man die Nullhypothese ablehnt oder nicht.

Connected to the different errors are the following probabilities:

$$\begin{aligned} \mathcal{P}(\text{type I error}) &= \mathcal{P}(\mathbf{X} \in C_1 | H_0); \\ \mathcal{P}(\text{type II error}) &= \mathcal{P}(\mathbf{X} \in C_0 | H_1). \end{aligned}$$

We can only control one of the both probabilities.

Definition 4.1.6 (Significance level)

The significance level α is the probability for the error of type I.

To construct a test, we proceed as follows:

Remark 4.1.7 (Construction of a test)

1. Formulate a model with some free parameters $\vartheta \in \Theta$.
2. Formulate the null hypothesis H_0 and the alternative hypothesis H_1 .
3. Choose the significance level α .
4. Choose a test statistic T and determine the distribution of T under the null hypothesis.
5. Out of the distribution of T under the null hypothesis, determine the acceptance region C_0 and critical region C_1 . Afterwards, we calculate the test statistic for the given data and decide, whether we reject the null hypothesis or not.

Bemerkung 4.1.8

Typische Werte für α sind 10%, 5% oder 1%.
Je kleiner α ist, umso sicherer sind wir, dass bei einem Ablehnen der Nullhypothese, diese auch wirklich nicht zutrifft. Allerdings passiert es bei kleinerem α immer seltener, dass wir die Nullhypothese ablehnen.

Da wir den Fehler erster Art kontrollieren, ist unser eigentliches Ziel, die Nullhypothese abzulehnen, d.h. wir wollen an sich die Alternative beweisen.

Wenn wir die Nullhypothese nicht ablehnen können, dann wissen wir nichts. Formal akzeptieren wir die Nullhypothese, aber die Wahrscheinlichkeit für den Fehler zweiter Art kann sehr gross sein!

Wie die Nullhypothese und die Alternative formuliert werden, hängt also entscheidend davon ab, was der Benutzer des Tests erreichen möchte:

Beispiel 4.1.9

Betrachten wir einen Feuermelder, bei dem man die Empfindlichkeit einstellen kann.

Remark 4.1.8

Typical values for α are 10%, 5% or 1%.
The smaller α , the more sure we are that the null hypothesis is false, if we reject it. However, the smaller α , the less often we will be able to reject the null hypothesis.

Since we control the type I error, our real target is to reject the null hypothesis, i.e. we want to show that the alternative holds.

If we cannot reject the null hypothesis, then we know nothing. Formally, we accept the null hypothesis, but the probability for the type II error can be very high!

How the null hypothesis and the alternative hypothesis are formulated, depends heavily on what the user of the test would like to reach:

Remark 4.1.9

Let us consider a fire-alarm, where the sensitiveness can be chosen.

Der Hausbesitzer würde die Empfindlichkeit sehr niedrig einstellen, weil es ihn nicht so sehr stört, falls ein Fehlalarm ausgelöst wird, aber es katastrophal ist, falls bei einem Feuer kein Alarm ausgelöst wird.

Dies entspricht der Nullhypothese

H_0 : „Feuer“

da er sich dann bei einem Ablehnen relativ sicher sein kann, dass es auch wirklich nicht brennt.

Die Feuerwehr vertritt eher die andere Meinung, d.h. sie wollen nicht so oft umsonst ausrücken.

Sie wählen also als Nullhypothese

H_0 : „kein Feuer“

da sie dann nur Ausrücken, wenn die Nullhypothese abgelehnt wird, und in diesem Fall es relativ sicher auch brennt.

Bemerkung 4.1.10

Aus dem vorherigen Beispiel wird ersichtlich, dass man als Nullhypothese wählen soll, was beim fälschlichen Ablehnen zu schlimmeren Konsequenzen führt!

Um die Güte eines Punktschätzers zu beurteilen, führten wir das Risiko ein. Dieselbe Vorgehensweise wird auch bei Tests gemacht. Allerdings mit einer anderen Verlustfunktion und es wird auch nicht Risiko, sondern „Macht“ bzw. „Güte“ genannt.

The owner of the house would choose a low sensitiveness, because he does not care so much if there is a false alarm, but it would be disastrous if there is no alarm in the case of a fire.

This corresponds to the null hypothesis

H_0 : “Fire”

because in the case of the rejection of the null hypothesis, he can be rather sure that it does not burn.

The fire-brigade has more likely the other opinion, i.e. they do not want to be called so often if there is no fire.

They choose for the null hypothesis

H_0 : “No fire”

because in that case, they are only called, if the null hypothesis is rejected and in that case it is rather sure that it burns.

Remark 4.1.10

We learn from the previous example, that we should choose for the null hypothesis what leads to worse consequences in the case of a false rejection!

We introduced the risk in order to judge the quality of a point estimator. The same is done for tests. However, the loss function is different and we do not call it risk but “power”.

Definition 4.1.11 (Macht / Güte)

Die Wahrscheinlichkeit, dass H_0 richtigerweise abgelehnt wird,

$$Q(T, \vartheta) = P_{\vartheta}(T(\mathbf{X}) \in C_1) \quad \vartheta \in \Theta_1,$$

wird die Macht des Tests T genannt.

Definition 4.1.11 (Power)

The probability that H_0 is correctly rejected,

is called the power of the test T .

Bemerkung 4.1.12

Wir wollen am liebsten $Q(T, \vartheta) = 1$ haben!
Dies ist gleichbedeutend mit $\mathcal{P}(\text{Fehler 2. Art}) = 0$.

Die Analogie zum Risiko bei Punktschätzern wird durch folgende Rechnung klar:

$$Q(T, \vartheta) = E_{\vartheta}(1_{\{T(\mathbf{X}) \in C_1\}}).$$

Eng verknüpft mit der Macht ist die sogenannte Operationscharakteristik.

Remark 4.1.12

We would like to have $Q(T, \vartheta) = 1$!
This is the same as $\mathcal{P}(\text{type II error}) = 0$.

The analogy to the risk for point estimators becomes clear by the following calculation:

Closely connected to the power is the so-called operation characteristic.

Definition 4.1.13 (Operationscharakteristik)

Die Operationscharakteristik des Tests T ist

$$\beta(T, \vartheta) = P_{\vartheta}(T(\mathbf{X}) \notin C_1).$$

Bei Konfidenzintervallen berechneten wir bereits die Wahrscheinlichkeit, dass eine Statistik in einem gegebenen Bereich liegt. Tatsächlich kann man dies auch für Tests nutzen.

Definition 4.1.13 (Operation characteristic)

The operation characteristic of the test T is

We calculated the probability of a statistic to lie in a given region already in the case of confidence intervals. In fact, we can use this also for tests.

Bemerkung 4.1.14 (Beziehung zwischen Hypothesentests und Konfidenzintervallen)

Sei $[g(X), h(X)]$ ein $(1 - \alpha)$ -Konfidenzintervall für ϑ .

Wir nehmen folgende Nullhypothese und Alternative:

$$H_0 : \vartheta = \vartheta_0$$

$$H_1 : \vartheta \neq \vartheta_0$$

In diesem Fall ist der Annahmebereich für die Teststatistik $T(\mathbf{X}) = \mathbf{X}$ gleich

$$C_0 = \{\mathbf{x} \mid \vartheta_0 \in [g(\mathbf{x}); h(\mathbf{x})]\}.$$

Dies führt zu einem Test mit Niveau α , weil

$$P_{\vartheta_0}(\mathbf{X} \notin C_0) = 1 - P_{\vartheta_0}(\vartheta_0 \in [g(\mathbf{X}); h(\mathbf{X})]) \leq 1 - (1 - \alpha) = \alpha.$$

Remark 4.1.14 (Link between hypothesis tests and confidence intervals)

Let $[g(X), h(X)]$ be a $(1 - \alpha)$ -confidence interval for ϑ .

We take the following null hypothesis and alternative hypothesis:

In this case, the acceptance region for the test statistic $T(\mathbf{X}) = \mathbf{X}$ is equal to

This leads to a test with level α , because

**4.2 Tests für normalverteilte Daten
Tests for normally distributed data**

Viele Hypothesentests basieren auf der Annahme, dass die Zufallsvariablen einer Normalverteilung folgen. Hier beschreiben wir einige gewöhnliche Tests, die diese Annahme benutzen.

Der erste Test ist der sogenannte t-Test, bei dem bzgl. des Mittelwerts getestet wird und die Varianz als unbekannt betrachtet wird:

Many hypothesis tests are based on the assumption that the random variables follow a normal distribution. Here we describe some common tests using this assumption.

The first test is the so-called t-test, where we test for the mean and the variance is regarded as unknown:

Definition 4.2.1 (Einseitiger Einstichproben t-Test (Student's Test))

Modell:

X_1, \dots, X_n u.i. $\mathcal{N}(\mu, \sigma^2)$ -verteilt.

Problem:

$$H_0 : \mu \in \Theta_0 = \{\mu_0\}$$

$$H_1 : \mu \in \Theta_1 = (\mu_0, \infty) = \{\mu > \mu_0\}$$

Die Nullhypothese ist also einseitig und σ^2 als unbekannt betrachtet.

Teststatistik:

$$T(\mathbf{X}) = \sqrt{n} \cdot \frac{\bar{X}_n - \mu_0}{\hat{s}_n} \sim t_{n-1},$$

wobei

$$\hat{s}_n^2 = \frac{1}{n-1} \cdot \sum_{j=1}^n (X_j - \bar{X}_n)^2.$$

Annahmeregion:

$$C_0 = \{T(\mathbf{x}) \leq \tau_\alpha\},$$

wobei τ_α das $(1 - \alpha)$ -Quantil der t_{n-1} -Verteilung ist, wie in Abbildung 4.2.1 illustriert.

Definition 4.2.1 (One-sided one-sample t-test (Student's test))

Model:

X_1, \dots, X_n i.i. $\mathcal{N}(\mu, \sigma^2)$ -distributed.

Problem:

The null hypothesis is thus one-sided and σ^2 is regarded as unknown.

Test statistic:

where

Acceptance region:

where τ_α is the $(1 - \alpha)$ -quantile of the t_{n-1} -distribution, as illustrated in Figure 4.2.1.

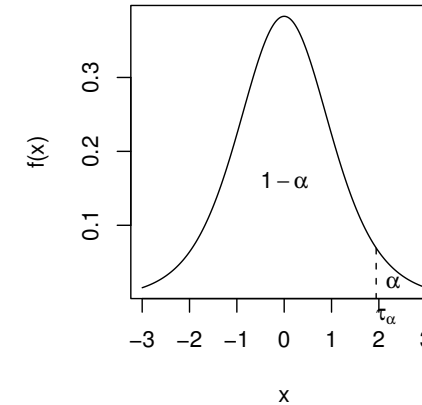


Figure 4.2.1: t-test

Weil die t-Statistik $T(\mathbf{X})$ unter der Nullhypothese H_0 gleich t_{n-1} -verteilt ist, wird das Signifikanzniveau α eingehalten:

Since the t-statistic $T(\mathbf{X})$ is t_{n-1} -distributed under the null hypothesis H_0 , the significance level α is obtained:

$$P_{\mu_0, \sigma^2}(T(\mathbf{X}) > \tau_\alpha) = \alpha \quad \forall \sigma^2 > 0.$$

Bemerkung 4.2.2

Es macht keinen Unterschied im Beispiel 4.2.1, ob wir als Nullhypothese $\mu \in \Theta_0 = \{\mu_0\}$ oder $\mu \in \Theta_0 = (-\infty, \mu_0]$ nehmen, da die Alternative nur einseitig ist und dadurch μ_0 der „kritische“ Wert ist.

Ist nämlich der wahre Wert μ kleiner als μ_0 , so erwarten wir für die Teststatistik erst recht einen kleineren Wert.

Remark 4.2.2

It makes no difference in Example 4.2.1 whether we use the null hypothesis $\mu \in \Theta_0 = \{\mu_0\}$ or $\mu \in \Theta_0 = (-\infty, \mu_0]$, because the alternative hypothesis is only single-sided and therefore, μ_0 is the “critical” value.

If the true value μ is smaller than μ_0 , then we expect for the test statistic even a smaller value.

Beispiel 4.2.3

Wir betrachten wieder Beispiel 1.1.9 auf Seite 8. Wir nehmen an, dass die StudentInnen die folgende Hypothese bezüglich des Gewichts der Schnitzel (X_i u.i. $\mathcal{N}(\mu, \sigma^2)$) testen wollen:

$$H_0 : \mu \geq \mu_0 = 180\text{g}$$

$$H_1 : \mu < 180\text{g}$$

Angenommen, dass $n = 50$ Beobachtungen aufgenommen wurden und dass der beobachtete Mittelwert $\bar{x}_n = 184,6\text{g}$ ist und die Standardabweichung $\hat{s}_n = 6\text{g}$. Dann ist der Wert der t-Statistik gleich 5,42. Das 1%-Quantil der t_{49} -Verteilung ist $-2,40$ und somit kann die Nullhypothese zum Signifikanzniveau $\alpha = 1\%$ nicht verworfen werden.

Zu beachten ist, dass hier die Alternative anderst herum formuliert wurde als in Beispiel 4.2.1!

Example 4.2.3

We recall Example 1.1.9 on page 8. We assume that the students want to test the following hypothesis concerning the weight of the steaks (X_i i.i. $\mathcal{N}(\mu, \sigma^2)$):

Suppose that $n = 50$ observations have been recorded and that the observed mean is $\bar{x}_n = 184.6\text{g}$ and the standard deviation $\hat{s}_n = 6\text{g}$. Then the value of the t-statistic is $t = 5.42$. The 1%-quantile of the t_{49} -distribution is -2.40 , so the null hypothesis cannot be rejected at the significance level $\alpha = 1\%$.

We must be aware that the alternative hypothesis is exactly the other way round than in Example 4.2.1!

Definition 4.2.4 (Gepaarter t-Test)

Angenommen, wir haben Paare von Zufallsvariablen (X_i, Y_i) und dass $D_i = X_i - Y_i$, $i = 1, \dots, n$ u.i. normal verteilt sind.

Dann kann man denselben Test wie in Beispiel 4.2.1 anwenden. Da wir aber von Paaren von Zufallsvariablen ausgegangen sind, nennen wir den Test nun einen gepaarten t-Test.

Manchmal sind wir interessiert an einer Abweichung vom Mittelwert μ_0 , d.h. eine positive und eine negative Abweichung sollen betrachtet werden:

Definition 4.2.5 (Zweiseitiger Einstichproben t-Test)**Modell:**

X_1, \dots, X_n u.i. $\mathcal{N}(\mu, \sigma^2)$ -verteilt und σ^2 unbekannt.

Problem:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \in \Theta_1 = (-\infty, \mu_0) \cup (\mu_0, \infty)$$

Teststatistik:

$$T(\mathbf{X}) = \sqrt{n} \cdot \frac{\bar{X}_n - \mu_0}{\hat{s}_n} \sim t_{n-1},$$

Definition 4.2.4 (Paired t-test)

Suppose that we have pairs of random variables (X_i, Y_i) and that $D_i = X_i - Y_i$, $i = 1, \dots, n$ are i.i. normally distributed.

Then, we can proceed as in Example 4.2.1. However, we started from pairs of random variables and therefore, we now call the test a paired t-test.

Sometimes, we are interested in a deviation from the mean value μ_0 , i.e. a positive and a negative deviation shall be considered:

Definition 4.2.5 (Two-sided one-sample t-test)**Model:**

X_1, \dots, X_n i.i. $\mathcal{N}(\mu, \sigma^2)$ -distributed and σ^2 unknown.

Problem:**Test statistic:**

wobei

where

$$\hat{s}_n^2 = \frac{1}{n-1} \cdot \sum_{j=1}^n (X_j - \bar{X}_n)^2.$$

Annahmebereich:

Acceptance region:

$$C_0 = \{ |T(\mathbf{x})| \leq \tau_\alpha \},$$

wobei τ_α das $(1 - \frac{\alpha}{2})$ -Quantil einer t_{n-1} -Verteilung ist.

where τ_α is the $(1 - \frac{\alpha}{2})$ -quantile of the t_{n-1} -distribution.

Wegen der Symmetrie der t -Verteilung erhalten wir:

Due to the symmetry of the t -distribution, we obtain:

$$P_{\mu_0, \sigma^2}(|T(\mathbf{X})| > \tau_\alpha) = \alpha \quad \forall \sigma^2 > 0.$$

Demnach hat dieser zweiseitige Test auch das Signifikanzniveau α .

Thus, this two-sided test has also the significance level α .

Manchmal ist man auch daran interessiert, bis zu welchem Signifikanzniveau man die Hypothese ablehnen kann. Dies ist der sogenannte p -Wert. Da es wesentlich einfacher ist, dies für einen t -Test zu formulieren, machen wir dies hier, obwohl es wesentlich allgemeiner geht.

We are sometimes interested up to which significance level, we could reject the hypothesis. This is the so-called p -value. Since it is much easier to formulate this for a t -test, we do it here, although it can be done more generally.

Definition 4.2.6 (p -Wert)

Sei $T(\mathbf{x})$ die Teststatistik eines zweiseitigen t -Tests zur Nullhypothese $H_0 : \mu = \mu_0$ für die beobachteten Daten x_1, \dots, x_n . Y_1, \dots, Y_n seien u.i. $\mathcal{N}(\mu_0, \sigma^2)$ -verteilt. Der p -Wert ist nun:

Definition 4.2.6 (p -value)

Let $T(\mathbf{X})$ be the test statistic of a two-sided t -test with null hypothesis $H_0 : \mu = \mu_0$ for the given data x_1, \dots, x_n . Y_1, \dots, Y_n are i.i. $\mathcal{N}(\mu_0, \sigma^2)$ -distributed. The p -value is now:

$$P_{\mu_0}(T(\mathbf{Y}) > |T(\mathbf{x})|).$$

Im folgenden Test benutzen wir Daten von zwei unterschiedlichen Stichproben:

In the following test, we use data from two different samples:

Definition 4.2.7 (Zweiseitiger Zweistichproben- t -Test)

Modell:

X_1, \dots, X_n u.i. $\mathcal{N}(\mu_1, \sigma^2)$ -verteilt und Y_1, \dots, Y_m u.i. $\mathcal{N}(\mu_2, \sigma^2)$ -verteilt.

Problem:

Definition 4.2.7 (Two-sided two-sample t -test)

Model:

X_1, \dots, X_n i.i. $\mathcal{N}(\mu_1, \sigma^2)$ -distributed and Y_1, \dots, Y_m i.i. $\mathcal{N}(\mu_2, \sigma^2)$ -distributed.

Problem:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

Teststatistik:

Test statistic:

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{1}{n} + \frac{1}{m}} \hat{s}} \sim t_{n+m-2}$$

mit

with

$$\hat{s}^2 = \frac{(n-1) \cdot \hat{s}_x^2 + (m-1) \cdot \hat{s}_y^2}{n+m-2},$$

wobei

where

$$\hat{s}_x^2 = \frac{1}{n-1} \cdot \sum_{j=1}^n (X_j - \bar{X}_n)^2;$$

$$\hat{s}_y^2 = \frac{1}{m-1} \cdot \sum_{j=1}^m (Y_j - \bar{Y}_m)^2.$$

Annahmebereich:**Acceptance region:**

$$C_0 = \{T(\mathbf{x}, \mathbf{y}) \leq \tau_\alpha\},$$

wobei $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ und τ_α das $(1 - \alpha)$ -Quantil der t_{n+m-2} -Verteilung ist.

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ and τ_α is the $(1 - \alpha)$ -quantile of the t_{n+m-2} -distribution.

Beispiel 4.2.8

Während ihrer Wanderung zum Humbergturm gehen die StudentInnen am Erholungsgebiet Bremerhof vorbei. Dort treffen sie den Bauer, der für die Hirsche, die am Bremerhof leben, verantwortlich ist. Er erzählt den StudentInnen, dass er seit zwei Monaten ein neues Futter mit extra Proteinen für die Hirsche testet. 9 von 22 Hirschen (Gruppe 1) essen immer noch das alte Futter und die anderen 13 Hirsche (Gruppe 2) essen das neue proteinreiche Futter.

Example 4.2.8

During their hike to the Humberg tower, the students pass the recreation area Bremerhof. There they meet the farmer responsible for the deers living at Bremerhof. He tells the students that since two months he is testing new feedstuff with extra proteins for the deers. 9 of the 22 deers (called Group 1) still eat the old food and the other 13 deers (Group 2) eat the new protein-rich food.

Der Bauer will jetzt die Gewichtszunahme der Tiere während den zwei Testmonaten benutzen, um herauszufinden, ob es einen Unterschied zwischen dem neuen und alten Futter gibt, außer dem höheren Preis für das neue Futter; d.h. er will das neue Futter in Zukunft nur dann einsetzen, wenn es wirklich eine größere Gewichtszunahme bringt.
Die folgende Gewichtszunahme wurde für die 22 Tiere gemessen (in g):

The farmer wants to use the weight increase of the animals during the two test months to conclude if there is a difference between the new food and the old one, except that the new one is more expensive, i.e. he wants to use the new feedstuff in the future only, if it really leads to a higher increase in the weight.
The following weight increase was recorded for the 22 animals (in g):

Gruppe/Group 1: 710; 1170; 1020; 830; 1080; 1320; 950; 780; 1110;
Gruppe/Group 2: 1330; 1450; 1050; 1190; 1280; 1600; 1040; 910; 1310; 1010; 1280; 1100; 950;

Die StudentInnen helfen dem Bauer einen Hypothesentest zu konstruieren und benutzen folgende Annahmen:

X_1, \dots, X_9 , die Gewichte der Tiere in Gruppe 1, sind u.i. $\mathcal{N}(\mu_1, \sigma^2)$ -verteilt.
Analog seien die Gewichte Y_1, \dots, Y_{13} der Tiere in Gruppe 2 u.i. $\mathcal{N}(\mu_2, \sigma^2)$ -verteilt.
Die Varianz σ^2 wird als unbekannt betrachtet, aber als gleich in beiden Gruppen, angenommen.

Now, the students help the farmer to construct a hypothesis test and they use the following assumptions:

X_1, \dots, X_9 are i.i. $\mathcal{N}(\mu_1, \sigma^2)$ -distributed, describing the weights of the deers in Group 1. Analogously, the weight of the deers in Group 2, Y_1, \dots, Y_{13} are i.i. $\mathcal{N}(\mu_2, \sigma^2)$ -distributed. The variance σ^2 is assumed to be unknown but assumed to be the same for both groups.

Sie benutzen einen Zweistichproben-t-Test mit der Nullhypothese $H_0: \mu_1 = \mu_2$ und der Alternative $H_1: \mu_1 < \mu_2$ und dem Signifikanzniveau 2,5%.

Mit dieser Wahl der Nullhypothese bzw. Alternative, kann der Bauer beim Ablehnen der Nullhypothese ziemlich sicher sein, d.h. mit 97,5% Wahrscheinlichkeit, dass das neue Futter besser ist.

They use a two-sample t-test with the null hypothesis $H_0: \mu_1 = \mu_2$ and the alternative $H_1: \mu_1 < \mu_2$ and the significance level 2.5%.

With this choice for the null hypothesis and alternative, the farmer can be rather sure, i.e. with probability 97.5%, that the new feedstuff is better, if the null hypothesis is rejected.

Sie berechnen die arithmetischen Mittelwerte $\bar{x} = 997$ und $\bar{y} = 1192$ und die Varianzen $\sigma_x^2 = 39250$ und $\sigma_y^2 = 42036$.
Der Wert der Teststatistik aus Definition 4.2.7 ist 2,23 und das 97,5%-Quantil der t_{20} -Verteilung ist 2,09.
Also kann die Nullhypothese auf dem Signifikanzniveau 2,5% verworfen werden.

They calculate the arithmetic means $\bar{x} = 997$ and $\bar{y} = 1192$ and the variances $\sigma_x^2 = 39250$ and $\sigma_y^2 = 42036$.
Using the test statistic of Definition 4.2.7, the value 2.23 is obtained. The 97,5%-quantile of the t_{20} -distribution is 2.09.
Thus, the null hypothesis can be rejected at the significance level 2.5%.

Bemerkung 4.2.9

In obigen Definitionen 4.2.7 und 4.2.8, nahmen wir an, dass $\sigma^2 > 0$ unbekannt, aber gleich für beide Stichproben ist.

Wenn wir diese Annahme testen wollen, so gibt es auch dafür Tests, z.B. der folgende F-Test.

Remark 4.2.9

In the definitions 4.2.7 and 4.2.8 above, we assumed that the variance $\sigma^2 > 0$ is unknown but the same in both samples.

If we want to test this assumption, then there are also tests, e.g. the following F-Test.

Definition 4.2.10 (F-Test)

Modell:

Seien X_1, \dots, X_m u.i. $\mathcal{N}(\mu_1, \sigma_1^2)$ -verteilt und Y_1, \dots, Y_n u.i. $\mathcal{N}(\mu_2, \sigma_2^2)$ -verteilt. Alle Parameter $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ werden als unbekannt angenommen.

Definition 4.2.10 (F-test)

Model:

Let X_1, \dots, X_m be i.i. $\mathcal{N}(\mu_1, \sigma_1^2)$ -distributed and Y_1, \dots, Y_n i.i. $\mathcal{N}(\mu_2, \sigma_2^2)$ -distributed. All the parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ are assumed to be unknown.

Problem:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Problem:

Teststatistik:

Mit Definition 3.1.3 und Satz 3.1.10 erhalten wir

$$T(\mathbf{X}, \mathbf{Y}) = \frac{s_m^2}{s_n^2} \sim F_{m-1, n-1},$$

wobei

Test statistic:

With Definition 3.1.3 and Theorem 3.1.10, we obtain

where

$$s_m^2 = \frac{1}{m-1} \cdot \sum_{i=1}^m (X_i - \bar{X}_m)^2;$$
$$s_n^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (Y_i - \bar{Y}_n)^2.$$

Annahmehereich:

Acceptance region:

$$C_0 = \{f_{\frac{\alpha}{2}} \leq T(\mathbf{x}, \mathbf{y}) \leq f_{1-\frac{\alpha}{2}}\},$$

wobei $f_{\frac{\alpha}{2}}$ und $f_{1-\frac{\alpha}{2}}$ die $\frac{\alpha}{2}$ - bzw. $(1 - \frac{\alpha}{2})$ -Quantile der $F_{m-1, n-1}$ -Verteilung sind.

where $f_{\frac{\alpha}{2}}$ and $f_{1-\frac{\alpha}{2}}$ are the $\frac{\alpha}{2}$ - and $(1 - \frac{\alpha}{2})$ -quantiles, respectively, of the $F_{m-1, n-1}$ -distribution.

Beispiel 4.2.11

Als die MathematikstudentInnen atemlos am Humberturm ankommen, entdecken sie, dass sie nicht die einzigen StudentInnen sind, die eine Humberturmexpedition machen:

Einige PhysikstudentInnen sind dort, um die Varianzen von zwei verschiedenen Meßgeräten M_1 und M_2 zu vergleichen. Dafür haben sie die Höhe des Humberturms mehrmals gemessen. Dies sind die Ergebnisse der Meßungen (in Dezimeter):

- M_1 251; 301; 323; 345; 310; 321; 292; 309; 321; 310;
- M_2 237; 304; 343; 301; 312; 302; 310; 295; 305; 313; 288; 300;

Example 4.2.11

Breathless arriving to the Humbert tower, the mathematics students discover that they are not the only students making a Humbert tower expedition:

Some physics students are there to compare the variances of two kind of measuring instruments, called M_1 and M_2 . To do so, they have measured the height of the Homberg tower a number of times. These are the results of the measurements (in decimeter):

Die StudentInnen machen folgende Annahmen:
Die gemessenen Höhen X_1, \dots, X_{10} mit M_1 sind u.i. $\mathcal{N}(\mu_1, \sigma_1^2)$ -verteilt. Analog für M_2 seien Y_1, \dots, Y_{12} u.i. $\mathcal{N}(\mu_2, \sigma_2^2)$ -verteilt.

Die Nullhypothese $H_0 : \sigma_1^2 = \sigma_2^2$ wird gegen die Alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$ auf dem Signifikanzniveau 5% getestet.

Die Teststatistik

$$T(\mathbf{X}, \mathbf{Y}) = \frac{s_{10}^2}{s_{12}^2} \sim F_{9,11}$$

von Definition 4.2.10 führt zu einem Wert von 1,04.

Das 2,5%-Quantil der $F_{9,11}$ -Verteilung ist 0,26 und das 97,5%-Quantil ist 3,59, d.h. die Nullhypothese kann nicht verworfen werden.

The students make the following assumptions:
The measured heights X, \dots, X_{10} with M_1 are i.i. $\mathcal{N}(\mu_1, \sigma_1^2)$ -distributed. Analogously for M_2 , Y_1, \dots, Y_{12} are i.i. $\mathcal{N}(\mu_2, \sigma_2^2)$ -distributed.

The null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ is tested against the alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$ at the significance level 5%.

The test statistic

from Definition 4.2.10 leads to the value 1.04.

The 2.5%-quantile of the $F_{9,11}$ -distribution is 0.26 and the 97.5%-quantile is 3.59, i.e. the null hypothesis cannot be rejected.

4.3 Likelihood-Quotienten-Tests Likelihood ratio tests

4.3.1 Gleichmäßig beste Tests Uniformly most powerful tests

In Definition 4.1.11 führten wir die Macht eines Tests als Äquivalent zum Risiko für Punktschätzer ein, d.h. die Güte soll für viele $\vartheta \in \Theta_1$, laut Bemerkung 4.1.12, groß sein. Daraus ergibt sich folgende Definition für gleichmäßig beste Tests:

In Definition 4.1.11, we introduced the power of a test as equivalent to the risk of a point estimator. According to Remark 4.1.12, the power should be large for many $\vartheta \in \Theta_1$. This leads us to the following definition for uniformly most powerful tests:

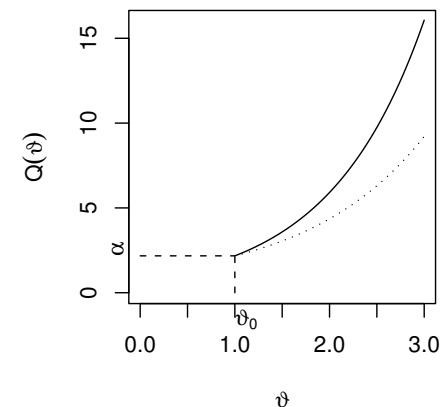


Figure 4.3.1: Gleichmäßig bester Test
Uniformly most powerful test

Definition 4.3.1 (Gleichmäßig bester Test)

Die Macht Q eines gleichmäßig besten Tests mit Signifikanzniveau α ist für jedes $\vartheta \in \Theta_1$ mindestens so groß wie die Macht eines beliebigen anderen Tests zum Signifikanzniveau α .

In Abbildung 4.3.1 wird dies illustriert, wobei die durchgezogene Kurve die Macht eines gleichmäßig besten Tests darstellt und die gestrichelte Kurve die Güte eines beliebigen anderen Tests zum selben Signifikanzniveau.

Definition 4.3.1 (Uniformly most powerful test)

The power Q of a uniformly most powerful test with significance level α must for each $\vartheta \in \Theta_1$ be at least as high as the power of any other test with significance level α .

This is illustrated in Figure 4.3.1, where the solid curve represents the power of a uniformly most powerful test and the dotted curve relates to any other test at the same significance level.

Bemerkung 4.3.2

In Bemerkung 2.1.8 haben wir gesehen, dass es keine gleichmäßig beste Punktschätzer gibt. Wie soll es dann einen gleichmäßig besten Test geben, wenn die Macht so sehr mit dem Risiko verwandt ist?

Die Antwort ist, dass die Alternative nur einpunktig ist, d.h. die Macht wird nur an einem einzigen Punkt ausgewertet. Dies entspricht dem Fall bei Punktschätzern, dass die Parametermenge nur einelementig ist. Im Gegensatz zu Punktschätzern, ist hier die Situation doch noch interessant, da man immer noch zwischen Nullhypothese und Alternative wählen muss.

Die Situation ist somit:

$$H_0 : \vartheta \in \Theta_0 = \{\vartheta_0\}$$

$$H_1 : \vartheta \in \Theta_1 = \{\vartheta_1\}.$$

Satz 4.3.3 (Neyman-Pearson Lemma)

Seien X_1, \dots, X_n Zufallsvariablen mit gemeinsamer Verteilung \mathcal{P}_ϑ und $\vartheta \in \Theta = \{\vartheta_0; \vartheta_1\}$. Sei $L(\vartheta|\mathbf{X})$ die Likelihood-Funktion. Als Teststatistik betrachten wir

$$T(\mathbf{X}) = \frac{L(\vartheta_0|\mathbf{X})}{L(\vartheta_1|\mathbf{X})}.$$

Remark 4.3.2

We have seen in Remark 2.1.8, that there is no uniformly best point estimator. How should it be possible that there exists a uniformly most powerful test, if the power is so closely related to the risk?

The answer is that the alternative consists of one point, i.e. the power is evaluated just at a single point. This corresponds for point estimators to the case that the parameter space has only one element. In contrast to point estimators, the situation here is still interesting, because we still have to choose between null hypothesis and alternative.

The situation is therefore:

Theorem 4.3.3 (Neyman-Pearson Lemma)

Let X_1, \dots, X_n be random variables with common distribution \mathcal{P}_ϑ , where $\vartheta \in \Theta = \{\vartheta_0, \vartheta_1\}$. Let $L(\vartheta|\mathbf{x})$ be the likelihood function. For the test statistic, we use

Wenn es ein k gibt, so dass

If there is a k , such that

$$C_0 := \{T(\mathbf{x}) \geq k\}$$

$$C_1 := \{T(\mathbf{x}) < k\}$$

zu einem Test mit Signifikanzniveau α führt, dann ist dies der gleichmäßig beste Test.

leads to a test with significance level α , then this is the uniformly most powerful test.

Beweis:

Sei A_0 und A_1 der Annahme- bzw. kritische Bereich eines anderen Tests S zum Signifikanzniveau α . Da wir annehmen, dass k existiert, gilt

$$\mathcal{P}_{\vartheta_0}(T(\mathbf{X}) \in C_1) = \mathcal{P}_{\vartheta_0}(S(\mathbf{X}) \in A_1) = \alpha \tag{4.3.4}$$

Zu zeigen ist

We must show

$$\mathcal{P}_{\vartheta_1}(T(\mathbf{X}) \in C_1) \geq \mathcal{P}_{\vartheta_1}(S(\mathbf{X}) \in A_1).$$

Wir führen folgende vereinfachende Notation ein:

We introduce the following simplifying notation:

$$\tilde{C}_i := \{\omega \in \Omega \mid T(\mathbf{X}(\omega)) \in C_i\} \quad i = 0; 1$$

$$\tilde{A}_i := \{\omega \in \Omega \mid S(\mathbf{X}(\omega)) \in A_i\} \quad i = 0; 1$$

Damit stellt sich unser Ziel folgendermaßen dar:

Our aim is now with the new notation:

$$0 \stackrel{?}{\leq} \mathcal{P}_{\vartheta_1}(\tilde{C}_1) - \mathcal{P}_{\vartheta_1}(\tilde{A}_1) = [\mathcal{P}_{\vartheta_1}(\tilde{C}_1 \cap \tilde{A}_1) + \mathcal{P}_{\vartheta_1}(\tilde{C}_1 \setminus \tilde{A}_1)] - [\mathcal{P}_{\vartheta_1}(\tilde{C}_1 \cap \tilde{A}_1) + \mathcal{P}_{\vartheta_1}(\tilde{A}_1 \setminus \tilde{C}_1)]$$

$$= \mathcal{P}_{\vartheta_1}(\tilde{C}_1 \setminus \tilde{A}_1) - \mathcal{P}_{\vartheta_1}(\tilde{A}_1 \setminus \tilde{C}_1).$$

Aus der Konstruktion des Tests T folgt

From the construction of the test T follows

$$\omega \in \tilde{C}_1 \setminus \tilde{A}_1 \subset \tilde{C}_1 \Rightarrow L(\vartheta_1 | \mathbf{X}(\omega)) > \frac{1}{k} \cdot L(\vartheta_0 | \mathbf{X}(\omega))$$

$$\omega \in \tilde{A}_1 \setminus \tilde{C}_1 \subset \tilde{C}_0 \Rightarrow L(\vartheta_1 | \mathbf{X}(\omega)) \leq \frac{1}{k} \cdot L(\vartheta_0 | \mathbf{X}(\omega))$$

Unabhängig davon, ob die Likelihood-Funktion gleich der Dichte oder einer diskreten Wahrscheinlichkeit ist, ergibt sich damit umgekehrt

It plays no role whether the likelihood function is the density or a discrete probability, in order to obtain reversely

$$\mathcal{P}_{\vartheta_1}(\tilde{C}_1 \setminus \tilde{A}_1) \geq \frac{1}{k} \cdot \mathcal{P}_{\vartheta_0}(\tilde{C}_1 \setminus \tilde{A}_1)$$

$$\mathcal{P}_{\vartheta_1}(\tilde{A}_1 \setminus \tilde{C}_1) \leq \frac{1}{k} \cdot \mathcal{P}_{\vartheta_0}(\tilde{A}_1 \setminus \tilde{C}_1).$$

Damit erhalten wir

Using this, we receive

$$\begin{aligned} \mathcal{P}_{\vartheta_1}(\tilde{C}_1 \setminus \tilde{A}_1) - \mathcal{P}_{\vartheta_1}(\tilde{A}_1 \setminus \tilde{C}_1) &\geq \frac{1}{k} \cdot \{[\mathcal{P}_{\vartheta_0}(\tilde{C}_1 \setminus \tilde{A}_1) + \mathcal{P}_{\vartheta_0}(\tilde{C}_1 \cap \tilde{A}_1)] - [\mathcal{P}_{\vartheta_1}(\tilde{A}_1 \setminus \tilde{C}_1) + \mathcal{P}_{\vartheta_0}(\tilde{C}_1 \cap \tilde{A}_1)]\} \\ &= \frac{1}{k} \cdot [\mathcal{P}_{\vartheta_0}(\tilde{C}_1) - \mathcal{P}_{\vartheta_0}(\tilde{A}_1)] = 0, \end{aligned}$$

laut Gleichung 4.3.4.

by equation 4.3.4.

Beispiel 4.3.6

Wir erinnern uns an die Elwedritschefalle aus Beispiel 1.2.2 auf Seite 23.

33 Fallen wurden im Wald ausgesetzt und die Tiere, die während einer bestimmter Zeit gefangen wurden, wurden gezählt.

Die Tabelle gibt die Anzahl der Fallen, die die entsprechende Anzahl an Tieren fingen, an:

Example 4.3.6

Recall the elwedritsche trap in Example 1.2.2 on page 23.

33 traps were set out in the forest and the numbers of animals caught in a fixed time interval were counted.

The table gives the number of traps which contained the corresponding numbers of animals:

Anzahl Fallen	Anzahl Tiere pro Falle	No of traps	Count of animals per trap
9	0	9	0
11	1	11	1
4	2	4	2
5	3	5	3
1	4	1	4
2	5	2	5
1	6	1	6
0	≥ 7	0	≥ 7

Wir nehmen an, dass die Anzahlen gefangener Tiere pro Falle unabhängig und Poisson-verteilt mit Parameter bzw. Mittelwert μ sind.
Wir testen

We assume that the numbers of caught animals per trap are independent and that they are Poisson distributed with parameter or mean μ .
We test

$$H_0 : \mu = \mu_0 = 1$$

$$H_1 : \mu = \mu_1 = 3$$

Die Likelihood-Funktion ist

The Likelihood function is

$$L(\mu | \mathbf{X}) = \frac{e^{-33\mu} \cdot \mu^{\sum_{i=1}^{33} X_i}}{\prod (X_i!)},$$

sodass der Ablehnbereich des Neyman-Pearson Lemmas 4.3.3 gleich

such that the critical region of the Neyman-Pearson Lemma 4.3.3 is

$$\left(\frac{e^{-33\mu_0} \cdot (\mu_0)^{\sum_{i=1}^{33} X_i}}{\prod (X_i!)} \right) \Bigg/ \left(\frac{e^{-33\mu_1} \cdot (\mu_1)^{\sum_{i=1}^{33} X_i}}{\prod (X_i!)} \right) < k$$

ist, bzw.

or

$$\frac{e^{-33\mu_0} \cdot (\mu_0)^{\sum_{i=1}^{33} X_i}}{e^{-33\mu_1} \cdot (\mu_1)^{\sum_{i=1}^{33} X_i}} < k.$$

Durch Logarithmieren von beiden Seiten erhalten wir

Taking logs of both sides, we obtain

$$-33 \cdot \mu_0 + 33 \cdot \mu_1 + [\log(\mu_0) - \log(\mu_1)] \cdot \sum_{i=1}^{33} X_i < \log(k),$$

und durch Umordnung

and by rearranging

$$[\log(\mu_0) - \log(\mu_1)] \cdot \sum_{i=1}^{33} X_i < \log(k) + 33 \cdot (\mu_0 - \mu_1).$$

Weil $\mu_1 > \mu_0$ ist, ist der optimale kritische Bereich von der Form

Since $\mu_1 > \mu_0$, the optimal critical region is of the form

$$\sum_{i=1}^{33} X_i > C,$$

für eine Konstante C , abhängig von der Nullhypothese und dem Signifikanzniveau, aber nicht mehr von der Alternative!

for a constant C , depending on the null hypothesis and the significance level, but not from the alternative!

Wir haben $\sum_{i=1}^{33} X_i \sim \text{Poisson}(33 \cdot \mu)$, d.h. unter H_0 gilt $\sum_{i=1}^{33} X_i \sim \text{Poisson}(33)$.

We have $\sum_{i=1}^{33} X_i \sim \text{Poisson}(33 \cdot \mu)$, i.e. under H_0 holds $\sum_{i=1}^{33} X_i \sim \text{Poisson}(33)$.

Wir berechnen nun den p -Wert (siehe Definition 4.2.6).

Now we calculate the p -value (see Definition 4.2.6).

Da 54 Tiere gefangen wurden, ist der Wert unserer Teststatistik $\sum_{i=1}^{33} X_i$ gleich 54 und damit ist der p -Wert gleich

Since 54 animals were caught, we have exactly the same value for our test statistic $\sum_{i=1}^{33} X_i$. Therefore, we obtain for the p -value

$$\mathcal{P}_{\mu_0} \left(\sum_{i=1}^{33} X_i > 54 \right) \approx 3 \cdot 10^{-4},$$

d.h. wir können die Nullhypothese für jedes Signifikanzniveau größer als $3 \cdot 10^{-4}$ ablehnen!

i.e. we can reject the null hypothesis for any significance level greater than $3 \cdot 10^{-4}$!

Wir können natürlich die Hypothese und Alternative auch vertauschen, d.h. der kritische Bereich ist dann von der Form $\sum_{i=1}^{33} X_i < \tilde{C}$ für eine andere Konstante \tilde{C} , die wiederum aber nur von der jetzigen Nullhypothese $H_0 : \mu = \mu_1 = 3$ und dem Signifikanzniveau abhängt.

In unserem jetzigen Beispiel ist $\sum_{i=1}^{33} X_i \sim \text{Poisson}(99)$ unter der Nullhypothese und damit ergibt sich hier als p -Wert

$$\mathcal{P}_{\mu_1} \left(\sum_{i=1}^{33} X_i < 54 \right) \approx 5 \cdot 10^{-7}.$$

Dieses Resultat ist doch sehr verblüffend, denn wir würden beide Hypothesen ablehnen!

Wenn man den Mittelwert schätzt, so erhält man $\frac{54}{33} \approx 1,63$. Die beiden Werte 1 und 3 liegen soweit davon entfernt, dass man beide Hypothesen ablehnt.

Insgesamt war also das Modell $\mu \in \{\mu_0; \mu_1\}$ misspezifiziert und dies führte zu dieser zuerst doch verblüffenden Situation!

Of course, we can interchange the hypothesis and alternative, i.e. then the critical region is of the form $\sum_{i=1}^{33} X_i < \tilde{C}$ for another constant \tilde{C} , which again solely depends on the actual null hypothesis $H_0 : \mu = \mu_1 = 3$ and the significance level.

In our example now, $\sum_{i=1}^{33} X_i \sim \text{Poisson}(99)$ under the null hypothesis and we observe the following p -value

This result is very striking, because we would reject both hypotheses!

If we estimate the mean, then we arrive at $\frac{54}{33} \approx 1.63$. Both values 1 and 3 are so far away from the estimated mean, that we reject both hypotheses.

Finally, the model $\mu \in \{\mu_0; \mu_1\}$ was totally misspecified and this led to this, at the first glance, amazing situation!

Bemerkung 4.3.7 (Randomisierte Tests)

Im vorherigen Beispiel haben wir verschwiegen, dass es nicht zu jedem Signifikanzniveau α ein entsprechendes C gibt, womit man dieses Signifikanzniveau exakt erreicht. In solch einem Fall gilt das Neyman–Pearson Lemma **nicht**.

Man muss dann C so wählen, dass $\mathcal{P}_{\mu_0}(\sum X_i > C) < \alpha$ und $\mathcal{P}_{\mu_0}(\sum X_i \geq C) > \alpha$.

Remark 4.3.7 (Randomized tests)

We did not reveal in the previous example that there are significance levels α for which we cannot find C , such that we exactly receive the significance level. Therefore, in such a case, the Neyman–Pearson Lemma does **not** hold.

Then, we must determine C such that $\mathcal{P}_{\mu_0}(\sum X_i > C) < \alpha$ and $\mathcal{P}_{\mu_0}(\sum X_i \geq C) > \alpha$.

Um doch noch das Niveau α exakt zu erreichen, kann man den Test **randomisieren**:

Ist $\sum X_i = C$, so führe unabhängig davon ein Bernoulli-Experiment mit

$$\text{Erfolgswahrscheinlichkeit } p = \frac{\alpha - p_{\mu_0}(\sum X_i > C)}{p_{\mu_0}(\sum X_i = C)}$$

durch. Bei Erfolg dieses Bernoulli-Experiments lehnen wir die Nullhypothese ab und sonst akzeptieren wir sie.

Damit erhalten wir als Signifikanzniveau

$$P_{\mu_0}(H_0 \text{ rejected}) = P_{\mu_0}(\sum X_i > C) + p \cdot P_{\mu_0}(\sum X_i = C) = \alpha.$$

4.3.2 Allgemeine Likelihood-Quotienten-Tests General likelihood ratio tests

Intuitiv basiert der Neyman-Pearson Ansatz auf der Idee, dass die Hypothese verworfen wird, wenn die Beobachtung $X = x$ unter der Alternativhypothese $\vartheta = \vartheta_1$ um ein Vielfaches wahrscheinlicher ist, als unter der Hypothese $\vartheta = \vartheta_0$.

Wir haben gesehen, dass der Neyman-Pearson Test nur bei einer einfachen Nullhypothese, d.h. nur einelementig, gegen eine einfache Alternative anwendbar ist.

Dies ist nicht sonderlich realitätsnah, weil fast alle Anwendungen zusammengesetzte Alternativen, d.h. nicht nur einelementig, und manchmal auch eine zusammengesetzte Nullhypothese beinhalten.

We can **randomize** the test in order to reach the level α exactly:

If $\sum X_i = C$, then we perform an independent Bernoulli experiment with success probability

$$p = \frac{\alpha - p_{\mu_0}(\sum X_i > C)}{p_{\mu_0}(\sum X_i = C)}. \text{ We reject the null hypothesis}$$

in the case of a success and accept it otherwise.

Thus, we obtain the significance level

Intuitively, the Neyman-Pearson approach is based on the idea that the hypothesis is rejected, when the observation $X = x$ under the alternative hypothesis $\vartheta = \vartheta_1$ is more probable than under the hypothesis $\vartheta = \vartheta_0$.

We have seen that the Neyman-Pearson test applies only to a single point null hypothesis against a single point alternative.

This is not very realistic, because almost all applications involve composite alternatives and even sometimes a composite null hypothesis as well.

Um den Neyman-Pearson Ansatz auf zusammengesetzte Alternativen bzw. Hypothesen zu erweitern, benutzen wir folgende Idee:

Basierend auf den Daten \mathbf{X} , schätzen wir ϑ innerhalb Θ_0 , bezeichnet mit $\hat{\vartheta}_0$, und innerhalb Θ_1 , bezeichnet mit $\hat{\vartheta}_1$. Danach führen wir folgenden Neyman-Pearson Test durch:

$$H_0 : \vartheta = \hat{\vartheta}_0$$

$$H_1 : \vartheta = \hat{\vartheta}_1.$$

Da wir für den Neyman-Pearson Test die Likelihood verwenden, bietet es sich an für die Schätzung Maximum-Likelihood-Schätzer mit den entsprechenden Nebenbedingungen ($\vartheta \in \Theta_0$ bzw. $\vartheta \in \Theta_1$) zu verwenden.

Damit erhält man als Akzeptanzbereich

$$C_0 = \left\{ \frac{\sup_{\vartheta \in \Theta_0} L(\vartheta|\mathbf{X})}{\sup_{\vartheta \in \Theta_1} L(\vartheta|\mathbf{X})} \geq c_\alpha \right\}.$$

Definition 4.3.8

(Likelihood-Quotienten-Test; LQ-Test)

$\frac{\sup_{\vartheta \in \Theta_0} L(\vartheta|\mathbf{X})}{\sup_{\vartheta \in \Theta_1} L(\vartheta|\mathbf{X})}$ wird *Likelihood-Quotient* genannt und der Test mit obigem Annahmebereich C_0 *Likelihood-Quotienten-Test*.

Man verwendet das Supremum und nicht das Maximum, wie es die vorherige Beschreibung suggeriert, da es dieses Maximum nicht geben muss, falls z.B. Θ_1 offen ist.

To extend the Neyman-Pearson approach to cases with composite alternatives or hypotheses, we use the following idea:

Based on the data \mathbf{X} , we estimate ϑ within Θ_0 , denoted by $\hat{\vartheta}_0$, and within Θ_1 , denoted by $\hat{\vartheta}_1$. Afterwards, we perform the following Neyman-Pearson test:

Since we use the likelihood for the Neyman-Pearson test, it seems to be natural to use maximum likelihood estimators with the appropriate constraints ($\vartheta \in \Theta_0$ and $\vartheta \in \Theta_1$, respectively) for the estimators.

Therefore, we obtain the following acceptance region

Definition 4.3.8 (Likelihood ratio test; LQ test)

$\frac{\sup_{\vartheta \in \Theta_0} L(\vartheta|\mathbf{X})}{\sup_{\vartheta \in \Theta_1} L(\vartheta|\mathbf{X})}$ is called *likelihood ratio* and the test with the acceptance region C_0 from above is the *likelihood ratio test*.

We use the supremum and not the maximum, as announced in the previous description, because the maximum does not have to exist, if e.g. Θ_1 is open.

Sehr oft ist $\Theta_0 \subset \overline{\Theta_1}$, z.B. falls $\Theta_0 = \{\vartheta_0\}$ und $\Theta_0 \cup \Theta_1 = \mathbb{R}^n$. In diesem Fall ist dann die Nebenbedingung $\sup_{\vartheta \in \Theta_1}$ irrelevant, d.h. dies ist gleichbedeutend mit $\sup_{\vartheta \in \Theta}$ mit $\Theta = \Theta_0 \cup \Theta_1$. In diesem Fall ist dann offensichtlich die Teststatistik kleiner oder gleich eins.

Ist $\Theta_0 = \{\vartheta_0\}$ und $\Theta_0 \subset \overline{\Theta_1}$, so reduziert sich die Teststatistik zu

$$\left\{ \frac{L(\vartheta_0|\mathbf{X})}{\sup_{\vartheta \in \Theta} L(\vartheta|\mathbf{X})} \geq c_\alpha \right\}$$

Very often holds $\Theta_0 \subset \overline{\Theta_1}$, e.g. for $\Theta_0 = \{\vartheta_0\}$ and $\Theta_0 \cup \Theta_1 = \mathbb{R}^n$. In this case the constraint $\sup_{\vartheta \in \Theta_1}$ is irrelevant, i.e. it is equivalent to $\sup_{\vartheta \in \Theta}$ with $\Theta = \Theta_0 \cup \Theta_1$. In this case it is obvious that the test statistic is less than or equal to one.

If $\Theta_0 = \{\vartheta_0\}$ and $\Theta_0 \subset \overline{\Theta_1}$, then the test statistic reduces to

Bemerkung 4.3.9

Prinzipiell benutzen wir den Quotienten der Likelihood-Werte. Allerdings können wir auch die Teststatistik insgesamt logarithmieren, womit wir zu der Differenz der Log-Likelihood-Werte übergehen.

Remark 4.3.9

In principal, we use the ratio of the likelihood values. However, we can also take the logarithm of the whole test statistic, which leads to the difference of the log-likelihood values.

Beispiel 4.3.10

Seien X_1, \dots, X_n u.i. exponentialverteilt mit Parameter ϑ .
Wir wollen testen:

$$\begin{aligned} H_0 : \vartheta \in \Theta_0 &= \{\vartheta_0\} \\ H_1 : \vartheta \in \Theta_1 &= (\vartheta_0; \infty) \end{aligned}$$

Die Likelihood-Funktion ist

$$L(\vartheta|\mathbf{X}) = \prod_{i=1}^n f_\vartheta(X_i) = \vartheta^n \cdot \exp\left(-\vartheta \cdot \sum_{i=1}^n X_i\right).$$

Example 4.3.10

Let X_1, \dots, X_n be i.i. exponentially distributed with parameter ϑ .
We want to test:

The likelihood function is

Im Beispiel 2.4.6 auf Seite 65 haben wir den Maximum-Likelihood-Schätzer für die Exponentialverteilung berechnet, samt der Ableitung der Log-Likelihood-Funktion. Daraus ergibt sich unmittelbar

$$\sup_{\vartheta \in \Theta = [\vartheta_0; \infty)} L(\vartheta|\mathbf{X}) = \begin{cases} (\bar{X}_n)^{-n} \cdot e^{-n \cdot \frac{1}{\bar{X}_n}} & \frac{1}{\bar{X}_n} \geq \vartheta_0, \\ \vartheta_0^n \cdot e^{-n \cdot \vartheta_0 \cdot \bar{X}_n} & \frac{1}{\bar{X}_n} < \vartheta_0 \end{cases}$$

und für die Teststatistik ergibt sich

$$T(\mathbf{X}) = \frac{L(\vartheta_0|\mathbf{X})}{\sup_{\vartheta \in \Theta = [\vartheta_0; \infty)} L(\vartheta|\mathbf{X})} = \begin{cases} \vartheta_0^n \cdot (\bar{X}_n)^n \cdot \exp(-n \cdot (\vartheta_0 \cdot \bar{X}_n - 1)) & \frac{1}{\bar{X}_n} \geq \vartheta_0, \\ 1 & \frac{1}{\bar{X}_n} < \vartheta_0. \end{cases}$$

Weil

$$\frac{d}{d\bar{X}_n} ((\bar{X}_n)^n \cdot e^{-n \cdot \vartheta_0 \cdot \bar{X}_n}) = n \cdot (\bar{X}_n)^{n-1} \cdot e^{-n \cdot \vartheta_0 \cdot \bar{X}_n} \cdot (1 - \vartheta_0 \cdot \bar{X}_n)$$

positiv ist für $\frac{1}{\bar{X}_n} > \vartheta_0$, ist die Teststatistik $T(\mathbf{X})$ eine nicht-abnehmende Funktion von \bar{X}_n , wie in Abbildung 4.3.2 illustriert.

Deswegen hat der kritische Bereich des Likelihood-Quotienten-Tests die Form

$$C_1 = \left\{ \sum_{i=1}^n X_i \leq C \right\},$$

wobei $C = \frac{\tau_\alpha}{2 \cdot n \cdot \vartheta_0}$, τ_α das α -Quantil der $\chi_{2 \cdot n}^2$ -Verteilung und α das gewünschte Signifikanzniveau ist, weil laut Satz 3.2.1 $2 \cdot n \cdot \vartheta_0 \cdot \bar{X}_n$ unter H_0 gleich $\chi_{2 \cdot n}^2$ -verteilt ist.

We have calculated the maximum likelihood estimator for the exponential distribution in Example 2.4.6 on page 65 and also the derivative of the log-likelihood function. Thus, we immediately obtain

and we obtain for the test statistic

Since

is positive for $\frac{1}{\bar{X}_n} > \vartheta_0$, the test statistic $T(\mathbf{X})$ is a non-decreasing function of \bar{X}_n , as can be seen in Figure 4.3.2.

Therefore the critical region of the likelihood ratio test has the form

where $C = \frac{\tau_\alpha}{2 \cdot n \cdot \vartheta_0}$ and τ_α is the α -quantile of the $\chi_{2 \cdot n}^2$ -distribution and α is the desired significance level. That is because of Theorem 3.2.1 which states that $2 \cdot n \cdot \vartheta_0 \cdot \bar{X}_n$ is $\chi_{2 \cdot n}^2$ -distributed under H_0 .

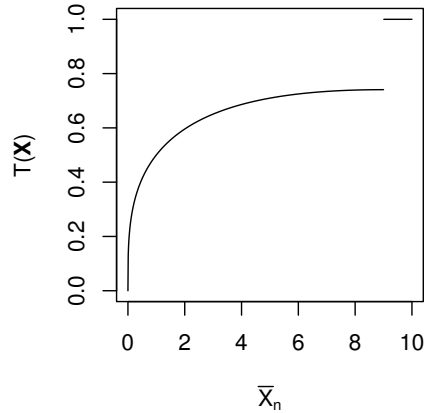


Figure 4.3.2: Teststatistik für Exponentialverteilung
Test statistic for exponential distribution

Proposition 4.3.11

Seien X_1, \dots, X_n u.i. $\mathcal{N}(\mu, \sigma^2)$ -verteilt.

Der t-Test aus Definition 4.2.1 bzw. 4.2.5 ist der Likelihood-Quotienten-Test der Hypothese

$H_0 : \mu = \mu_0$ gegen die Alternative $H_1 : \mu > \mu_0$ oder $H_1 : \mu \neq \mu_0$.

Beweis:

Wir beschränken uns auf den Fall der zweiseitigen Alternative $\mu \neq \mu_0$.

Es ist aus Satz 2.4.4 bekannt, dass

Proposition 4.3.11

Let X_1, \dots, X_n be i.i. $\mathcal{N}(\mu, \sigma^2)$ -distributed.

The t-test of Definition 4.2.1 or 4.2.5,

respectively, is the likelihood ratio test of the hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu > \mu_0$ or $H_1 : \mu \neq \mu_0$.

Proof:

We restrict ourselves to the two-sided alternative $\mu \neq \mu_0$.

It is known from Theorem 2.4.4 that

$$\bar{X}_n = \frac{1}{n} \cdot \sum_{j=1}^n X_j,$$

$$\hat{\sigma}^2 = \frac{1}{n} \cdot \sum_{j=1}^n (X_j - \bar{X}_n)^2$$

die Maximum-Likelihood-Schätzer für μ und σ^2 sind.

Für bekannten Mittelwert $\mu = \mu_0$ haben wir den Maximum-Likelihood-Schätzer

are the maximum likelihood estimators of μ and σ^2 .

For known mean $\mu = \mu_0$, we have the maximum likelihood estimator

$$\hat{\sigma}_0^2 = \frac{1}{n} \cdot \sum_{j=1}^n (X_j - \mu_0)^2.$$

Die Teststatistik bzw. der Likelihood-Quotient ist

The test statistic and the likelihood ratio, respectively, is

$$\begin{aligned} T(\mathbf{X}) &= \frac{L(\mu_0, \hat{\sigma}_0^2 | \mathbf{X})}{L(\bar{X}_n, \hat{\sigma}^2 | \mathbf{X})} \\ &= \prod_{j=1}^n \left\{ \frac{\frac{1}{\sqrt{2\pi\hat{\sigma}_0^2}} \cdot \exp\left(-\frac{(X_j - \mu_0)^2}{2\hat{\sigma}_0^2}\right)}{\frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \cdot \exp\left(-\frac{(X_j - \bar{X}_n)^2}{2\hat{\sigma}^2}\right)} \right\} \\ &= \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{\frac{n}{2}} \cdot \frac{\exp\left[-\frac{1}{2\hat{\sigma}_0^2} \cdot \sum_{j=1}^n (X_j - \mu_0)^2\right]}{\exp\left[-\frac{1}{2\hat{\sigma}^2} \cdot \sum_{j=1}^n (X_j - \bar{X}_n)^2\right]} \\ &= \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{\frac{n}{2}} \end{aligned}$$

wegen der Definition von $\hat{\sigma}^2$ und $\hat{\sigma}_0^2$.

Dies bedeutet, dass der Likelihood-Quotienten-Test den Annahmebereich

due to the definition of $\hat{\sigma}^2$ and $\hat{\sigma}_0^2$.

This means that the likelihood ratio test has the acceptance region

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \geq c_\alpha,$$

mit einem c_α gemäss dem Signifikanzniveau α . Das Ungleichheitszeichen muss in diese Richtung sein, da die Teststatistik eines Likelihood-Quotienten-Tests stets größer ist unter H_0 als unter H_1 . Weil $\hat{\sigma}_0^2 = \hat{\sigma}^2 + (\mu_0 - \bar{X}_n)^2$, ist dies gleichbedeutend mit

$$\frac{(\bar{X}_n - \mu_0)^2}{\hat{\sigma}^2} \leq \frac{1}{c_\alpha} - 1,$$

das heißt, wenn

$$\sqrt{n} \frac{|\bar{X}_n - \mu_0|}{\hat{s}} \leq \left\{ (n-1) \cdot \left(\frac{1}{c_\alpha} - 1 \right) \right\}^{\frac{1}{2}}$$

mit $\hat{s}^2 = \frac{n}{n-1} \cdot \hat{\sigma}^2$.

Der Likelihood-Quotienten-Test akzeptiert die Hypothese genau dann, wenn die zweiseitige t-Statistik unter einer gewissen Grenze ist, die vom Signifikanzniveau abhängt.

Betrachtet man die einseitige Alternative, so muss man lediglich noch die Fallunterscheidung $\bar{X}_n \leq \mu_0$ und $\bar{X}_n > \mu_0$ beachten. ■

Bemerkung 4.3.12

Obwohl die Likelihood-Quotienten-Tests vom Neyman-Pearson Test, der gleichmäßig am besten ist, abgeleitet sind, sind sie **nicht** gleichmäßig beste Tests!

with a c_α chosen according to the significance level α . It must be that inequality sign, because the test statistic of a likelihood ratio test is always higher under H_0 than under H_1 . Since $\hat{\sigma}_0^2 = \hat{\sigma}^2 + (\mu_0 - \bar{X}_n)^2$, this is equivalent to

that is when

with $\hat{s}^2 = \frac{n}{n-1} \cdot \hat{\sigma}^2$.

The likelihood ratio test accepts the hypothesis exactly when the two-sided t-statistic is below a certain bound, which is determined by the significance level.

If we investigate the one-sided alternative, then we just have to consider the two cases $\bar{X}_n \leq \mu_0$ and $\bar{X}_n > \mu_0$. ■

Remark 4.3.12

Although likelihood ratio tests were derived from the Neyman-Pearson test, which is uniformly most powerful, they are **not** uniformly most powerful.

Als Beispiel betrachten wir X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma^2)$ -verteilt mit bekannter Varianz $\sigma^2 > 0$. Testen wollen wir auf dem $\alpha = 0,1$ Signifikanzniveau:

$$H_0 : \mu = \mu_0,$$

$$H_1 : \mu \neq \mu_0.$$

Ähnlich wie in Proposition 4.3.11 können wir zeigen, dass der Likelihood-Quotienten-Test die Teststatistik $T(\mathbf{X}) = \sqrt{n} \cdot \frac{\bar{X}_n - \mu_0}{\sigma}$ und Annahmeregion $C_0^{(1)} = \left\{ \phi_{\frac{\alpha}{2}} \leq T(\mathbf{X}) \leq \phi_{1-\frac{\alpha}{2}} \right\}$ mit $\phi_{\frac{\alpha}{2}}$ und $\phi_{1-\frac{\alpha}{2}}$ die entsprechenden Quantile der $\mathcal{N}(0, 1)$ -Verteilung, hat.

Als zweiten Test betrachten wir die gleiche Teststatistik, aber der Annahmeregion sei $C_0^{(2)} = \left\{ \phi_{\frac{\alpha}{4}} \leq T(\mathbf{X}) \leq \phi_{1-\frac{3\alpha}{4}} \right\}$, wobei $\phi_{\frac{\alpha}{4}}$ und $\phi_{1-\frac{3\alpha}{4}}$ wiederum die entsprechenden Quantile der $\mathcal{N}(0, 1)$ -Verteilung sind.

Wir betrachten nun die Verteilung mit Mittelwert $\mu = \mu_1 = \mu_0 + \phi_{1-\frac{\alpha}{2}} > \mu_0$ und vergleichen die Macht der beiden Tests

$$\begin{aligned} & \mathcal{P}_{\mu_1, \sigma^2} \left(\sqrt{n} \cdot \frac{\bar{X}_n - \mu_0}{\sigma} \in C_0^{(2)} \right) - \mathcal{P}_{\mu_1, \sigma^2} \left(\sqrt{n} \cdot \frac{\bar{X}_n - \mu_0}{\sigma} \in C_0^{(1)} \right) \\ &= \sqrt{n} \cdot \left\{ \int_{\phi_{1-\frac{3\alpha}{4}}}^{\phi_{1-\frac{\alpha}{2}}} \frac{\bar{X}_n - \mu_0}{\sigma} d\mathcal{P}_{\mu_1, \sigma^2} - \int_{\phi_{\frac{\alpha}{2}}}^{\phi_{\frac{\alpha}{2}}} \frac{\bar{X}_n - \mu_0}{\sigma} d\mathcal{P}_{\mu_1, \sigma^2} \right\} \\ &= \left[\Phi(0) - \Phi \left(\phi_{1-\frac{3\alpha}{4}} - \phi_{1-\frac{\alpha}{2}} \right) \right] - \left[\Phi \left(\phi_{\frac{\alpha}{2}} - \phi_{1-\frac{\alpha}{2}} \right) - \Phi \left(\phi_{\frac{\alpha}{4}} - \phi_{1-\frac{\alpha}{2}} \right) \right] > 0, \end{aligned}$$

As an example, we consider X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma^2)$ -distributed with known variance $\sigma^2 > 0$. We want to test with significance level $\alpha = 0.1$:

Similarly to Proposition 4.3.11, we can show that the likelihood ratio test has the test statistic $T(\mathbf{X}) = \sqrt{n} \cdot \frac{\bar{X}_n - \mu_0}{\sigma}$ and the acceptance region $C_0^{(1)} = \left\{ \phi_{\frac{\alpha}{2}} \leq T(\mathbf{X}) \leq \phi_{1-\frac{\alpha}{2}} \right\}$ with $\phi_{\frac{\alpha}{2}}$ and $\phi_{1-\frac{\alpha}{2}}$ the corresponding quantiles of the $\mathcal{N}(0, 1)$ -distribution.

As a second test, we take the same test statistic, but now the acceptance region is $C_0^{(2)} = \left\{ \phi_{\frac{\alpha}{4}} \leq T(\mathbf{X}) \leq \phi_{1-\frac{3\alpha}{4}} \right\}$, where $\phi_{\frac{\alpha}{4}}$ and $\phi_{1-\frac{3\alpha}{4}}$ again are the corresponding quantiles of the $\mathcal{N}(0, 1)$ -distribution.

Now we investigate the distribution with mean $\mu = \mu_1 = \mu_0 + \phi_{1-\frac{\alpha}{2}} > \mu_0$ and compare the power of both tests

mit Φ der Verteilungsfunktion der Standardnormalverteilung, weil

$$\sqrt{n} \cdot \frac{\bar{X}_n - \mu_0}{\sigma} \sim \mathcal{N}(\mu_1 - \mu_0, 1) = \mathcal{N}(\phi_{1-\frac{\alpha}{2}}, \sigma^2) \text{ und}$$

$$\phi_{\frac{\alpha}{4}} - \phi_{1-\frac{\alpha}{2}} < \phi_{\frac{\alpha}{2}} - \phi_{1-\frac{\alpha}{2}} < \phi_{1-\frac{3\alpha}{4}} - \phi_{1-\frac{\alpha}{2}} < 0.$$

Damit haben wir gezeigt, dass unser zweiter Test eine größere Macht als der Likelihood-Quotienten-Test für μ_1 hat.

Wir können allerdings einen unverzerrten Test haben:

Definition 4.3.13 (Unverzerrter Test)

Ein Test T mit Signifikanzniveau α von $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$, wird unverzerrt genannt, wenn die Macht $Q(T, \vartheta) \geq \alpha$ für alle $\vartheta \in \Theta_1$.

Bemerkung 4.3.14

Dies bedeutet, dass die Macht nie kleiner ist als das Signifikanzniveau. Die Wahrscheinlichkeit für die Ablehnung von H_0 für $\vartheta \in \Theta_1$ ist mindestens so hoch, wie die für $\vartheta \in \Theta_0$.

with Φ the distribution function of the standard normal distribution, because

$$\sqrt{n} \cdot \frac{\bar{X}_n - \mu_0}{\sigma} \sim \mathcal{N}(\mu_1 - \mu_0, 1) = \mathcal{N}(\phi_{1-\frac{\alpha}{2}}, \sigma^2) \text{ and}$$

$$\phi_{\frac{\alpha}{4}} - \phi_{1-\frac{\alpha}{2}} < \phi_{\frac{\alpha}{2}} - \phi_{1-\frac{\alpha}{2}} < \phi_{1-\frac{3\alpha}{4}} - \phi_{1-\frac{\alpha}{2}} < 0.$$

Therefore, we have shown that the second test has a higher power for μ_1 than the likelihood ratio test.

We can, however, have an unbiased test:

Definition 4.3.13 (Unbiased test)

A test T at significance level α for testing $H_0 : \vartheta \in \Theta_0$ against $H_1 : \vartheta \in \Theta_1$ is said to be unbiased if the power $Q(T, \vartheta) \geq \alpha$ for all $\vartheta \in \Theta_1$.

Remark 4.3.14

This means that the power is never less than the significance level. The probability of rejection of H_0 for $\vartheta \in \Theta_1$ is at least as high as for $\vartheta \in \Theta_0$.

Bemerkung 4.3.15

Es gibt viele Tests, die unverzerrt sind, z.B. t-Tests.

Allerdings gibt es Likelihood-Quotienten-Tests, die nicht unverzerrt sind:

Seien X_1, \dots, X_n Zufallsvariablen mit gemeinsamer Verteilung \mathcal{P}_i bzw. Dichte f_i mit $i = 0; 1; 2$ und $A = [0; 1]^n$ und $B = [1; 2]^n$.

Die Dichten seien folgendermaßen:

$$f_0(\mathbf{x}) = \begin{cases} 0,95 & \mathbf{x} \in A \\ 0,05 & \mathbf{x} \in B \\ 0 & \mathbf{x} \notin A \cup B \end{cases} \quad f_1(\mathbf{x}) = \begin{cases} 0,5 & \mathbf{x} \in A \cup B \\ 0 & \mathbf{x} \notin A \cup B \end{cases} \quad f_2(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in A \\ 0 & \mathbf{x} \notin A \end{cases}$$

Wir wollen auf dem $\alpha = 5\%$ -Niveau die folgende Hypothese testen:

Remark 4.3.16

There are many tests which are unbiased, e.g. t-tests.

However, there are likelihood ratio test which are not unbiased:

Let X_1, \dots, X_n be random variables with common distribution \mathcal{P}_i which has the density f_i with $i = 0; 1; 2$ and $A = [0; 1]^n$ and $B = [1; 2]^n$.

The densities are defined as follows:

We want to test with significance level $\alpha = 5\%$ the following hypothesis:

$$H_0 : \mathcal{P} = \mathcal{P}_0 \\ H_1 : \mathcal{P} \in \{\mathcal{P}_1; \mathcal{P}_2\}$$

Der Likelihood-Quotient ist

The likelihood ratio is

$$\frac{f_0(\mathbf{x})}{\max\{f_1(\mathbf{x}); f_2(\mathbf{x})\}} = \begin{cases} 0,95 & \mathbf{x} \in A \\ 0,1 & \mathbf{x} \in B \end{cases}$$

Weil der Likelihood-Quotient für $\mathbf{x} \in B$ kleiner ist und $\mathcal{P}_0(B) = 0,05$, ist dies genau der Ablehnbereich C_1 .

Ist die wahre Verteilung \mathcal{P}_2 , so ist $\mathcal{P}_2(\mathbf{x} \in C_1) = \mathcal{P}_2(\mathbf{x} \in B) = 0 < \alpha!$

Since the likelihood ratio is smaller for $\mathbf{x} \in B$ and $\mathcal{P}_0(B) = 0,05$, this is exactly the rejection region C_1 .

If the true distribution is \mathcal{P}_2 , then $\mathcal{P}_2(\mathbf{x} \in C_1) = \mathcal{P}_2(\mathbf{x} \in B) = 0 < \alpha!$

4.4 χ^2 -Tests χ^2 -tests

In diesem Abschnitt werden wir den χ^2 -Test studieren. Zuerst motivieren wir und leiten den χ^2 -Test her. Danach betrachten wir zwei Anwendungen: Goodness-of-Fit-Tests und Unabhängigkeitstests.

In this section, we are going to study the χ^2 -test. First, we motivate and derive the χ^2 -test. Then, we show two applications: Goodness-of-Fit tests and independence tests.

4.4.1 Herleitung des χ^2 -Tests Derivation of the χ^2 -test

Als Ausgangspunkt für den χ^2 -Test benutzen wir die Multinomialverteilung.

As a starting point for the χ^2 -test, we use the multinomial distribution.

Definition 4.4.1

Modell:

Sei $\mathbf{Z} = (Z_1, \dots, Z_d)$ multinomialverteilt mit Parametern (n, p_1, \dots, p_d) , wobei $p_k \geq 0$, $p_1 + \dots + p_d = 1$ und $Z_1 + \dots + Z_d = n$.

Definition 4.4.1

Model:

Let $\mathbf{Z} = (Z_1, \dots, Z_d)$ be multinomially distributed with parameters (n, p_1, \dots, p_d) , where $p_k \geq 0$, $p_1 + \dots + p_d = 1$ and $Z_1 + \dots + Z_d = n$.

Problem:

Problem:

$$H_0 : \vartheta = (p_1, \dots, p_d) \in \Theta_0 = \left\{ \vartheta_0 = (p_1^{(0)}, \dots, p_d^{(0)}) \right\}$$

$$H_1 : \vartheta \neq \vartheta_0.$$

Teststatistik:

Test statistic:

$$\chi^2 = \sum_{k=1}^d \frac{(Z_k - n \cdot p_k^{(0)})^2}{n \cdot p_k^{(0)}}$$

Annahmehereich:

Acceptance region:

$$\chi^2 \leq c_\alpha,$$

wobei c_α das $(1 - \alpha)$ -Quantil der χ_{d-1}^2 -Verteilung ist.

where c_α is the $(1 - \alpha)$ -quantile of the χ_{d-1}^2 -distribution.

Bemerkung 4.4.2

Wir können die χ^2 -Statistik auf die folgende Weise motivieren:

Die Hypothese wird nicht abgelehnt, wenn die beobachteten Werte sich nicht zu sehr von den Erwartungswerten unter der Hypothese unterscheiden. Als Maß der Abweichung betrachten wir ein gewichtetes Mittel der quadratischen Differenz $(Z_k - n \cdot p_k^{(0)})^2$, wobei die Gewichte $1 / (n \cdot p_k^{(0)})$ so gewählt sind, dass wir eine möglichst einfache Verteilung der Teststatistik erhalten.

Im Satz 4.4.4 werden wir mit Hilfe von Proposition 4.4.3 beweisen, dass χ^2 die angegebene asymptotische Verteilung hat.

Remark 4.4.2

We can motivate the χ^2 -statistic in the following way:

The hypothesis is not rejected if the observed values do not differ too much from the expectations under the hypothesis. As measure of the deviation, we consider a weighted mean of the quadratic difference $(Z_k - n \cdot p_k^{(0)})^2$, where the weights $1 / (n \cdot p_k^{(0)})$ are chosen to give an as simple distribution of the test statistic as possible. We will show in Theorem 4.4.4, using Proposition 4.4.3, that χ^2 has the announced asymptotic distribution.

Proposition 4.4.3

Sei

Proposition 4.4.3

Let

$$U_k = \frac{1}{\sqrt{n}} \cdot (Z_k - n \cdot p_k^{(0)}), \quad k = 1, \dots, d$$

wobei $\mathbf{Z} = (Z_1, \dots, Z_d)$ multinomialverteilt ist mit Parametern $(n, p_1^{(0)}, \dots, p_d^{(0)})$.
Dann gilt für den Zufallsvektor $\mathbf{U} = (U_1, \dots, U_d)^T$:

where $\mathbf{Z} = (Z_1, \dots, Z_d)$ is multinomially distributed with parameters $(n, p_1^{(0)}, \dots, p_d^{(0)})$.
Then holds for the random vector $\mathbf{U} = (U_1, \dots, U_d)^T$:

$$\mathbf{U} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma), \quad n \rightarrow \infty,$$

wobei $\Sigma = (\sigma_{kl})_{1 \leq k, l \leq d}$ und

where $\Sigma = (\sigma_{kl})_{1 \leq k, l \leq d}$ and

$$\sigma_{kl} = \text{COV}_{\mathfrak{D}_0}(\mathbf{U}_k, \mathbf{U}_l) = \begin{cases} p_k^{(0)} \cdot (1 - p_k^{(0)}) & k = l \\ -p_k^{(0)} \cdot p_l^{(0)} & k \neq l. \end{cases}$$

Beweis:

Seien X_1, \dots, X_n u.i.v. Zufallsvariablen mit Werten in der endlichen Menge $\{\xi_1, \dots, \xi_d\}$ und

$$\mathcal{P}(X_j = \xi_k) = p_k^{(0)}; \quad j = 1, \dots, n; \quad k = 1, \dots, d.$$

Wir betrachten die Vektoren

$$\boldsymbol{\varepsilon}^{(j)} = (\varepsilon_1^{(j)}, \dots, \varepsilon_d^{(j)})^T,$$

gegeben durch

$$\varepsilon_k^{(j)} = \begin{cases} 1 & X_j = \xi_k \\ 0 & X_j \neq \xi_k. \end{cases}$$

Seien

$$\mathbf{Z}^* = \sum_{j=1}^n \boldsymbol{\varepsilon}_k^{(j)},$$

$$\mathbf{U}_k^* = \frac{1}{\sqrt{n}} \cdot (\mathbf{Z}_k^* - n \cdot p_k^{(0)}).$$

Proof:

Let X_1, \dots, X_n be i.i.d. random variables with values in the finite set $\{\xi_1, \dots, \xi_d\}$ and

We consider the vectors

given by

We let

(Z_1^*, \dots, Z_d^*) und (U_1^*, \dots, U_d^*) haben dieselbe Verteilung wie (Z_1, \dots, Z_d) und (U_1, \dots, U_d) .
Die Zufallsvektoren $\boldsymbol{\varepsilon}^{(1)}, \dots, \boldsymbol{\varepsilon}^{(n)}$ sind u.i.v. mit

(Z_1^*, \dots, Z_d^*) and (U_1^*, \dots, U_d^*) have the same distribution as (Z_1, \dots, Z_d) and (U_1, \dots, U_d) .
The random vectors $\boldsymbol{\varepsilon}^{(1)}, \dots, \boldsymbol{\varepsilon}^{(n)}$ are i.i.d. with

$$\mathcal{E}(\boldsymbol{\varepsilon}^{(j)}) = (p_1^{(0)}, \dots, p_d^{(0)})^T = \boldsymbol{\mathfrak{D}}_0^T.$$

Deswegen ist

Therefore is

$$\mathbf{U}^* = (U_1^*, \dots, U_d^*)^T = \frac{1}{\sqrt{n}} \cdot \sum_{j=1}^n (\boldsymbol{\varepsilon}^{(j)} - \boldsymbol{\mathfrak{D}}_0^T)$$

die standardisierte Summe von u.i.v. Zufallsvariablen mit Kovarianzmatrix Σ , gegeben durch

the standardized sum of i.i.d. random variables with covariance matrix Σ given by

$$\sigma_{kl} = \text{COV}(\boldsymbol{\varepsilon}_k^{(j)}, \boldsymbol{\varepsilon}_l^{(j)}) = \mathcal{E} \left[(\boldsymbol{\varepsilon}_k^{(j)} - p_k^{(0)}) \cdot (\boldsymbol{\varepsilon}_l^{(j)} - p_l^{(0)}) \right]$$

$$= \begin{cases} -p_k^{(0)} \cdot p_l^{(0)} & k \neq l, \quad (\boldsymbol{\varepsilon}_k^{(j)} \cdot \boldsymbol{\varepsilon}_l^{(j)} = 0) \\ p_k^{(0)} \cdot (1 - p_k^{(0)}) & k = l. \end{cases}$$

Vom multivariaten zentralen Grenzwertsatz folgt, dass

From the multivariate central limit theorem it follows that

$$\mathcal{L}(\mathbf{U}) = \mathcal{L}(\mathbf{U}^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

Satz 4.4.4

Unter der Hypothese, d.h. wenn $\mathbf{Z} = (Z_1, \dots, Z_d)$ multinomialverteilt ist mit Parameter $(n, p_1^{(0)}, \dots, p_d^{(0)})$, ist die χ^2 -Statistik asymptotisch χ_{d-1}^2 -verteilt:

Theorem 4.4.4

Under the hypothesis, i.e. if $\mathbf{Z} = (Z_1, \dots, Z_d)$ is multinomially distributed with parameter $(n, p_1^{(0)}, \dots, p_d^{(0)})$, is the χ^2 -statistic asymptotically χ_{d-1}^2 -distributed:

$$\chi^2 = \sum_{k=1}^d \frac{(Z_k - n \cdot p_k^{(0)})^2}{n \cdot p_k^{(0)}} \xrightarrow{\mathcal{L}} \chi_{d-1}^2, \quad n \rightarrow \infty.$$

Beweis:

Sei $Y_k = U_k / \sqrt{p_k^{(0)}}$ mit U_k wie in Proposition 4.4.3, so dass

$$\chi^2 = \sum_{k=1}^d Y_k^2$$

$$\mathbf{Y} = (Y_1, \dots, Y_d)^T \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tilde{\Sigma}), \quad \tilde{\Sigma} = (\tilde{\sigma}_{kl})_{1 \leq k, l \leq d}$$

mit

$$\tilde{\sigma}_{kl} = COV_{\vartheta_0}(Y_k, Y_l) = \begin{cases} 1 - p_k^{(0)} & k = l \\ -\sqrt{p_k^{(0)} \cdot p_l^{(0)}} & k \neq l. \end{cases}$$

Als Kovarianzmatrix ist $\tilde{\Sigma}$ symmetrisch und nicht-negativ definit, d.h. hat nicht-negative Eigenwerte $\lambda_1 \geq \dots \geq \lambda_d \geq 0$ und ein entsprechendes Orthonormalsystem von Eigenvektoren $\mathbf{e}_1, \dots, \mathbf{e}_d$.

Wir sehen sehr einfach, dass

$$\mathbf{e}_d = \left(\sqrt{p_1^{(0)}}, \dots, \sqrt{p_d^{(0)}} \right)^T$$

ein Eigenvektor ist, mit Eigenwert

$$\lambda_d = 0.$$

Sei $\mathbf{v} = (v_1, \dots, v_d)^T$ ein zu \mathbf{e}_d orthogonaler Vektor, d.h. $\sum_{i=1}^d \sqrt{p_i^{(0)}} \cdot v_i = 0$.

$$(\tilde{\Sigma} \cdot \mathbf{v})_k = (1 - p_k^{(0)}) \cdot v_k - \sum_{l \neq k} \sqrt{p_k^{(0)} \cdot p_l^{(0)}} \cdot v_l = v_k - \sqrt{p_k^{(0)}} \cdot \sum_{l=1}^d \sqrt{p_l^{(0)}} \cdot v_l = v_k,$$

d.h. ist stets ein Eigenvektor zum Eigenwert 1. Dies bedeutet, dass $\mathbf{e}_1, \dots, \mathbf{e}_{d-1}$ Eigenvektoren mit den Eigenwerten $\lambda_1 = \dots = \lambda_{d-1} = 1$ sind.

Proof:

We put $Y_k = U_k / \sqrt{p_k^{(0)}}$ with U_k from Proposition 4.4.3, such that

with

As a covariance matrix is $\tilde{\Sigma}$ symmetric and non-negative definite, i.e. has non-negative eigenvalues $\lambda_1 \geq \dots \geq \lambda_d \geq 0$ and a corresponding orthonormal system of eigenvectors $\mathbf{e}_1, \dots, \mathbf{e}_d$.

We see immediately that

is an eigenvector with eigenvalue

Let $\mathbf{v} = (v_1, \dots, v_d)^T$ be a vector which is orthogonal to \mathbf{e}_d , i.e. $\sum_{i=1}^d \sqrt{p_i^{(0)}} \cdot v_i = 0$.

i.e. it is always an eigenvector with eigenvalue 1. This means that $\mathbf{e}_1, \dots, \mathbf{e}_{d-1}$ are eigenvectors with eigenvalues $\lambda_1 = \dots = \lambda_{d-1} = 1$.

Sei O die Orthogonalmatrix mit Spalten $\mathbf{e}_1, \dots, \mathbf{e}_d$ und

Let O be the orthogonal matrix with columns $\mathbf{e}_1, \dots, \mathbf{e}_d$ and

$$\mathbf{V} = (V_1, \dots, V_d)^T = O^T \cdot \mathbf{Y}.$$

Es folgt wegen der Orthogonalität von O , dass:It follows from the orthogonality of O that:

$$\chi^2 = \sum_{k=1}^d Y_k^2 = \sum_{k=1}^d V_k^2$$

und

and

$$\mathbf{V}_d = \mathbf{e}_d^T \cdot \mathbf{Y} = \sum_{k=1}^d U_k = \frac{1}{\sqrt{n}} \cdot \sum_{k=1}^d (Z_k - n \cdot p_k^{(0)}) = 0,$$

weil

since

$$\begin{aligned} Z_1 + \dots + Z_d &= n, \\ p_1^{(0)} + \dots + p_d^{(0)} &= 1. \end{aligned}$$

Wir erhalten

We obtain

$$COV_{\vartheta_0}(V_k, V_l) = \mathcal{E}_{\vartheta_0}(\mathbf{e}_k^T \cdot \mathbf{Y} \cdot \mathbf{Y}^T \cdot \mathbf{e}_l) = \mathbf{e}_k^T \cdot \tilde{\Sigma} \cdot \mathbf{e}_l = \lambda_l \mathbf{e}_k^T \cdot \mathbf{e}_l = \begin{cases} 1 & k = l < d \\ 0 & \text{sonst / otherwise.} \end{cases}$$

Wir haben also gezeigt, dass

So, we have shown that:

$$\chi^2 = \sum_{k=1}^{d-1} V_k^2$$

mit

with

$$(V_1, \dots, V_{d-1})^T \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_{d-1}),$$

wobei I_{d-1} die $(d-1)$ -dimensionale Einheitsmatrix ist. Der Satz folgt nun unmittelbar aus der Definition der χ^2 -Verteilung. ■

where I_{d-1} is the $(d-1)$ -dimensional unit matrix. Now the theorem follows immediately from the definition of the χ^2 -distribution. ■

Bemerkung 4.4.5

Wir wollen jetzt zeigen, dass wir den χ^2 -Test aus dem Likelihood-Quotienten-Test für multinomialverteilte Zufallsvariablen herleiten können.

Seien X_1, \dots, X_n multinomialverteilt mit Parameter ϑ aus dem Parameterraum $\Theta = \{(p_1, \dots, p_d) \mid p_1, \dots, p_d \geq 0; p_1 + \dots + p_d = 1\}$, d.h. $X_i : \Omega \rightarrow \{1, \dots, d\}$.

$Z_k := \#\{i \mid X_i = k\}$ seien die Anzahl der X_i , die gleich k sind, mit $k \in \{1, \dots, d\}$.

Die Likelihood-Funktion ist

$$L(\vartheta|\mathbf{Z}) = \mathcal{P}_\vartheta(Z_1, \dots, Z_d) = \binom{n}{Z_1 \dots Z_d} \prod_{k=1}^d p_k^{Z_k}.$$

Analog zum Beispiel 2.4.5 auf Seite 64 ergibt sich mit Hilfe von $p_d = 1 - p_1 - \dots - p_{d-1}$, dass die Maximum-Likelihood-Schätzer p_k^* gegeben sind durch:

$$p_k^* = \frac{Z_k}{n}, \quad k = 1, \dots, d.$$

Daraus ergibt sich für den Likelihood-Quotient

$$\lambda(\mathbf{Z}) = \frac{L(\vartheta_0|\mathbf{Z})}{\max_{\vartheta \in \Theta} L(\vartheta|\mathbf{Z})} = \prod_{k=1}^d \left(\frac{n \cdot p_k^{(0)}}{Z_k} \right)^{Z_k}.$$

Remark 4.4.5

We will now show that we can derive the χ^2 -test from the likelihood ratio test for multinomially distributed random variables.

Let X_1, \dots, X_n be multinomially distributed with parameter ϑ from the parameter space $\Theta = \{\vartheta = (p_1, \dots, p_d) \mid p_1, \dots, p_d \geq 0; p_1 + \dots + p_d = 1\}$, i.e. $X_i : \Omega \rightarrow \{1, \dots, d\}$.

$Z_k := \#\{i \mid X_i = k\}$ is the number of X_i which are equal to k where $k \in \{1, \dots, d\}$.

The likelihood function is

If we use that $p_d = 1 - p_1 - \dots - p_{d-1}$, then we get the maximum likelihood estimators for p_k^* analogously as in Example 2.4.5 on page 64:

From this, we obtain for the likelihood ratio

Die Nullhypothese $H_0 : \vartheta = \vartheta_0$ wird nicht abgelehnt, falls der Likelihood-Quotient größer als eine gewisse Schranke liegt. Da die Logarithmusfunktion streng monoton steigend ist, kann man auch $\log[\lambda(\mathbf{Z})]$ betrachten.

The null hypothesis $H_0 : \vartheta = \vartheta_0$ is not rejected, if the likelihood ratio is greater than a certain threshold. Since the logarithm is a strictly increasing function, we can also consider $\log[\lambda(\mathbf{Z})]$.

$\log[\lambda(\mathbf{Z})]$ ist eine komplizierte Funktion von \mathbf{Z} , deren Verteilung schwierig zu bestimmen ist. Deswegen wird $\log[\lambda(\mathbf{Z})]$ jetzt durch eine einfachere Zufallsvariable approximiert, die auch eine ähnliche Verteilung hat. Wir benutzen die Taylorentwicklung von $\log(1+x)$ und erhalten

$\log[\lambda(\mathbf{Z})]$ is a complicated function of \mathbf{Z} whose distribution is difficult to determine. Therefore we now approximate $\log[\lambda(\mathbf{Z})]$ by a simpler random variable, which has a similar distribution. We use the Taylor expansion of $\log(1+x)$ and obtain

$$\begin{aligned} \log[\lambda(\mathbf{Z})] &= \sum_{k=1}^d Z_k \cdot \log\left(\frac{n \cdot p_k^{(0)}}{Z_k}\right) \\ &\approx \left\{ \sum_{k=1}^d Z_k \cdot \left(\frac{n \cdot p_k^{(0)}}{Z_k} - 1\right) \right\} - \frac{1}{2} \cdot \left\{ \sum_{k=1}^d Z_k \cdot \left(\frac{n \cdot p_k^{(0)}}{Z_k} - 1\right)^2 \right\} \\ &= -\frac{1}{2} \cdot \sum_{k=1}^d Z_k \cdot \left(\frac{n \cdot p_k^{(0)}}{Z_k} - 1\right)^2 \\ &\approx -\frac{1}{2} \cdot \sum_{k=1}^d \frac{(n \cdot p_k^{(0)} - Z_k)^2}{n \cdot p_k^{(0)}}. \end{aligned}$$

Der erste Term der Taylorreihenentwicklung verschwindet, weil $Z_1 + \dots + Z_d = n$ und $p_1^{(0)} + \dots + p_d^{(0)} = 1$. Die letzte Näherung folgt vom Gesetz der großen Zahlen, weil $\frac{Z_k}{n} \xrightarrow{n \rightarrow \infty} p_k^{(0)}$ unter der Hypothese $H_0 : p_k = p_k^{(0)}$. Somit haben wir gezeigt, dass die χ^2 -Teststatistik eine Approximation der Likelihood-Quotienten-Teststatistik für multinomialverteilte Zufallsvariablen ist.

The first term of the Taylor expansion vanishes due to $Z_1 + \dots + Z_d = n$ and $p_1^{(0)} + \dots + p_d^{(0)} = 1$. The last approximation follows from the law of large numbers, because $\frac{Z_k}{n} \xrightarrow{n \rightarrow \infty} p_k^{(0)}$ under the hypothesis $H_0 : p_k = p_k^{(0)}$. Thus, we have shown that the χ^2 -test statistic is an approximation of the likelihood ratio test statistic for multinomially distributed random variables.

Beispiel 4.4.6

Wir gehen zu Beispiel 1.2.4 auf Seite 24 zurück. Wir wollen mittels eines χ^2 -Tests untersuchen, ob die Phänotypen der Elwedritsche einer Multinomialverteilung folgen mit
 Ph. 1: $p_1^{(0)} = \frac{9}{16}$ (lange Haare, kleiner Mund);
 Ph. 2: $p_2^{(0)} = \frac{3}{16}$ (lange Haare, großer Mund);
 Ph. 3: $p_3^{(0)} = \frac{3}{16}$ (kurze Haare, kleiner Mund);
 Ph. 4: $p_4^{(0)} = \frac{1}{16}$ (kurze Haare, großer Mund).
 Im Laufe der Jahre wurden die folgenden Beobachtungen von Leuten, die im Pfälzer Wald wandern, berichtet:

Phänotyp	Anzahl Beobachtungen
1	480
2	171
3	166
4	79

Durch Benutzung des Schätzers $\hat{p}_i = \frac{z_i}{896}$, $i = 1, \dots, 4$ erhalten wir den Wert 10,67 für die χ^2 -Statistik. Das 95%-Quantil der χ^2_3 -Verteilung ist gleich 7,81, d.h. wir können die Hypothese, dass die Phänotypen der Elwedritschen der oben beschriebenen Multinomialverteilung folgen, auf dem 5%-Niveau ablehnen..

Example 4.4.6

We go back to Example 1.2.4 on page 24. We want to use a χ^2 -test to investigate if the phenotypes of elwedritsches follow a multinomial distribution with
 Ph. 1: $p_1^{(0)} = \frac{9}{16}$ (long-haired, small mouth);
 Ph. 2: $p_2^{(0)} = \frac{3}{16}$ (long-haired, big mouth);
 Ph. 3: $p_3^{(0)} = \frac{3}{16}$ (short-haired, small mouth);
 Ph. 4: $p_4^{(0)} = \frac{1}{16}$ (short-haired, big mouth).
 Over the years, the following observations have been reported from people hiking in the Palatinate forest:

Phenotype	No of observations
1	480
2	171
3	166
4	79

We obtain the value 10.67 for the χ^2 -statistic when we use the estimator $\hat{p}_i = \frac{z_i}{896}$, $i = 1, \dots, 4$. The 95% quantile of the χ^2_3 -distribution is equal to 7.81, i.e. we can reject the hypothesis that the phenotypes of elwedritsches follow the multinomial distribution described above at significance level 5%.

Bemerkung 4.4.7

Im allgemeinen ist die Hypothese nicht ein- sondern höherdimensional, d.h. wir betrachten Hypothesen der Form:

$$H_0: \vartheta \in \Theta_0 = \{(p_1, \dots, p_d) \mid p_k = p_k(\delta), k = 1, \dots, d, \delta \in \Delta \subseteq \mathbb{R}^q\}.$$

Wir nennen q die Dimensionalität der Hypothese, wenn es nicht möglich ist Θ_0 mit weniger als q Parametern zu parameterisieren. Klar ist, dass wir nun auch die Wahrscheinlichkeiten $p_k(\delta)$ unter der Hypothese per Maximum-Likelihood Methode schätzen müssen.

Wie würde also der χ^2 -Test im höherdimensionalen Fall aussehen?

Remark 4.4.7

In general, the hypothesis is not one-, but higher dimensional, i.e. we consider hypotheses of the form:

We call q the dimensionality of the hypothesis, if it is not possible to parameterize Θ_0 with less than q parameters. It is obvious that we now must estimate the probabilities $p_k(\delta)$ for the hypothesis by the maximum likelihood method.

What would the χ^2 -test be like in the higher dimensional case?

Satz 4.4.8

Seien $Z = (Z_1, \dots, Z_d)$ multinomialverteilt mit Parametern (n, p_1, \dots, p_d) . Unter der Hypothese $p_k = p_k(\delta)$, $k = 1, \dots, d$ haben wir, dass

$$\chi^2 = \sum_{k=1}^d \frac{[Z_k - n \cdot p_k(\hat{\delta})]^2}{n \cdot p_k(\hat{\delta})} \xrightarrow{\mathcal{L}} \chi^2_{d-q-1},$$

wobei q die Dimensionalität der Hypothese und $\hat{\delta}$ der Maximum-Likelihood-Schätzer von δ ist.

Theorem 4.4.8

Let $Z = (Z_1, \dots, Z_d)$ be multinomially distributed with parameters (n, p_1, \dots, p_d) . Under the hypothesis $p_k = p_k(\delta)$, $k = 1, \dots, d$, we have that

where q is the dimensionality of the hypothesis and $\hat{\delta}$ is the maximum likelihood estimator of δ .

Der Beweis des Satzes ist ähnlich des Beweises von Satz 4.4.4.

The proof of the theorem is similar to the proof of Theorem 4.4.4.

Bemerkung 4.4.9

Für jeden reellen Parameter $\delta_1 \dots, \delta_q$, den wir schätzen, nimmt die Anzahl der Freiheitsgrade um eins ab.

Allerdings bleibt die asymptotische Verteilung immer noch χ^2 .

Was wir hier nicht untersucht haben, ist die Situation, dass $q > d - 1$ ist!

Remark 4.4.9

For every real parameter $\delta_1 \dots, \delta_q$, which we are estimating, the number of degrees of freedom decreases with one.

However, the asymptotic distribution is still χ^2 .

Be aware of the fact that we did not consider the situation $q > d - 1$!

4.4.2 Goodness-of-Fit Tests

Goodness-of-Fit tests

Hiermit wird getestet, ob die Verteilung der Daten zu einer gegebenen Klasse von Verteilungen gehört.

Problem:

X_1, \dots, X_n u.i.v. mit Werten in X .
Stammt $\mathcal{L}(X_j)$ aus $\{\mathcal{P}_\vartheta \mid \vartheta \in \Theta\}$?

Um dies mit dem χ^2 -Test zu untersuchen, teilen wir X in d disjunkte Teilmenge X_1, \dots, X_d mit $X = \cup_{i=1}^d X_i$ auf und betrachten die Indikatorvariablen

$$\epsilon_k^{(j)} = \begin{cases} 1 & X_j \in X_k \\ 0 & X_j \notin X_k, \end{cases} \quad j = 1, \dots, n, \quad k = 1, \dots, d,$$

und die Zählvariablen

$$Z_k = \sum_{j=1}^n \epsilon_k^{(j)}, \quad k = 1, \dots, d.$$

Wenn $\mathcal{L}(X_j) \in \{\mathcal{P}_\vartheta \mid \vartheta \in \Theta\}$, dann ist $Z = (Z_1, \dots, Z_d)$ multinomialverteilt mit Parametern $(n, p_1(\vartheta), \dots, p_d(\vartheta))$ für ein $\vartheta \in \Theta$.

We test here, if the distribution of the data belongs to a given class of distributions.

Problem:

X_1, \dots, X_n i.i.d. with values in X .
Does $\mathcal{L}(X_j)$ belong to $\{\mathcal{P}_\vartheta \mid \vartheta \in \Theta\}$?

To investigate this with the χ^2 -test, we divide X into d disjunct subsets X_1, \dots, X_d with $X = \cup_{i=1}^d X_i$ and consider the indicator variables

and the counting variables

If $\mathcal{L}(X_j) \in \{\mathcal{P}_\vartheta \mid \vartheta \in \Theta\}$, then is $Z = (Z_1, \dots, Z_d)$ multinomially distributed with parameters $(n, p_1(\vartheta), \dots, p_d(\vartheta))$ for a $\vartheta \in \Theta$.

Wir sind somit in der Situation von Satz 4.4.8, wobei q die Dimension von Θ ist, d.h. wir können die Hypothese $\mathcal{L}(X_j) \in \{\mathcal{P}_\vartheta \mid \vartheta \in \Theta\}$ mittels eines χ^2 -Tests untersuchen.

Bemerkung 4.4.10

Die Anzahl d der Teilmengen soll zwar mit steigender Anzahl n von Daten wachsen, aber nicht ganz so schnell wie n . Ferner sollte $n \cdot p_i(\vartheta) \xrightarrow{n \rightarrow \infty} \infty$ für alle $\vartheta \in \Theta$ und alle $i = 1, \dots, d$ gelten. Falls $n \cdot p_i(\vartheta) < 5$ für ein i , so sollte man diese Teilmenge mit einer benachbarten zusammenfassen.

Bezüglich der genauen Wahl der Teilmengen gibt es kein allgemein gültiges Rezept. Hat man allerdings eine Ahnung, was als Alternativverteilungen in Frage kommt, dann sollte man viele Teilmengen in Bereichen haben, wo die Verteilungsgruppen sich stark unterscheiden.

Beispiel 4.4.11

Während der Wanderung zurück zur Universität Kaiserslautern, treffen die StudentInnen einen Biologen, der das Muster von gesunden und kranken Bäumen im Pfälzer Wald studiert. Die Forschung ist auf die Wurzelfäule, insbesondere die Länge verrotteter Wurzeln (in Meter) von $n = 110$ kranken Bäumen, konzentriert:

We are therefore in the situation of Theorem 4.4.8, where q is the dimension of Θ , i.e. we can test the hypothesis $\mathcal{L}(X_j) \in \{\mathcal{P}_\vartheta \mid \vartheta \in \Theta\}$ by a χ^2 -test.

Remark 4.4.10

The number d of subsets should increase with the number n of data, but not as fast as n . Furthermore should hold $n \cdot p_i(\vartheta) \xrightarrow{n \rightarrow \infty} \infty$ for all $\vartheta \in \Theta$ and all $i = 1, \dots, d$. If $n \cdot p_i(\vartheta) < 5$ for some i , then we should union this subset with one of the neighboring subsets.

There is no general valid recipe how to choose the subsets. However, if one has some idea about the possible alternative distributions, then there should be many subsets in areas where the groups of distributions are rather different.

Example 4.4.11

During their walk back to the university of Kaiserslautern, the students meet a biologist studying the pattern of healthy and diseased trees in the Palatinate forest. The research concentrates on the root rot in the trees and especially the lengths of the rotten root parts (in meter) of $n = 110$ diseased trees:

Verrottete Länge:	Anzahl Beobachtungen z_i :	Rotten root length:	No of observations z_i :
1	73	1	73
2	27	2	27
3	5	3	5
4	2	4	2
5	2	5	2
6	1	6	1

Der Biologe schlägt eine geometrische Verteilung vor:

The biologist suggests a geometric distribution:

$$p_i(\vartheta) = \mathcal{P}(X = i) = \vartheta \cdot (1 - \vartheta)^{i-1}, \quad i = 1, 2, \dots$$

wobei X die Länge der verrotteten Wurzel ist. Mittels eines Goodness-of-Fit-Tests wollen die StudentInnen untersuchen, ob die Annahme eines geometrischen Modell vernünftig ist.

where X is the length of the rotten roots. The students want to investigate with the help of a goodness-of-fit test, if the geometric model is reasonable.

Zuerst müssen sie den Maximum-Likelihood-Schätzer von ϑ bestimmen.

First, they must determine the maximum likelihood estimator of ϑ .

Die Log-Likelihood-Funktion ist

The log-likelihood function is

$$l(\vartheta|\mathbf{x}) = \sum_{j=1}^n (x_j - 1) \cdot \log(1 - \vartheta) + n \cdot \log(\vartheta)$$

und durch Ableiten erhalten sie

and by taking the derivative they obtain

$$\frac{\partial l}{\partial \vartheta}(\vartheta|\mathbf{x}) = -\frac{\sum_{j=1}^n x_j - n}{1 - \vartheta} + \frac{n}{\vartheta}$$

was zum Maximum-Likelihood-Schätzwert $\hat{\vartheta} = \frac{1}{\bar{x}} = 0,66$ führt.

which gives the maximum likelihood estimate $\hat{\vartheta} = \frac{1}{\bar{x}} = 0.66$.

Die χ^2 -Statistik hat in diesem Fall den Wert

The χ^2 -statistic has in this case the value

$$\chi^2 = \sum_{i=1}^6 \frac{[Z_i - n \cdot p_i(\hat{\vartheta})]^2}{n \cdot p_i(\hat{\vartheta})} = 4,40.$$

Da es sechs Gruppen sind und ein Parameter geschätzt wurde, wird laut Satz 4.4.8 eine χ^2 -Verteilung mit $6 - 1 - 1 = 4$ Freiheitsgraden benutzt.

Since there are six groups and one parameter was estimated, a χ^2 -distribution with $6 - 1 - 1 = 4$ degrees of freedom must be used according to Theorem 4.4.8.

Das 95%-Quantil dieser Verteilung ist 9,49, d.h. die Hypothese der geometrischen Verteilung kann nicht abgelehnt werden.

The 95% quantile of this distribution is 9.49, i.e. the hypothesis of a geometric distribution cannot be rejected.

4.4.3 Unabhängigkeitstest Test of independence

Der χ^2 -Test kann auch zum Untersuchen der Unabhängigkeit benutzt werden.

The χ^2 -test can be used to test for independence, too.

Modell:

Seien $(X_1, Y_1), \dots, (X_n, Y_n)$ u.i.v. mit Werten in $\{1, \dots, m_x\} \times \{1, \dots, m_y\}$ und die Verteilung spezifiziert durch

Model:

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. with values in $\{1, \dots, m_x\} \times \{1, \dots, m_y\}$ and distribution specified by

$$\mathcal{P}(X_j = k, Y_j = l) = p_{kl}, \quad k = 1, \dots, m_x, l = 1, \dots, m_y.$$

Seien $p_1^{(x)}, \dots, p_{m_x}^{(x)}$ und $p_1^{(y)}, \dots, p_{m_y}^{(y)}$ die Marginalverteilungen von X_k und Y_l :

Let $p_1^{(x)}, \dots, p_{m_x}^{(x)}$ and $p_1^{(y)}, \dots, p_{m_y}^{(y)}$ be the marginal distributions of X_i and Y_j :

$$p_k^{(x)} = \mathcal{P}(X_i = k) = \sum_{l=1}^{m_y} p_{kl}, \quad k = 1, \dots, m_x,$$

$$p_l^{(y)} = \mathcal{P}(Y_j = l) = \sum_{k=1}^{m_x} p_{kl}, \quad l = 1, \dots, m_y.$$

Problem:

Problem:

$H_0 : X_j, Y_j$ unabhängig,
 $H_1 : X_j, Y_j$ abhängig.

$H_0 : X_j, Y_j$ independent,
 $H_1 : X_j, Y_j$ dependent.

Die Hypothese ist äquivalent zu

The hypothesis is then equivalent to

$$H_0 : \forall(k, l) \quad p_{kl} = p_k^{(x)} \cdot p_l^{(y)},$$

$$H_1 : \exists(k, l) \quad p_{kl} \neq p_k^{(x)} \cdot p_l^{(y)}.$$

Der Parameterraum

The parameter space

$$\Theta = \{(p_{11}, p_{12}, \dots, p_{m_x m_y}) \mid \sum_{k=1}^{m_x} \sum_{l=1}^{m_y} p_{kl} = 1; p_{kl} \geq 0, k = 1, \dots, m_x, l = 1, \dots, m_y\}$$

hat Dimension $m_x \cdot m_y - 1$.

has dimension $m_x \cdot m_y - 1$.

Der Hypothesenraum

The hypothesis space

$$\Theta_0 = \left\{ (p_{11}, p_{12}, \dots, p_{m_x m_y}) \mid p_{kl} = p_k^{(x)} \cdot p_l^{(y)}; \sum_{k=1}^{m_x} p_k^{(x)} = 1; \sum_{l=1}^{m_y} p_l^{(y)} = 1; p_1^{(x)}, \dots, p_{m_x}^{(x)}, p_1^{(y)}, \dots, p_{m_y}^{(y)} \geq 0 \right\}$$

für die Unabhängigkeitsannahme hat Dimension
 $q = (m - 1) + (n - 1)$.

for the independence assumption has dimension
 $q = (m - 1) + (n - 1)$.

Teststatistik:

Sei Z_{kl} die Anzahl der Paare (X_j, Y_j) mit
 $X_j = k, Y_j = l$.

Wir benutzen die Notation

Test statistic:

Let Z_{kl} be the number of pairs (X_j, Y_j) with
 $X_j = k, Y_j = l$.

We use the notation

$$Z_k = \sum_{l=1}^{m_y} Z_{kl},$$

$$Z_l = \sum_{k=1}^{m_x} Z_{kl}.$$

Der Maximum-Likelihood-Schätzer unter der Hypothese ist

The maximum likelihood estimator under the hypothesis is

$$\hat{p}_k^{(x)} = \frac{1}{n} \cdot Z_k,$$

$$\hat{p}_l^{(y)} = \frac{1}{n} \cdot Z_l$$

und deswegen $\hat{p}_{kl} = \hat{p}_k^{(x)} \cdot \hat{p}_l^{(y)}$.

and therefore $\hat{p}_{kl} = \hat{p}_k^{(x)} \cdot \hat{p}_l^{(y)}$.

Die χ^2 -Statistik von Satz 4.4.8 hat die Form

The χ^2 -statistic from Theorem 4.4.8 has the form

$$\chi^2 = \sum_{k=1}^{m_x} \sum_{l=1}^{m_y} \frac{(Z_{kl} - n \cdot \hat{p}_k^{(x)} \cdot \hat{p}_l^{(y)})^2}{n \cdot \hat{p}_k^{(x)} \cdot \hat{p}_l^{(y)}} = \sum_{k=1}^{m_x} \sum_{l=1}^{m_y} \frac{(Z_{kl} - \frac{Z_k \cdot Z_l}{n})^2}{\frac{Z_k \cdot Z_l}{n}}$$

und ist unter der Hypothese asymptotisch

and is under the hypothesis asymptotically

χ^2 -verteilt mit

χ^2 -distributed with

$$m_x \cdot m_y - (m_x - 1) - (m_y - 1) - 1 =$$

$$(m_x - 1) \cdot (m_y - 1) \text{ Freiheitsgraden.}$$

$$m_x \cdot m_y - (m_x - 1) - (m_y - 1) - 1 =$$

$$(m_x - 1) \cdot (m_y - 1) \text{ degrees of freedom.}$$

Annahmebereich:

Acceptance region:

$$C_0 = \{\mathbf{Z} \mid \chi^2(\mathbf{Z}) < c_\alpha\},$$

wobei c_α das $(1 - \alpha)$ -Quantil der
 $\chi^2_{(m_x - 1) \cdot (m_y - 1)}$ -Verteilung ist.

where c_α is the $(1 - \alpha)$ -quantile of the
 $\chi^2_{(m_x - 1) \cdot (m_y - 1)}$ -distribution.

Beispiel 4.4.12

Zurück bei der Universität, sind die verbliebenen StudentInnen der Wandergruppe ein bißchen enttäuscht, weil sie keine richtige Elwedritsche im Wald gesehen haben.

Example 4.4.12

Back at the university again, the students still remaining from the hiking group are of course a bit disappointed, since they never saw a real elwedritsche in the forest.

Allerdings lesen sie danach die neueste Auswertung einer Umfrage, die die Beziehung zwischen Freizeitaktivitäten und Studienergebnisse der Mathematics International behandelt. Sofort sind sie ein bißchen glücklicher:

However, afterwards they read the new evaluation of a survey concerning the connection between leisure activities and study results of the mathematics international students. At once, they become a bit happier.

In der Umfrage wurden 100 Mathematics International StudentInnen über ihre Freizeitaktivitäten befragt.

In the survey, 100 mathematics international students were asked about their leisure activities.

Sie wurden wie folgt in drei Gruppen aufgeteilt: Gruppe A besteht aus StudentInnen, die sich völlig auf Mathematik konzentrieren, d.h. ohne Freizeitaktivitäten.

They were divided into three groups as follows: Group A consists of students, who are totally concentrating on mathematics, i.e. without leisure activities.

Die StudentInnen in Gruppe B machen einmal oder zweimal pro Woche etwas anderes als studieren.

The students in Group B do once or twice a week something else than studying.

Schließlich die StudentInnen in Gruppe C, die natürlich in Deutschland sind, um Mathematik zu studieren, aber auch um den deutschen Lebensstil zu erleben.

Finally, the students in Group C are those, who are of course in Germany to study mathematics, but also experience the German way of living.

Ferner wurden dieselben StudentInnen über ihre Durchschnittsnoten \bar{g} gefragt.

Furthermore, the same students were asked about their average grades, called \bar{g} .

Das Ergebnis ist in der folgenden Tabelle vermerkt:

The result can be seen in the following table:

	$\bar{g} > 3$	$2 \leq \bar{g} \leq 3$	$\bar{g} < 2$
Gr. A:	22	6	2
Gr. B:	11	12	7
Gr. C:	17	12	11

Bei Benutzung eines χ^2 -Unabhängigkeitstests erhält man 10,74 für die Teststatistik. Das 95%-Quantil der χ^2_4 -Verteilung ist 9,49.

Somit kann die Hypothese, dass das Studienergebnis und Freizeitaktivitäten unabhängig sind, abgelehnt werden!

Using a χ^2 -independence test, the test statistic is 10.74. The 95% quantile of the χ^2_4 -distribution is 9.49. So, the hypothesis that study result and leisure activities are independent can be rejected!

5 Literatur

Literature

Breiman, Leo. *Statistics: with a view toward applications*. Boston, 1973.

Bickel, Peter J. and Doksum, Kjell A. *Mathematical statistics: basic ideas and selected topics*. San Francisco, 1977.

Schlittgen, R. *Statistische Inferenz*. München/Wien, 1996.

Serfling, Robert J. *Approximation Theorems of Mathematical Statistics*. New York, 1980.