Wahrscheinlichkeit und Statistik

Skript zur Vorlesung:
Elementare Wahrscheinlichkeitstheorie und Statistik, WS 2006.

Prof. Dr. Michael Kohler
Fachrichtung 6.1 - Mathematik
Universität des Saarlandes
Postfach 151150
D–66041 Saarbrücken

kohler@math.uni-sb.de
http://www.uni-sb.de/ag-statistik/

“Those who ignore Statistics are condemned to reinvent it.”

Brad Efron

“Was war das für eine Stimme?” schrie Arthur.

“Ich weiß es nicht”, brüllte Ford zurück, “ich weiß es nicht. Es klang wie eine Wahrscheinlichkeitsrechnung.”

“Wahrscheinlichkeit? Was willst du damit sagen?”

“Eben Wahrscheinlichkeit. Verstehst du, so was wie zwei zu eins, drei zu eins, fünf zu vier. Sie sagte, zwei hoch einhunderttausend zu eins. Das ist ziemlich unwahrscheinlich, verstehst du?”

Ein Fünf-Millionen-Liter-Bottich Vanillesoße ergoß sich ohne Warnung über sie.

“Aber was soll das denn?” rief Arthur.

“Was, die Vanillesoße?”

“Nein, die Wahrscheinlichkeitsrechnung!”

Douglas Adams

Inhaltsverzeichnis

Kapitel 1

Motivation

Im vorliegenden Buch wird eine Einführung in die Wahrscheinlichkeitstheorie und die Statistik gegeben. Eine naheliegende Frage, bevor man sich mit einem neuen – und wie im vorliegenden Fall nicht völlig trivialen – Stoffgebiet befasst, ist, ob man das dabei (unter Umständen mühsam) erlernte Wissen jemals wirklich brauchen wird.

Diese Frage ist im Falle der Statistik (deren gründliches Verständnis Kenntnisse in Wahrscheinlichkeitstheorie voraussetzt) ganz klar mit Ja zu beantworten, da Statistikwissen bei vielen Aussagen im täglichen Leben benötigt wird. Dies soll im Folgenden mit Hilfe einiger weniger der vielen Anwendungsbeispiele von Statistikwissen illustriert werden.

1.1 Statistik-Prüfung, Herbst 2001

Im Sommersemester 2001 wurde an der Universität Stuttgart die Vorlesung *Statistik für Ingenieure* abgehalten. Diese gehörte zum Pflichtprogramm für das Vordiplom im Studienfach Elektrotechnik und wurde am 27.09.2001 im Rahmen einer zweistündigen Klausur abgeprüft. Nach Korrektur der 59 abgegebenen Klausuren stellte sich die Frage, wie denn nun die Prüfung ausgefallen ist. Dazu kann man natürlich die Noten aller 59 Klausuren einzeln betrachten, verliert aber dabei schnell den Überblick.

Hilfreich ist hier die *deskriptive (oder beschreibende) Statistik*, die Verfahren bereitstellt, mit denen man - natürlich nur unter Verlust von Information - die 59 Einzelnoten in wenige Zahlen zusammenfassen kann, wie z.B.

Notendurchschnitt	:	1,9
Durchfallquote	:	3,4 %

Dies kann man auch für Teilmengen der abgegebenen Klausuren tun. Betrachtet man z.B. die Menge aller Teilnehmer, die (den übrigens freiwillig zu erwerbenden) Übungsschein zur Vorlesung erworben haben, so erhält man:

Anzahl Teilnehmer mit Übungsschein	:	46
Notendurchschnitt	:	1,7
Durchfallquote	:	0 %

Dagegen erhält man für die Teilnehmer, die diesen Schein nicht erworben haben:

Anzahl Teilnehmer ohne Übungsschein	:	13
Notendurchschnitt	:	2,7
Durchfallquote	:	15,4 %

Hierbei fällt auf, dass sowohl der Notendurchschnitt als auch die Durchfallquote bei der ersten Gruppe von Studenten deutlich günstiger ausfällt als bei der zweiten Gruppe. Dies führt auf die Vermutung, dass auch bei zukünftigen Studenten der Vorlesung Statistik für Ingenieure der Erwerb des Übungsscheines sich günstig auf das Bestehen und die Note der Prüfung auswirken wird.

Die Fragestellung, ob man aus den oben beschriebenen Daten eine solche Schlussfolgerung ziehen kann, gehört zur *induktiven (oder schließenden) Statistik*.

Problematisch an dieser Schlussweise ist vor allem der Schluss von der beobachteten Gleichzeitigkeit (d.h., vom gleichzeitigen Auftreten des Erwerb des Übungsscheines und des guten Abschneidens bei der Prüfung) auf die Kausalität (d.h., auf die Behauptung, dass Studenten deshalb bessere Noten haben, weil sie den Übungsschein erworben haben). Ein bekanntes Beispiel für diese im täglichen Leben häufig auftretende Schlussweise wird im nächsten Abschnitt vorgestellt.

1.2 Sex und Herzinfarkt

In einer Studie an der Universität Bristol wurde versucht, Risikofaktoren für das Auftreten eines Herzinfarktes zu bestimmen. Dazu wurden 2400 gesunde Männer unter anderem zu ihrem Sexualleben befragt und über einen Zeitraum von 10 Jahren beobachtet.

Ein Resultat dieser Studie war, dass in der Gruppe der Männer, die angegeben hatten, mindestens 3 bis 4 Orgasmen die Woche zu haben, prozentual nur halb so häufig ein Herzinfarkt aufgetreten ist wie beim Rest.

Die gängige Interpretation dieses Ergebnisses in Tageszeitungen (die darüber in der Vergangenheit ausführlich berichtet haben) ist, dass man durch Änderung seines Sexualverhaltens das Risiko, einen Herzinfarkt zu erleiden, beeinflussen kann. Beschäftigt man sich aber etwas näher mit der Interpretation von Studien (z.B. durch Lesen von Kapitel 2 dieses Buches), so sieht man leicht, dass die hier vorgenommene Schlussweise von der beobachteten Gleichzeitigkeit auf die behauptete Kausalität nicht zulässig ist.

1.3 Die Challenger-Katastrophe

Am 28. Jannuar 1986 explodierte die Raumfähre Challenger genau 73 Sekunden nach ihrem Start. Dabei starben alle 7 Astronauten. Auslöser dieser Katastrophe war, dass zwei Dichtungsringe an einer der beiden Raketentriebwerke der Raumfähre aufgrund der sehr geringen Außentemperatur beim Start ihre Elastizität verloren hatten und undicht geworden waren.

Einen Tag vor dem Start hatten Experten von Morton Thiokol, dem Hersteller der Triebwerke, angesichts der geringen vorhergesagten Außentemperatur beim Start von unter 0 Grad Celsius Bedenken hinsichtlich der Dichtungsringe und empfahlen, den Start zu verschieben. Als Begründung dienten in der Vergangenheit beobachtete Materialermüdungen an den Dichtungsringen (unter anderem gemessen durch das Vorhandensein von Ruß hinter den Dichtungen). Eine wichtige Rolle in der Argumentation spielten die in Tabelle 1.1 dargestellten Daten, die sich auf die Flüge beziehen, bei denen eine nachträgliche Untersuchung Materialermüdungen an einem der sechs Dichtungsringe ergeben hatten.

Flugnummer	Datum	Temperatur (in Grad Celsius)
STS-2	12.11.81	21,1
41-B	03.02.84	13,9
41-C	06.04.84	17,2
41-D	30.08.84	21,1
51-C	24.01.85	11,7
61-A	30.10.85	23,9
61-C	12.01.86	14,4

Tabelle 1.1: Flüge mit Materialermüdung an den Dichtungsringen.

Der Zusammenhang zwischen dem Auftreten von Schädigungen und der Außentemperatur war für die Experten von der NASA leider nicht nachvollziehbar.

Insbesondere wurde argumentiert, dass ja auch bei hohen Außentemperaturen Schädigungen aufgetreten waren. Daher wurde der Start nicht verschoben.

Bemerkenswert ist daran, dass der wahre Grund für die spätere Katastrophe bereits vor dem Unfall bekannt war und ausgiebig diskutiert wurde. Unglücklicherweise waren die Techniker von Morton nicht in der Lage, ihre Bedenken genau zu begründen. Neben einer Vielzahl von Fehlern bei der graphischen Darstellung der in der Vergangenheit beobachteten Messdaten hatten diese erstens vergessen, auch die Flüge ohne Schädigungen am Dichtungsring zusammen mit ihrer Außentemperatur mit darzustellen. Dies hätte das obige Argument der Schädigungen bei hohen Außentemperaturen relativiert, indem es gezeigt hätte, dass zwar einerseits bei einigen Starts bei hohen Außentemperaturen Schädigungen auftraten, aber andererseits bei allen Starts bei niedrigen Außentemperaturen Schädigungen auftraten (vgl. Abbildung 1.1).

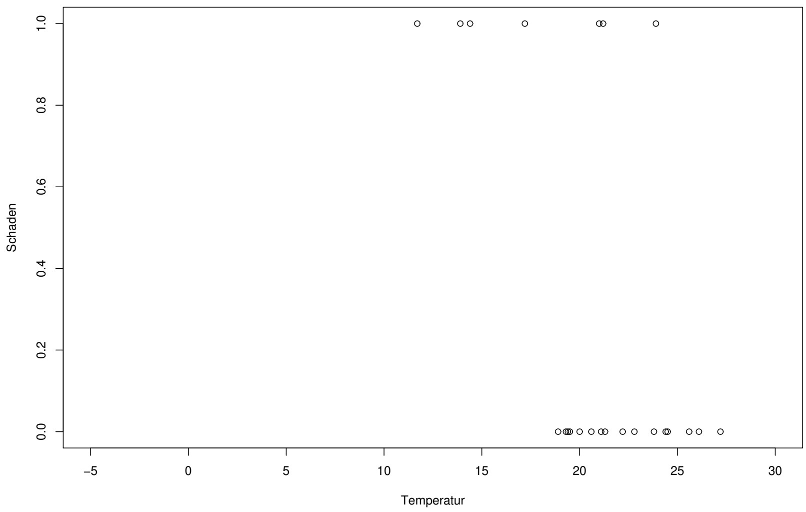

Abbildung 1.1: Auftreten von Schäden bei früheren Flügen.

Zweitens war das Auftreten von Materialermüdung nicht das richtige Kriterium zur Beurteilung der Schwere des Problems. Hätte man z.B. die aufgetretenen Abnutzungen der Dichtungsringe zusammen mit dem Auftreten von Ruß in einem Schadensindex zusammengefasst und diesen in Abhängigkeit der Temperatur dargestellt, so hätte man die Abbildung 1.2 erhalten.

Diese hätte klar gegen einen Start bei der vorhergesagten Außentemperatur von unter 0 Grad Celsius gesprochen.

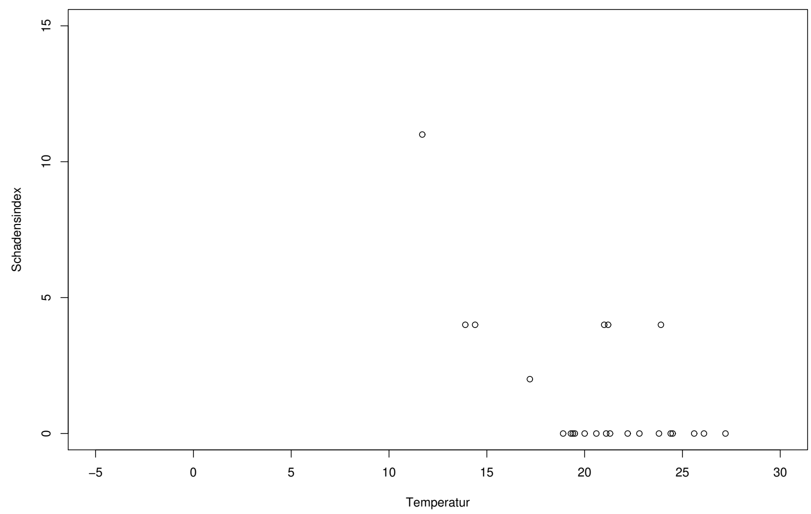

Abbildung 1.2: Schadensindex in Abhängigkeit von der Temperatur.

1.4 Genetischer Fingerabdruck

Beim genetischen Fingerabdruck handelt es sich um ein in der jüngeren Vergangenheit oft sehr erfolgreich angewandtes Hilfsmittel zur Aufklärung von Kapitalverbrechen. Dabei wird am Tatort gefundenes DNA Material (das z.B. aus Hautpartikeln des Täters stammt) mit dem eines Verdächtigen verglichen. Da die DNA für jeden Menschen eindeutig ist, weiß man, dass bei völliger Übereinstimmung des DNA Materials der Verdächtige der Täter sein muss, und dass bei Nichtübereinstimmung der Verdächtige nicht der Täter sein kann.

Leider ist es unmöglich, festzustellen, ob das DNA Material völlig übereinstimmt. Dies liegt daran, dass (vereinfacht gesprochen) die DNA eine lange Kette aus mehr als 1.000.000 Mononukleotiden ist. Jedes dieser Mononukleotide nimmt eine von vier möglichen Formen an, so dass die DNA selbst mehr als $4^{1.000.000}$ mögliche Formen annehmen kann. Um eine völlige Übereinstimmung der DNA feststellen zu können, müsste man alle, d.h. mehr als 1.000.000, Mononukleotide vergleichen. Dies ist leider zu aufwendig.

Statt dessen vergleicht man nur eine (kurze) Sequenz von Mustern in der DNA und damit nur einen Teil der Mononukleotidkette. Ergibt dieser Vergleich keine Übereinstimmung, so weiß man sicher, dass der Verdächtige nicht der Täter ist. Schwieriger ist aber die Schlussweise bei Vorliegen einer Übereinstimmung.

Um auch in diesem Fall zu einer Aussage zu kommen, verwendet man ein stochastisches Modell: Man schätzt, wie häufig eine Übereinstimmung auftritt, wenn man Menschen zufällig auswählt und ihre DNA mit dem am Tatort gefundenen Material vergleicht. Falls dabei eine Übereinstimmung nur sehr selten auftritt, so schließt man, dass der Verdächtige der Täter ist.

Problematisch bei diesem Vorgehen ist die Schätzung der Häufigkeit einer Übereinstimmung. Die Häufigkeit des Auftretens bestimmter genetischer Muster variiert stark zwischen verschiedenen rassischen und ethnischen Gruppen von Menschen. Insofern hängt obige Schätzung auch stark davon ab, ob man die Auswahl von Menschen aus der gesamten Menschheit, aus einer Großfamilie oder aus einem abgeschiedenen Dorf betrachtet.

1.5 Präsidentschaftswahl in den USA, Herbst 2000

In den USA wird der Präsident indirekt gewählt: Pro Bundesstaat werden die gültigen abgegebenen Stimmen pro Kandidat ermittelt. Wer die meisten Stimmen erhält, bekommt die Wahlmänner bzw. -frauen zugesprochen, die für diesen Bundesstaat zu vergeben sind. Diese wählen dann den Präsidenten.

Bei der Präsidentschaftswahl im Herbst 2000 trat der Fall auf, dass George Bush - einer der beiden aussichtsreichen Kandidaten - die 25 Wahlmänner bzw. -frauen des Bundesstaates Florida (und damit die Mehrheit der Wahlmänner bzw. -frauen) mit einem Vorsprung von nur 537 Stimmen gewann. Al Gore - der unterlegene andere aussichtsreiche Kandidat - versuchte danach in einer Reihe von Prozessen, die Auszählung der Stimmen in Florida (und damit die Präsidentschaftswahl) doch noch zu seinen Gunsten zu entscheiden.

Die Abgabe der Stimmen erfolgte in Florida größtenteils durch Lochung von Lochkarten, die anschließend maschinell ausgezählt wurden. Es ist bekannt, dass bei diesem Verfahren mit ca. $1,5\%$ der Stimmen deutlich mehr versehentlich ungültig abgegebene (da z.B. unvollständig gelochte) Stimmen auftreten als bei optoelektronischen Verfahren (hier treten ca. $0,5\%$ versehentlich ungültige Stimmen auf). Zentraler Streitpunkt bei den Prozessen war, ob man z.B. im Wahlbezirk Tallahasse, wo allein 10.000 ungültige Stimmen abgegeben wurden, diese manuell nachzählen sollte.

Im Prozess vor dem Supreme Court in Florida hat Statistik Professor Nicholas Hengartner aus Yale für Al Gore ausgesagt. Dessen zentrales Argument war, dass eine unabsichtliche unvollständige Lochung bei Kandidaten, die wie Al Gore auf der linken Seite der Lochkarte stehen, besonders häufig auftritt. Zur Begründung wurde auf die Senats- und Gouverneurswahl in Florida im Jahre 1998 verwiesen. Dabei waren bei einer der beiden Wahlen deutlich mehr ungültige Stimmen aufgetreten als bei der anderen. Diese Argumentation war aber nicht haltbar, da - wie die Anwälte von George Bush durch Präsentation eines Stimmzettels der damaligen Wahl überzeugend begründeten - damals die Kandidaten für beide Wahlen auf der gleichen Seite des Stimmzettels standen.

Dennoch hätte eine vollständige manuelle Nachzählung der Stimmen in Florida unter Umständen das Ergebnis der Wahl verändert: Lochkarten wurden vor allem in ärmeren Wahlbezirken eingesetzt, während in reicheren Gegenden (teurere und genauere) optoelektronische Verfahren verwendet wurden. Da der Anteil der Stimmen für Al Gore in den ärmeren Gegenden besonders hoch war, steht zu vermuten, dass unter den versehentlich für ungültig erklärten Stimmen mehr für Al Gore als für George Bush waren. Um dies aber sicher festzustellen, hätten man nicht nur in einem, sondern in allen Wahlbezirken Floridas manuell nachzählen müssen, was zeitlich nicht möglich war.

1.6 Personalisierung von Internetseiten

Beim Versuch des Einkaufens von Waren im Internet steht der potentielle Käufer häufig vor dem Problem, dass es gar nicht so einfach ist, das gewünschte Produkt zu finden. Dies könnte deutlich einfacher (und damit für den Betreiber der Seite lukrativer) gemacht werden, wenn sich die Internetseite automatisch dem Wunsch des Besuchers anpassen würde, d.h. wenn der jeweilige Nutzer individuell auf seine Wünsche zugeschnittene Seiten präsentiert bekäme.

Um die Wünsche des Besuchers vorherzusagen steht zum einen das bisher beobachtete Navigationsverhalten des aktuellen Besuchers, sowie zum anderen das in der Vergangenheit beobachtete Navigationsverhalten anderer Besucher (inklusive Kaufentscheidung) zur Verfügung. Nach Bestimmung des Wunsches eines Besuchers kann eine personalisierte Internetseite dann z.B. durch Einblendung von spezieller Werbung auf diesen zugeschnitten werden.

Als Beispiel für eine personalisierte Internetseite sei auf

www.k1010.de

verwiesen. Dort wird ein Quizspiel mit Gewinnmöglichkeiten angeboten. Um die Besucher dabei möglichst lange auf der Seite (und damit bei der auf dieser Seite eingeblendeten Werbung) festzuhalten, wird hier der Schwierigkeitsgrad der

Fragen den bisherigen Antworten des Besuchers angepasst.

Kapitel 2

Erhebung von Daten

Die Statistik beschäftigt sich mit der Analyse von Daten, in denen gewisse zufällige Strukturen vorhanden sind. Manchmal kann der Statistiker auf die Erhebung dieser Daten, z.B. in Form von Studien oder Umfragen, Einfluss nehmen. Was dabei zu beachten ist, wird in diesem Kapitel erläutert. Die Kenntnis dieser Sachverhalte ist insofern wichtig, da sie hilfreich bei der Beurteilung der Aussagekraft von Ergebnissen von Studien und Umfragen ist.

2.1 Kontrollierte Studien

Kontrollierte Studien werden im Folgenden anhand des Vorgehens bei der Überprüfung der Wirksamkeit der Anti-Grippe-Pille Tamiflu eingeführt.

Grippe (oder Influenza) ist eine durch Tröpfcheninfektion übertragene Infektionskrankheit, die durch Viren ausgelöst wird. Allein in den USA, Japan und Westeuropa erkranken jedes Jahr rund 100 Millionen Menschen an Grippe, in den USA sterben jährlich ca. 20.000 meist ältere Menschen an den Folgen einer Grippeerkrankung. In Abständen von (mehreren) Jahrzehnten bricht eine besonders tückische Grippeepidemie aus, z.B. 1968-69 die sogenannte Hongkong-Grippe, 1957-58 die sogenannte asiatische Grippe oder 1918-20 die sogenannte spanische Grippe. An Letzterer starben weltweit 22 Millionen Menschen.

An Grippe erkranken Menschen aller Alterstufen. Die Grippe-Viren greifen die Schleimhäute im Atembereich (Nase bis Bronchien) an, was die Gefahr von Sekundärinfektionen (insbesondere Lungenentzündung, Ursache von mehr als 80% der Grippetodesfällen) birgt. Typisch an Grippe ist der plötzliche Beginn mit hohem Fieber, Halsweh, Schnupfen und Gliederschmerzen. Bei unkompliziertem Verlauf ist die Erkrankung nach ca. einer Woche vorüber, unter Umständen ist man aber noch längere Zeit danach geschwächt.

Wirksamster Schutz vor einer Grippeinfektion ist eine Impfung. Da sich der Erreger ständig verändert, muss diese jährlich wiederholt werden. Nach Ausbruch der Erkrankung werden heutzutage meist nur die Symptome oder eventuell auftretende Begleitinfektionen bekämpft, nicht aber das Virus selbst. Zur Bekämpfung des Virus gab es bis Mitte der 90er Jahre nur zwei Präparate, die beide starke Nebenwirkungen hatten und nur bei speziellen Grippeviren wirksam waren.

Wie alle Viren vervielfältigt sich das Grippevirus, indem es in Körperzellen eindringt und diese veranlasst, neue Viren herzustellen. Beim Verlassen der Wirtszelle zerstören diese die Zelle und befallen dann weitere Körperzellen. Um ein Klebenbleiben an der Wirtszelle zu vermeiden, muss vorher die auf deren Oberfläche befindliche Salinsäure aufgelöst werden. Dies macht das Enzym Neuraminidase, das auf der Oberfläche des Grippevirus sitzt.

Australische Wissenschaftler entschlüsselten 1983 den komplexen räumlichen Aufbau des Neuraminidase-Moleküls. Wie auch die Oberfläche des Grippevirus verändert sich auch dessen Oberfläche von Jahr zu Jahr stark. Entdeckt wurde aber eine Stelle, die immer gleich bleibt: eine tiefe Spalte, in der die Salinsäure aufgelöst wurde. Die Idee bei der Entwicklung einer neuen Behandlungsmethode für Grippe war nun, ein Molekül zu finden, das diese Spalte verstopft und damit die Auflösung der Salinsäure verhindert. Gleichzeitig musste es vom Körper einfach aufgenommen werden können, ungiftig sein, und es durfte nur die Neuraminidase der Grippeviren, nicht aber andere Enzyme, blockieren.

Potenzielle Stoffe wurden zuerst im Reagenzglas getestet. Dabei wurde festgestellt, ob sie wirklich die Neuraminidase blockieren und ob sie in Gewebekulturen die Vermehrung von Grippeviren verhindern. Anschließend wurde die Wirksamkeit an Mäusen und Iltisen getestet. Nach dreijähriger Arbeit hatte man Anfang 1996 einen Stoff gefunden, der das Grippevirus in Mäusen und Iltisen erfolgreich bekämpfte.

Zur Zulassung als Medikament musste die Wirksamkeit am Menschen nachgewiesen werden. Dabei ist ein Vorgehen in drei Phasen üblich: In Phase I wird an einer kleinen Gruppe gesunder Menschen getestet, ob es unerwartete Nebenwirkungen gibt und was die beste Dosierung ist. In Phase II wird die Wirksamkeit des Medikaments an einer kleinen Gruppe Grippekranker überprüft. Abschließend erfolgt in Phase III ein Test unter realistischen Bedingungen an Hunderten von Menschen.

Die Überprüfung der Wirksamkeit eines Medikaments in den Phasen II und III erfolgt im Rahmen einer *Studie*. Die Grundidee dabei ist der *Vergleich:* Man vergleicht eine sogenannte *Studiengruppe*, die mit dem Medikament behandelt wurde, mit einer sogenannten *Kontrollgruppe*, die nicht mit dem Medikament behandelt wurde. Um dabei von Unterschieden im Verhalten der Studien- und der Kontrollgruppe (z.B. hinsichtlich der Dauer der Erkrankung) auf die Wirksamkeit des Medikaments schließen zu können, muss dabei (abgesehen von der Behandlung mit dem Medikament) die Kontrollgruppe möglichst ähnlich zur Studiengruppe sein.

Für die Wahl von Studien- und Kontrollgruppe gibt es verschiedene Möglichkeiten. Bei einer *retrospektiv kontrollierten Studie* wird die Studiengruppe mit in der Vergangenheit gesammelten Daten verglichen.

Im obigen Beispiel bedeutet dies, dass man als Studiengruppe eine größere Anzahl von Personen auswählt, die gerade an Grippe erkrankt sind, und diese alle (bzw. nur diejenigen, die mit der Behandlung einverstanden sind) mit dem neuen Medikament behandelt. Dann wartet man einige Zeit ab und bestimmt die durchschnittliche Krankheitsdauer bei den behandelten Patienten. Diese vergleicht man mit der durchschnittlichen Krankheitsdauer von in der Vergangenheit an Grippe erkrankten Personen. Aufgrund der Betrachtung der durchschnittlichen Krankheitsdauer kann man dabei eventuelle Unterschiede bei den Gruppengrößen vernachlässigen.

Problematisch an diesem Vorgehen ist, dass sich das Grippevirus jedes Jahr stark verändert und immer wieder neue Varianten des Virus für Erkrankungen verantwortlich sind. Stellt man also fest, dass die durchschnittliche Krankheitsdauer bei den mit dem neuen Medikament behandelten Personen geringer ist als bei den in der Vergangenheit traditionell behandelten Personen, so weiß man nicht, ob das an dem neuen Medikament liegt, oder ob der Grund dafür ist, dass das Grippevirus in diesem Jahr vergleichsweise harmlos ist.

Im Gegensatz zu retrospektiv kontrollierten Studien stammen bei *prospektiv kontrollierten Studien* Studiengruppe und Kontrollgruppe beidesmal aus der Gegenwart. Je nachdem, ob man die Testpersonen dabei *deterministisch* oder *mittels eines Zufallsexperiments* in Studien- und Kontrollgruppe unterteilt, spricht man von *prospektiv kontrollierten Studien ohne* oder *mit Randomisierung.*

Im vorliegenden Beispiel könnte man eine prospektiv kontrollierte Studie ohne Randomisierung so durchführen, dass man zuerst eine größere Anzahl von an Grippe erkrankten Personen auswählt, und dann alle diejenigen, die der Behandlung zustimmen, mit dem neuen Medikament behandelt. Diese Personen würden die Studiengruppe bilden, der Rest der ausgewählten Personen wäre die Kontrollgruppe. Nach einiger Zeit würde man die durchschnittliche Krankheitsdauer in beiden Gruppen vergleichen.

Bei diesem Vorgehen entscheiden die Erkrankten, ob sie zur Studiengruppe oder zur Kontrollgruppe gehören. Das führt dazu, dass sich die Kontrollgruppe nicht nur durch die Behandlung von der Studiengruppe unterscheidet. Zum Beispiel ist es denkbar, dass besonders viele ältere Menschen der Behandlung zustimmen. Bei diesen führt Grippe besonders häufig zu Komplikationen (wie z.B. Lungenentzündung), so dass für diese eine möglicherweise verbesserte Behandlungsmethode besonders attraktiv ist. Darüberhinaus wird bei diesen Personen die Grippe auch im Durchschnitt länger dauern als bei jungen Menschen. Daher tritt das Problem auf, dass hier der Einfluss der Behandlung *konfundiert* (sich vermengt) mit dem Einfluss des Alters. Insofern kann man nicht sagen, inwieweit ein möglicher Unterschied bei den durchschnittlichen Krankheitsdauern auf die Behandlung zurückzuführen ist (bzw. ein eventuell nicht vorhandener Unterschied nur aufgrund der Unterschiede beim Alter auftritt).

Als möglicher Ausweg bietet sich an, als Kontrollgruppe nur einen Teil der Erkrankten auszuwählen, die der Behandlung mit dem neuen Medikament nicht zustimmen, und diesen Teil so zu bestimmen, dass er z.B. hinsichtlich des Alters möglichst ähnlich zur Studiengruppe ist. Dies ist aber sehr fehleranfällig, da man dazu sämtliche Faktoren kennen muss, die Einfluss auf die Krankheitsdauer haben. Da Grippe weltweit in Epidemien auftritt, wäre ein weiterer solcher Faktor z.B. der Wohnort der Erkrankten.

Dieses Problem wird bei einer *prospektiv kontrollierten Studie mit Randomisierung* vermieden. Denn dabei werden nur solche Testpersonen betrachtet, die sowohl für die Studien- als auch für die Kontrollgruppe in Frage kommen. Diese werden dann zufällig (z.B. durch Münzwurf) in Studien- und Kontrollgruppe unterteilt.

Im Falle des obigen Beispiels heißt das, dass nur die Erkrankten betrachtet werden, die der Behandlung zustimmen. Diese werden zufällig (z.B. durch Münzwürfe) in Studien- und Kontrollgruppe aufgeteilt. Anschließend werden die Personen in der Studiengruppe mit dem neuen Medikament behandelt, die in der Kontrollgruppe traditionell behandelt und nach einiger Zeit werden die durchschnittlichen Krankheitsdauern verglichen.

Wie zuletzt beschrieben wurde die Studie in den Jahren 1997/98 durchgeführt. Dabei traten jedoch eine Vielzahl praktischer Probleme auf. Z.B. war es nicht einfach, genügend an Grippe erkrankter Personen zu finden. Für die Studie in Phase II konnte dieses Problem leicht gelöst werden, indem man auf gesunde Versuchspersonen zurückgriff, die bereit waren, sich künstlich mit einer relativ harmlosen Variante des Grippevirus infizieren zu lassen.

Da die Studie in Phase III die Wirksamkeit des Medikaments unter realistischen Bedingungen (wozu auch die Auswahl der zu behandelnden Patienten durch einen Arzt rein aufgrund der beobachteten Symptome gehörte) erforderte, war dieses Vorgehen in Phase III nicht möglich. Hier stellte sich auch das Problem, dass die Studiengruppe einen möglichst hohen Prozentsatz an Grippekranke enthalten musste, denn nur bei diesen verkürzt das Medikament die Krankheitsdauer. Die Diagnose einer Grippe ist schwierig, da eine Vielzahl von bakteriellen Infektionen (sog. grippale Infekte) anfangs ähnliche Symptome zeigen. Eine sichere Diagnose der Grippe kann über einen Halsabstrich erfolgen, dessen Auswertung aber in aller Regel länger als die Erkrankung dauert. Um dieses Problem zu lösen, wurden nur in solchen Gegenden Testpersonen rekrutiert, wo in der vergangenen Woche (über Halsabstriche) mindestens zwei Grippefälle nachgewiesen wurden.

Weiter wurde den Personen in der Kontrollgruppe anstelle des Medikaments eine gleich aussehende Kapsel ohne Wirkstoff (sog. *Placebo*) verabreicht. Dies sollte verhindern, dass es den Personen in der Studiengruppe allein durch Einnahme einer Tablette besser geht als denen in der Kontrollgruppe (sog. *Placebo–Effekt*). Um eine Beeinflussung der (manchmal schwierig zu beurteilenden) Symptome durch die Verordnung des Medikaments zu vermeiden, wurde darüberhinaus den behandelnden Ärzten nicht mitgeteilt, ob ein Patient zur Studien– oder zur Kontrollgruppe gehörte (sog. *doppelblinde Studie*).

Anfang 1998 war die Studie abgeschlossen. Insgesamt wurden 1355 Versuchspersonen rekrutiert. Die Auswertung von Halsabstrichen ergab, dass davon 70% wirklich an Grippe erkrankt waren. Wichtigstes Ergebnis war, dass die Einnahme des neuen Medikaments innerhalb von 36 Stunden nach Auftreten der ersten Symptome dazu führte, dass die Grippe etwa eineinhalb Tage früher abgeklungen war. Aufgrund dieses Ergebnisses wurde das Medikament zugelassen und ist heute unter dem Namen Tamiflu in Apotheken erhältlich.

Die Durchführung einer prospektiv kontrollierten Studie mit Randomisierung ist deutlich aufwendiger als die einer retrospektiv kontrollierten Studie. Dennoch lohnt sich der Aufwand, wie die folgenden beiden Beispiele zeigen.

Das erste Beispiel betrifft die Einführung eines Polio-Impfstoffes in den USA im Jahre 1954. Polio (genauer: Poliomyelitis, auf deutsch: Kinderlähmung) ist eine fäkal–oral übertragene Infektionskrankheit, die durch Viren ausgelöst wird. Sie ist in Europa und Nordamerika heutzutage wegen des dort häufig vorhandenen Impfschutzes nicht mehr stark verbreitet, in tropischen Ländern aber relativ häufig. Aufgrund von nachlassender Impfbereitschaft sind aber in den letzten Jahren auch in Europa und Nordamerika wieder einzelne Fälle aufgetreten.

An Polio erkranken vor allem Kleinkinder. Es handelt sich um eine Entzündung von Nervenzellen, die in Phasen verläuft. Anfangs hat man dabei grippeähnliche Symptome, dann treten Erkältungssymptome und Durchfall auf, schließlich kommt es zu Lähmungserscheinungen. An Polio sterben zwischen 20% und 60% der Erkrankten.

In den USA wurde in den 50er Jahren des letzten Jahrhunderts ein Impfstoff entwickelt. Nachdem dieser im Labor erfolgreich getestet worden war, wurde dessen Wirksamkeit im Rahmen einer prospektiv kontrollierten Studie mit Randomisierung überprüft. Das Resultat der so durchgeführten Studie ist in Tabelle 2.1 beschrieben.

	Größe	# Fälle	Infektionsrate
SG	200.000	56	28
KG	200.000	142	71
KZdE	350.000	161	46

Tabelle 2.1: Infektionsraten mit und ohne Impfung.

Dabei steht SG für Studiengruppe, KG für Kontrollgruppe, KZdE ist die Gruppe aller Kindern, bei denen die Eltern der Impfung nicht zugestimmt haben und Infektionsrate ist die Anzahl Polio–Fälle pro 100.000 Kinder. Die Bildung von Studien- und Kontrollgruppe erfolgte durch zufälliges Aufteilen der Kinder, deren Eltern einer Impfung zugestimmt hatten.

Vergleicht man die Infektionsraten in Studien– und Kontrollgruppe, so sieht man, dass die Impfung die Wahrscheinlichkeit, an Polio zu erkranken, senkt.

Vergleicht man darüberhinaus die Infektionsraten bei KG und KZdE, so sieht man, dass eine prospektiv kontrollierte Studie ohne Randomisierung das offensichtlich unsinnige Resultat ergeben würde, dass eine Impfung mit Salzlösung die Wahrscheinlichkeit, an Polio zu erkranken, erhöht. Dies lässt sich dadurch erklären, dass viele Eltern mit geringem Einkommen die Impfung verweigerten. Deren Kinder wachsen häufig in vergleichsweise unhygienischen Verhältnissen auf, kommen daher häufig schon in den ersten Lebensjahren mit einer abgeschwächten Variante des Polio–Erregers in Kontakt und sind deshalb weniger anfällig für Polio. Daher tritt das Problem auf, dass hier der Einfluss der Impfung mit der Salzlösung konfundiert mit dem Einfluss des Einkommens der Eltern.

Das zweite Beispiel zur Illustration der Vorteile einer prospektiv kontrollierten Studie mit Randomisierung betrifft Studien zu Bypass–Operationen. Zu Bypass–

	prospektiv, randomisiert	retrospektiv
Operation	87.6 %	90.9 %
keine Operation	83.2 %	71.1 %

Tabelle 2.2: Überlebensrate nach drei Jahren bei Studien zu Bypass-Operationen.

Operationen wurden mehrere Studien durchgeführt, die zu unterschiedlichen Resultaten kamen. Dabei äußerten sich von 8 prospektiv kontrollierten Studien mit Randomisierung 7 negativ und eine positiv über den Nutzen der Operation, während sich von 21 retrospektiv kontrollierten Studien 16 (und damit die Mehrzahl) positiv und nur 5 negativ äußerten. Der Unterschied läßt sich leicht erklären, wenn man die Überlebensraten nach drei Jahren betrachtet. Diese wurden bei 6 der prospektiv kontrollierten Studien und bei 9 der retrospektiv kontrollierten Studien angegeben. Das Resultat ist in Tabelle 2.2 dargestellt.

Man sieht, dass die Überlebensraten bei den operierten Patienten ungefähr gleich sind, bei den nicht operierten Patienten aber bei den retrospektiven Studien viel geringer als bei den prospektiven Studien ausfallen. Der Grund dafür ist, dass für die Operation nur die nicht zu kranken Patienten in Frage kommen. Daher konnten Studien– und Kontrollgruppe bei den prospektiv kontrollierten Studien mit Randomisierung nur aus nicht zu kranken Patienten bestehen, während diese Einschränkung bei der Kontrollgruppe der retrospektiv kontrollierten Studien nicht bestand.

2.2 Beobachtungsstudien

Bei den im letzten Abschnitt behandelten kontrollierten Studien wurde der Einfluss einer Einwirkung (z.B. Impfung) auf Objekte (z.B. Kinder) untersucht. Dabei konnte der Statistiker entscheiden, auf welche Objekte eingewirkt wird und auf welche nicht. Entsprechend der Entscheidung des Statistikers wurden dann die Objekte in Studien– und Kontrollgruppe unterteilt.

Nicht bei allen Fragestellungen ist es möglich, dass der Statistiker die Objekte in Studien– und Kontrollgruppe unterteilt. Möchte man z.B. eine Studie durchführen, die klären soll, ob Rauchen Krankheiten verursacht, so wird man kaum Teilnehmer finden, die bereit sind, je nach Anweisung des Statistikers die nächsten zehn Jahre intensiv bzw. gar nicht zu rauchen.

Studien, bei denen es prinzipiell unmöglich ist, dass der Statistiker die Objekte in Studien– und Kontrollgruppe einteilt, und daher die Objekte diese Einteilung selbst vornehmen, bezeichnet man als *Beobachtungsstudien*. Hauptproblem bei dieser Art von Studien ist, dass man nicht weiß, ob die Kontrollgruppe wirklich ähnlich zur Studiengruppe ist oder nicht.

Zur Illustration der Probleme, die bei Beobachtungsstudien auftreten können, werden im Folgenden einige Beispiele vorgestellt.

Zuerst wird nochmals die Frage betrachtet, ob Rauchen Krankheiten verursacht. Im Rahmen einer Beobachtungsstudie könnte man dazu die Todesraten von Rauchern und Nichrauchern vergleichen. Leider unterscheidet sich hierbei die Studiengruppe (bestehend aus allen Rauchern) nicht nur hinsichtlich des Rauchens von der Kontrollgruppe (bestehend aus allen Nichtrauchern). Da besonders viele Männer rauchen, sind nämlich z.B. Männer überproportional häufig in der Studiengruppe vertreten. Die Todesrate bei Männern ist, wegen dem häufigeren Auftreten von Herzerkrankungen, höher als die von Frauen. Damit ist das Geschlecht ein *konfundierter Faktor*, d.h. eine Einflussgröße, deren Einfluss auf die Todesrate sich mit dem des Rauchens vermengt. Ist nun die Todesrate in der Studiengruppe deutlich höher als in der Kontrollgruppe, so weiß man nicht, ob dies am Rauchen oder an dem konfundierten Faktor liegt.

Wie bei prospektiv kontrollierten Studien ohne Randomisierung kann man wieder versuchen, dieses Problem zu lösen, indem man nur Gruppen vergleicht, die bzgl. dieses konfundierten Faktors übereinstimmen. Dazu würde man im obigen Beispiel die Todesrate von männlichen Rauchern mit der von männlichen Nichtrauchern und die von weiblichen Rauchern mit der von weiblichen Nichtrauchern vergleichen. Dies löst das Problem aber nicht vollständig, da es weitere konfundierte Faktoren gibt, wie z.B. Alter (ältere Menschen unterscheiden sich sowohl hinsichtlich der Rauchgewohnheiten als auch bezüglich des Risikos, an Lungenkrebs zu erkranken, von jüngeren Menschen). Nötig ist daher die Erkennung aller konfundierter Faktoren und die Bildung von vielen Untergruppen.

Dass dies nicht immer richtig durchgeführt wird (bzw. werden kann), sieht man am nächsten Beispiel: In den 80er Jahren des letzten Jahrhunderts wurde am John Hopkins Krankenhaus in Baltimore (USA) im Rahmen einer Beobachtungsstudie untersucht, ob eine Ultraschalluntersuchung während der Schwangerschaft das Geburtsgewicht eines Kindes beeinflusst. Da zu der damaligen Zeit eine Ultraschalluntersuchung vor allem bei Risikoschwangerschaften durchgeführt wurde, war das durchschnittliche Geburtsgewicht der Kinder, bei denen im Verlauf der Schwangerschaft die Untersuchung durchgeführt wurde, natürlich geringer als bei den Kindern, bei denen diese Untersuchung nicht durchgeführt worden war. Das

Überraschende daran war aber, dass dieser Effekt auch nach Berücksichtigung einer Vielzahl von konfundierten Faktoren wie z.B. Rauchen, Alkoholgenuss, Ausbildung der Mutter, etc., d.h. nach Bildung einer Vielzahl von Untergruppen gemäß diesen Faktoren, noch bestand. Dies wurde anschließend im Rahmen einer kontrollierten Studie mit Randomisierung widerlegt: Diese ergab, dass bei den Schwangerschaften, bei denen eine Ultraschalluntersuchung durchgeführt worden war, das Geburtsgewicht im Schnitt sogar noch etwas höher war als beim Rest. Der Unterschied beim Geburtsgewicht lässt sich dadurch erklären, dass in der Studiengruppe überproportional viele Mütter das Rauchen aufgaben, nachdem sie bei der Ultraschalluntersuchung ihr Kind gesehen hatten.

Was für widersprüchliche Effekte konfundierte Faktoren verursachen können, lässt sich auch anhand von Daten belegen, die bei der Zulassung von Studenten an die Universität Berkeley im Herbst 1973 erhoben wurden. Dort hatten sich für das Master-/PhD-Programm 8442 Männer und 4321 Frauen beworben. Zugelassen wurden 44% der Männer und 35% der Frauen. Dies scheint zu belegen, dass Männer im Rahmen des Zulassungsverfahrens bevorzugt wurden.

Die einzelnen Fächer entschieden unabhängig voneinander, welche Studenten sie zulassen und welche nicht. Betrachtet man daher wie in Tabelle 2.3 die Zulassungsdaten nach Fachrichtungen getrennt, so sollte man ablesen können, bei welchen Fächern Frauen bei der Zulassung am meisten diskrimminiert werden.

Fach	#Männer	Zugel.	#Frauen	Zugel.
A	825	62%	108	**82%**
B	560	63%	25	**68 %**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
D	191	**28%**	393	24%
F	373	6%	341	**7%**

Tabelle 2.3: Zulassung zum Studium in Berkeley im Herbst 1973.

Diese Zahlen belegen aber, dass in allen Fächern entweder prozentual mehr Frauen oder aber prozentual fast so viele Frauen wie Männer zugelassen wurden. Dieser scheinbare Widerspruch lässt sich dadurch erklären, dass hier der Einfluss des Geschlechts auf die Zulassung konfundiert mit dem Einfluss der Wahl des Faches: Frauen haben sich vor allem für Fächer beworben, in denen nur wenige zugelassen wurden.

Eine Übersicht über die verschiedenen Arten von Studien findet man in Abbildung 2.1.

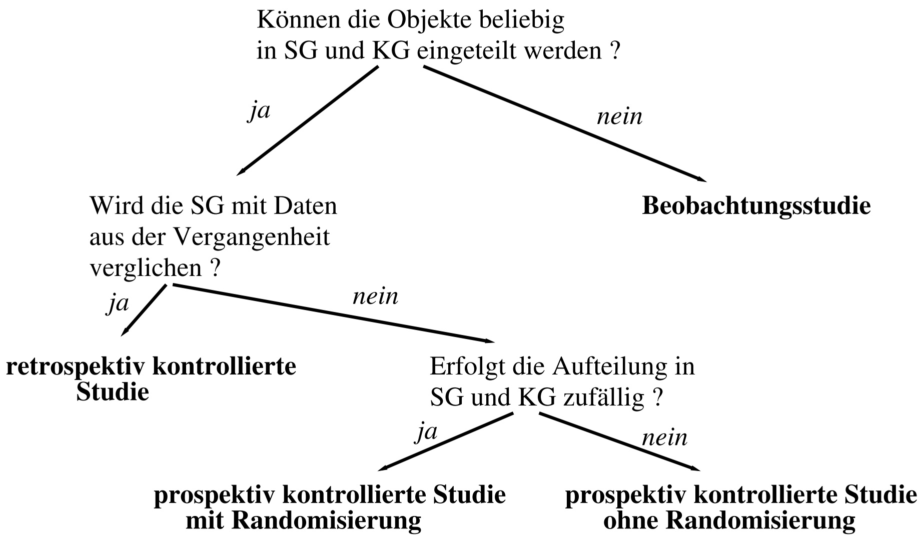

Abbildung 2.1: Übersicht über die verschiedenen Arten von Studien.

Zusammenfassend kann man sagen, dass eigentlich bei allen Studien zuerst einmal nur das gleichzeitige Auftreten (sogenante *Assoziation*) zweier Dinge nachgewiesen wird. Daraus möchte man auf einen kausalen Zusammenhang schließen. Insbesondere bei Beobachtungsstudien, retrospektiv kontrollierten Studien und bei prospektiv kontrollierten Studien ohne Randomisierung kann dieses gleichzeitige Auftreten aber auch an der Existenz konfundierter Faktoren liegen. Diese haben Einfluss sowohl auf die Aufteilung in Studien– und Kontrollgruppe als auch auf das beobachtete Resultat.

2.3 Umfragen

Bei einer Umfrage betrachtet man eine Menge von Objekten (*Grundgesamtheit*), wobei jedes der Objekte eine Reihe von Eigenschaften besitzt. Feststellen möchte man, wie viele Objekte der Grundgesamtheit eine gewisse vorgegebene Eigenschaft haben.

Ein Beispiel dafür ist die sogenannte Sonntagsfrage, über die regelmäßig in den Medien berichtet wird. Dabei möchte man wissen, wie viele der Wahlberechtigten in der BRD für die aktuelle Bundesregierung stimmen würden, wenn nächsten Sonntag Bundestagswahl wäre.

Tabelle 2.4 beinhaltet die Ergebnisse von Wahlumfragen, die von fünf verschiedenen Meinungsforschungsinstituten ca. drei Wochen vor der Bundestagswahl 2002 durchgeführt wurden, sowie das amtliche Endergebnis der Bundestagswahl am 22.09.2002. Wie man sieht, weichen die Umfrageergebnisse zum Teil erheblich vom tatsächlichen Wahlergebnis ab. Daraus kann man allerdings nicht auf Fehler bei den Umfragen schließen, da sich das Wahlverhalten der Deutschen in den letzten drei Wochen vor der Wahl noch geändert haben könnte. Allerdings sieht man an den Schwankungen der Umfrageergebnisse der verschiedenen Institute, dass zumindest bei einigen davon doch erhebliche Ungenauigkeiten bei der Vorhersage auftraten.

	SPD	CDU/CSU	FDP	GRÜNE	PDS
Allensbach	35,2	38,2	11,2	7,2	4,9
Emnid	37	39	8	6	5
Forsa	39	39	9	7	4
Forschungsgruppe Wahlen	38	38	8	7	4
Infratest-dimap	38	39,5	8,5	7,5	4
amtliches Endergebnis	**38,5**	**38,5**	**7,4**	**8,6**	**4,0**

Tabelle 2.4: Umfragen zur Budestagswahl 2002.

Wie man Umfragen durchführen kann und warum genaue Prognosen häufig schwierig sind, wird im Folgenden behandelt.

Die Bestimmung der Anzahl der Objekte einer Grundgesamtheit mit einer gewissen vorgegebenen Eigenschaft ist zunächst einmal eine rein deterministische Fragestellung, die man im Prinzip durch reines Abzählen entscheiden könnte. Bei vielen Fragestellungen (insbesondere bei der oben erwähnten Sonntagsfrage) ist die Betrachtung **aller** Objekte der Grundgesamtheit aber nicht möglich bzw. viel zu aufwendig.

Als Ausweg bietet sich an, nur für eine "kleine" Teilmenge (der Statistiker spricht hier von einer *Stichprobe*) der Grundgesamtheit zu ermitteln, wieviele Objekte darin die interessierende Eigenschaft haben, und dann zu versuchen, mit Hilfe dieses Resultats die gesuchte Größe näherungsweise zu bestimmen (der Statistiker spricht hier von *schätzen*). Dazu muss man erstens festlegen, wie man die Stichprobe wählt, und zweitens ein Verfahren entwickeln, das mit Hilfe der Stichprobe die gesuchte Größe schätzt.

Für die oben angesprochene Sonntagsfrage könnte man dazu wie folgt vorgehen: Zuerst wählt man "rein zufällig" n Personen (z.B. $n = 2000$) aus der Menge aller Wahlberechtigten aus und befragt diese bzgl. ihrem Wahlverhalten. Anschließend schätzt man den prozentualen Anteil der Stimmen für die aktuelle Bundesregierung in der Menge aller Wahlberechtigten durch den entsprechenden prozentualen Anteil in der Stichprobe. Wie wir in den weiteren Kapiteln dieses Skriptes sehen werden, liefert dies zumindest dann eine gute Schätzung, sofern die Stichprobe wirklich “rein zufällig” ausgewählt wurde. Damit steht man nur noch vor dem Problem, wie man letzteres durchführt. Dazu werden im weiteren die folgenden fünf Vorgehensweisen betrachtet:

Vorgehen 1: Befrage die Studenten einer Statistik-Vorlesung.

Vorgehen 2: Befrage die ersten n Personen, die Montag morgens ab 10 Uhr einen festen Punkt der Königsstraße in Stuttgart passieren.

Vorgehen 3: Erstelle eine Liste aller Wahlberechtigten (mit Adresse). Wähle aus dieser "zufällig" n Personen aus und befrage diese.

Vorgehen 4: Wähle aus einem Telefonbuch für Deutschland rein zufällig Nummern aus und befrage die ersten n Personen, die man erreicht.

Vorgehen 5: Wähle zufällig Nummern am Telefon, und befrage die ersten n Privatpersonen, die sich melden.

Betrachtet man diese bzgl. der praktischen Durchführbarkeit, so stellt sich Vorgehen 3 als sehr aufwendig heraus: Die zu befragenden Personen sind dabei im allgemeinen nämlich über die gesamte BRD verstreut, zudem werden die Adressen nicht immer aktuell sein. Darüberhinaus gibt es Länder (wie z.B. die USA), wo Listen aller Wahlberechtigten gar nicht erst existieren.

Bei allen anderen Vorgehensweisen tritt eine sogenannte *Verzerrung durch Auswahl* (*sampling bias*) auf. Diese beruht darauf, dass die Stichprobe nicht *repräsentativ* ist, d.h. dass bestimmte Gruppen der Wahlberechtigten, deren Wahlverhalten vom Durchschnitt abweicht, überrepräsentiert sind. Z.B sind dies bei Vorgehen 1 die Studenten, bei Vorgehen 2 die Einwohner von Stuttgart sowie Personen, die dem Interviewer sympathisch sind, bei Vorgehen 4 Personen mit Eintrag im Telefonbuch und bei Vorgehen 5 Personen, die telefonisch leicht erreichbar sind sowie Personen, die in einem kleinen Haushalt leben. Bei Vorgehen 5 lässt sich dieses Problem teilweise umgehen, indem man dort bei einzelnen Nummern mehrmals anruft, sofern man nicht sofort jemanden erreicht, und in dem man die Person, die man unter dieser Nummer befragt, nach demographischen Aspekten auswählt (wie z.B. "befrage jüngsten Mann, der älter als 18 ist und zu Hause ist").

Bei allen fünf Vorgehensweisen tritt darüberhinaus noch eine *Verzerrung durch Nicht–Antworten* (*non–response bias*) auf. Diese beruht darauf, dass ein Teil der Befragten die Antwort verweigern wird, und dass das Wahlverhalten dieser Personen unter Umständen vom Rest abweicht. Außerdem werden im allgemeinen nur sehr wenige Personen zugeben, dass sie nicht zur Wahl gehen, und auch deren Wahlverhalten kann vom Rest abweichen.

In den USA werden vom Meinungsforschungsinstitut Gallup seit 1988 telefonische Wahlumfragen durchgeführt. Dabei wird die USA zuerst gemäß Zeitzone und Bevölkerungsdichte unterteilt, dann wird für jeden Teil eine Umfrage mit Hilfe von zufälliger Wahl von Telefonnummern durchgeführt. Aus den Angaben der Personen in der Stichprobe wird durch *gewichtete Mittelung* die Schätzung bestimmt. Dabei gehen bei der Wahl der Gewichte auch demographische Faktoren ein, weiter wird dadurch versucht zu vermeiden, dass Personen, die in kleinen Haushalten leben, ein zu großes Gewicht in der Stichprobe bekommen.

Kapitel 3

Deskriptive und explorative Statistik

In diesem Kapitel werden einige Methoden der *deskriptiven* (oder *beschreibenden*) und der *explorativen* (oder *erforschenden*) Statistik eingeführt. Ausgangspunkt im Folgenden ist eine sogenannte **Messreihe** (auch *Stichprobe* oder *Datensatz* genannt), die mit

$$x_1, \ldots, x_n$$

bezeichnet wird. Hierbei ist n der Stichprobenumfang ist. Die Aufgabe der deskriptiven Statistik ist die übersichtliche Darstellung von Eigenschaften dieser Messreihe. Die explorative Statistik stellt Methoden zum Auffinden von (unbekannten) Strukturen in Datensätzen zur Verfügung.

Als Beispiel wird im Folgenden die Ankunftszeit von Studenten in der Vorlesung Statistik I für WirtschaftswissenschaftlerInnen am 26.10.01 betrachtet. Die Veranstaltung begann für alle Studenten um 8.45 Uhr mit Vortragsübungen. Diese gingen bis 9.30 Uhr, um 9.45 Uhr begann die eigentliche Vorlesung. Von 40 zufällig ausgewählten Studenten wurde im Rahmen einer Umfrage die Ankunftszeit ermittelt. Man erhielt

> -5, -5, -45, -15, 55, -15, 65, 55, -15, 0, -61, -15, 10, 65, -2, -35, 0, 47, 5, -30, 50, -30, 45, -65, -10, -15, -45, 5, 55, -30, 55, 35, 55, 45, -45, -55, 75, -15, -10, -45

wobei hier die Angabe in Minuten relativ zu Beginn der Vortragsübungen um 8.45 Uhr erfolgt. In diesem Beispiel ist $n = 40$, $x_1 = -5$, $x_2 = -5$, $\ldots$,$x_{40} = -45$. Betrachtet man alle diese Zahlen zusammen, so verliert man aufgrund der Vielzahl

	Abstandbegriff vorhanden ?	Ordnungsrelation vorhanden ?
reell	ja	ja
ordinal	nein	ja
zirkulär	ja	nein
nominal	nein	nein

Tabelle 3.1: Typen von Messgrößen.

der Zahlen leicht den Überblick. Die deskriptive Statistik stellt nun Verfahren bereit, wie man die in solchen Zahlenreihen vorhandene Information in wenige Zahlen oder Abbildungen zusammenfassen kann.

Bevor darauf näher eingegangen werden soll, werden zunächst die Typen von **Messgrößen** (oder auch *Merkmalen*, *Variablen*), die auftreten können, betrachtet. Hierbei gibt es verschiedene Unterteilungsmöglichkeiten. Z.B. kann man sie gemäß der Anzahl der auftretenden Ausprägungen unterteilen: Treten nur endlich oder abzählbar unendlich viele Ausprägungen auf, so spricht man von einer *diskreten* Messgröße, treten dagegen alle Werte eines Intervalls als Werte auf, so spricht man von einer *stetigen* Messgröße.

Eine andere mögliche Unterteilung erfolgt gemäß der Struktur des Wertebereichs der Messgröße. Dabei betrachtet man, ob für alle Paare von Werten dieser Messgröße ein Abstand (Entfernung zwischen den beiden Werten) und / oder eine Ordnungsrelation (Anordnung der Werte der Größe nach) definiert ist. Wie in Tabelle 3.1 dargestellt spricht man dann von *reellen, ordinalen, zirkulären* oder *nominalen* Messgrößen. Beispiel für eine reelle Messgröße ist die oben betrachtete Ankunftszeit bei der Statistik–Vorlesung relativ zu Beginn der Vortragsübungen, Beispiel einer ordinalen Messgröße sind z.B. Noten (die sicher der Größe nach geordnet werden können, bei denen aber z.B. der Abstand von 1 und 2 nicht so groß ist wie der zwischen 4 und 5 und daher nicht als Differenz der Noten festgelegt werden kann), Beispiel einer zirkulären Messgröße ist die Uhrzeit und Beispiel einer nominalen Messgröße ist die Parteizugehörigkeit einer Person.

Die Beachtung der Typen von Messgrößen ist insofern wichtig, da viele statistischen Verfahren zunächst einmal nur für reelle Messgrößen entwickelt wurden. Wendet man diese auf nicht–reelle Messgrößen an, so kann es sein, dass die implizite Annahme der Existenz eines Abstandsbegriffes und einer Ordnungsrelation zu einem unsinnigen Ergebnis führt.

3.1 Histogramme

Ausgangspunkt zur Erstellung eines Histogrammes ist eine sogenannte *Häufigkeitstabelle*. Bei dieser wird der Wertebereich der betrachteten reellen oder ordinalen Messgröße in k disjunkte (d.h. nicht überlappende) Klassen unterteilt, und in einer Tabelle wird für jede der Klassen die Anzahl n_i der Datenpunkte der Messreihe, die in dieser Klasse liegen, angegeben ($i = 1, \ldots, k$).

Klasse	Häufigkeit
1	n_1
2	n_2
$\vdots$	$\vdots$
k	n_k

Für die Wahl der Anzahl k von Klassen existieren Faustregeln wie z.B. $k \approx \sqrt{n}$ oder $k \approx 10 \cdot \log_{10} n$. Oft erfolgt diese aber subjektiv, insbesondere bei Verwendung graphischer Darstellungen wie z.B den unten beschriebenen Säulendiagrammen bzw. Histogrammen.

Im Beispiel oben erhält man bei Unterteilung der Ankunftszeiten in 8 Klassen als Häufigkeitstabelle

Zeit	Häufigkeit
$[-80, -60)$	2
$[-60, -40)$	5
$[-40, -20)$	4
$[-20, 0)$	13
$[0, 20)$	3
$[20, 40)$	1
$[40, 60)$	9
$[60, 80)$	3

Dabei steht das Intervall $[a, b) = \{x \in \mathbb{R} \; : \; a \leq x < b\}$ für die Klasse aller Ankunftszeiten, die in diesem Intervall liegen.

Die Häufigkeitstabelle lässt sich graphisch recht übersichtlich als *Säulendiagramm* darstellen. Dazu trägt man über jeder Klasse einen Balken mit Höhe gleich der Anzahl Datenpunkte in der Klasse ab. Im Beispiel oben erhält man das in Abbildung 3.1 dargestellte Säulendiagramm.

Diese graphische Darstellung ist aber irreführend, falls die Klassen nicht alle gleich lang sind. Möchte man z.B. wissen, wieviele Studenten in der Vorlesung pünktlich zur Vortragsübung bzw. pünktlich zur Vorlesung erschienen sind und

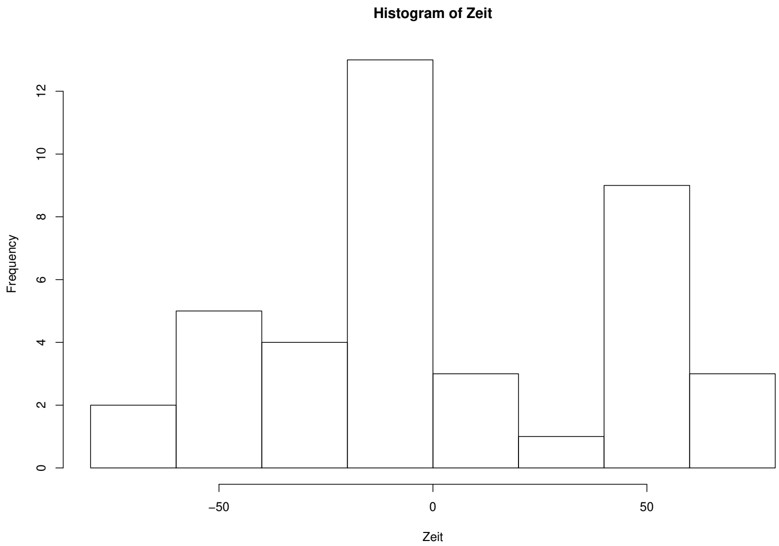

Abbildung 3.1: Säulendiagramm der Ankunftszeiten.

wieviele vermutlich fälschlicherweise gedacht haben, dass die Vortragsübungen schon um 8:00 Uhr beginnen, so kann man die Ankunftszeiten in Klassen wie in der unten stehenden Häufigkeitstabelle unterteilen.

Zeit	Häufigkeit
$[-65, -45)$	7
$[-45, 0)$	17
$[0, 15)$	3
$[15, 60)$	10
$[60, 80]$	3

Das zugehörige Säulendiagramm ist in Abbildung 3.2 dargestellt.

Betrachtet man nun nur dieses Säulendiagramm, so ist der Flächeninhalt des zur Klasse $[-45, 0)$ gehörenden Rechtecks mehr als fünfmal so groß wie der Flächeninhalt des zur Klasse $[-65, -45)$ gehörenden Rechtecks. Dadurch entsteht der falsche Eindruck, dass die Klasse $[-45, 0)$ mehr als fünfmal so viele Datenpunkte enthält wie die Klasse $[-65, -45)$.

Diesen falschen Eindruck kann man vermeiden, indem man bei der graphischen Darstellung nicht die Höhe sondern den Flächeninhalt proportional zur Anzahl (oder zur relativen Häufigkeit) der Datenpunkte in einer Klasse wählt. Dies führt auf das sogenannte **Histogramm.**

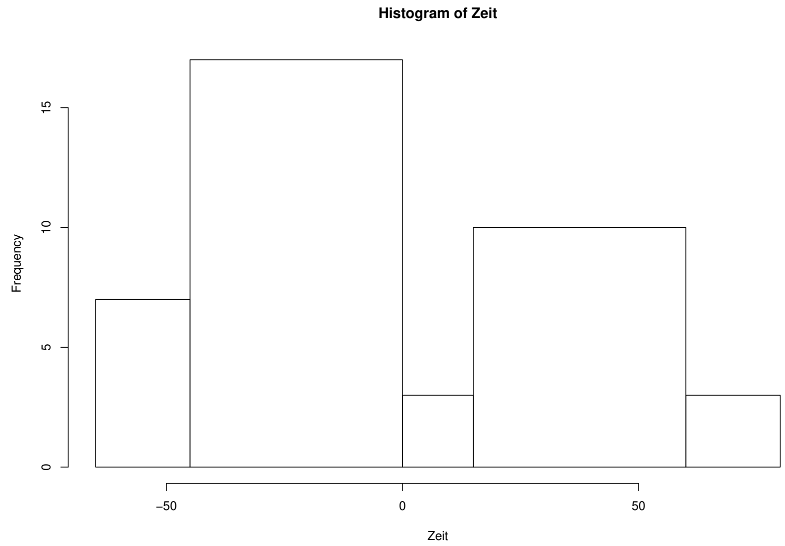

Abbildung 3.2: Säulendiagramm der Ankunftszeiten.

Dabei unterteilt man wieder den Wertebereich der (reellen oder ordinalen) Messgröße in k Intervalle $I_1, \dots I_k$, bestimmt für jedes dieser Intervall I_j die Anzahl n_j der Datenpunkte in diesem Intervall und trägt dann über I_j den Wert

$$\frac{n_j}{n \cdot \lambda(I_j)}$$

auf. Dabei bezeichnet $\lambda(I_j)$ die Länge von I_j.

Im Beispiel oben erhält man das in Abbildung 3.3 dargestellte Histogramm.

Wie man sieht, gibt hier der Flächeninhalt eines Rechtecks den prozentualen Anteil der Datenpunkte im zugrunde liegenden Intervall an.

3.2 Dichteschätzung

Beim Histogramm wird die Lage der Messreihe auf dem Zahlenstrahl durch eine stückweise konstante Funktion beschrieben. Die Vielzahl der Sprungstellen dieser Funktion erschwert häufig die Interpretation der zugrunde liegenden Struktur. Dies lässt sich durch Anpassung einer "glatten" Funktion (z.B. einer differenzierbaren Funktion) vermeiden. Dabei wird wieder wie beim Histogramm gefordert, dass die Funktion nichtnegativ ist, dass ihr Flächeninhalt Eins ist, und dass die

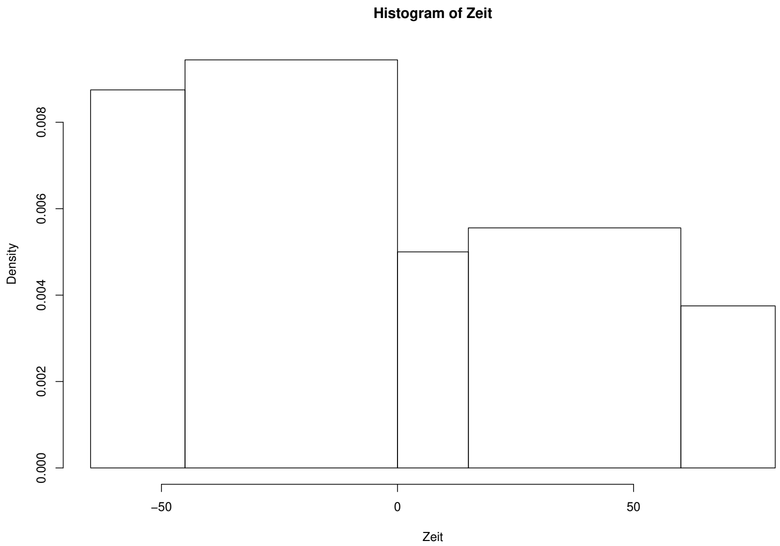

Abbildung 3.3: Histogramm der Ankunftszeiten.

Anzahl der Datenpunkte in einem Intervall proportional zum Flächeninhalt zwischen der Funktion und diesem Intervall ist. Funktionen mit den ersten beiden Eigenschaften heißen *Dichten*.

Definition 3.1 *Eine Funktion* $f : \mathbb{R} \to \mathbb{R}$ *mit*

$$f(x) \geq 0 \quad \text{für alle } x \in \mathbb{R}$$

und

$$\int_{\mathbb{R}} f(x)\, dx = 1$$

heißt **Dichte**.

Die Konstruktion einer Dichte, die eine Menge von Datenpunkten im obigen Sinne beschreibt, kann z.B. durch Bildung eines Histogrammes erfolgen. Im Folgenden soll dessen Konstruktion so abgeändert werden, dass glatte Dichten entstehen. Dazu wird zuerst das sogenannte *gleitende Histogramm* eingeführt. Bei diesem werden zur Bestimmung des Funktionswertes an einer Stelle x alle Datenpunkte betrachtet, die im Intervall $[x-h, x+h]$ ($h > 0$ fest) enthalten sind. Analog zum Histogramm wird der Funktionswert berechnet durch

$$\begin{aligned} f_h(x) &= \frac{\frac{1}{n} \cdot \text{Anzahl Datenpunkte } x_i \text{ in } [x-h, x+h]}{2h} \\ &= \frac{1}{n \cdot h} \sum_{i=1}^{n} \frac{1}{2} \cdot 1_{[x-h,x+h]}(x_i). \end{aligned} \tag{3.1}$$

Hierbei ist 1_A die Indikatorfunktion zu einer Menge A, d.h., $1_A(x) = 1$ für $x \in A$ und $1_A(x) = 0$ für $x \notin A$. Im Unterschied zum Histogramm hängt hierbei das der Berechnung zugrunde liegende Intervall $[x-h, x+h]$ von x ab und ist um x zentriert. Letzteres hat den Vorteil, dass Datenpunkte, die gleichweit von x entfernt sind, den gleichen Einfluss auf den Funktionswert an der Stelle x haben. Mit

$$
\begin{aligned}
1_{[x-h,x+h]}(x_i) = 1 \qquad &\Leftrightarrow \qquad x-h \leq x_i \leq x+h \quad \Leftrightarrow \quad -1 \leq \frac{x_i - x}{h} \leq 1 \\
&\Leftrightarrow \qquad -1 \leq \frac{x - x_i}{h} \leq 1
\end{aligned}
$$

folgt, dass sich das gleitende Histogramm $f_h(x)$ kompakter schreiben lässt gemäß

$$
f_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{3.2}
$$

wobei $K : \mathbb{R} \to \mathbb{R}$ gegeben ist durch $K(u) = \frac{1}{2} \cdot 1_{[-1,1]}(u)$. Wegen $K(u) \geq 0$ für alle $u \in \mathbb{R}$ und $\int_{\mathbb{R}} K(u)\, du = 1$ ist K selbst eine Dichtefunktion.

(3.2) kann gedeutet werden als arithmetisches Mittel von Dichtefunktionen, die um die $x_1, \ldots, x_n$ konzentriert sind. In der Tat sieht man leicht, dass mit K auch

$$
u \mapsto \frac{1}{h} K\left(\frac{u - x_i}{h}\right) \tag{3.3}
$$

eine Dichtefunktion ist. Diese entsteht aus K durch Verschiebung des Ursprungs an die Stelle x_i und anschließende Stauchung (im Falle $h < 1$) bzw. Streckung (im Falle $h > 1$).

Mit $K = \frac{1}{2} 1_{[-1,1]}$ sind auch (3.3) sowie das arithmetische Mittel (3.2) unstetig. Dies lässt sich vermeiden, indem man für K stetige Dichtefunktionen wählt, wie z.B.

$$
K(u) = \begin{cases} \frac{3}{4}(1-u^2) & \text{für} \quad -1 \leq u \leq 1, \\ 0 & \text{für} \quad u < -1 \text{ oder } u > 1, \end{cases}
$$

(sog. *Epanechnikov-Kern*) oder

$$
K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/2\right)
$$

(sog. *Gauss-Kern*).

Die Funktion

$$
f_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \qquad (x \in \mathbb{R})
$$

ist der sogenannte *Kern-Dichteschätzer* von Nadaraya und Watson. Sie hängt von K (einer Dichtefunktion, sog. *Kernfunktion*) und h (einer reellen Zahl größer als Null, sog. *Bandbreite*) ab.

Das Ergebnis der Anwendung des Kern-Dichteschätzers auf die Ankunftszeiten aus Abbildung 3.2 ist in Abbildung 3.4 dargestellt. Dabei werden der Gauss-Kern sowie verschiedene Werte für die Bandbreite verwendet.

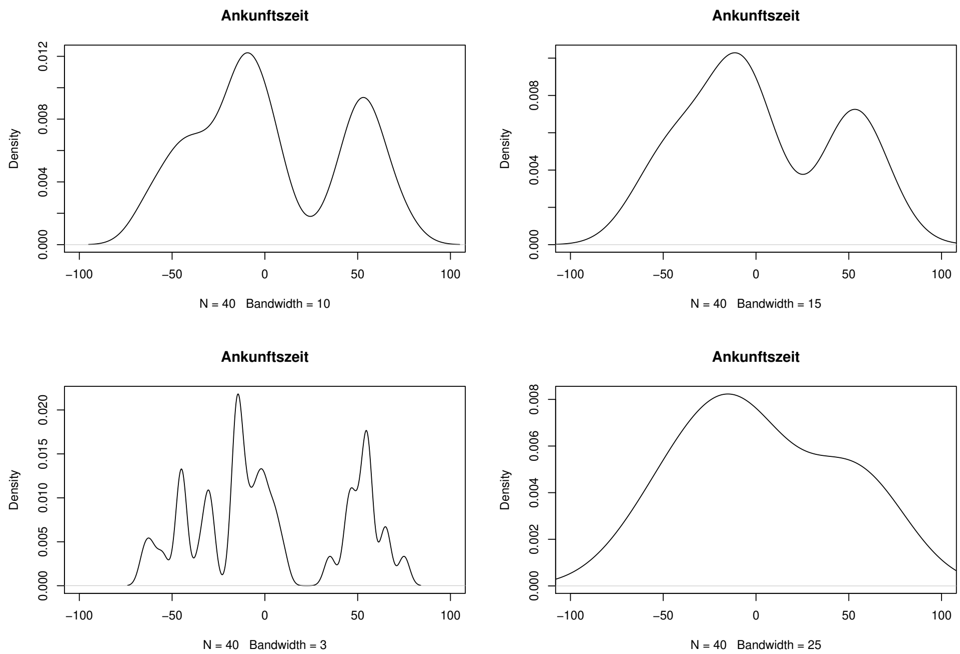

Abbildung 3.4: Kerndichteschätzer für Ankunftszeiten.

Wie man sieht, lässt sich mittels h die "Glattheit" des Kern-Dichteschätzers $f_h(x)$ kontrollieren: Ist h sehr klein, so wird $f_h(x)$ als Funktion von x sehr stark schwanken, ist dagegen h groß, so variiert $f_h(x)$ als Funktion von x kaum noch.

Es ist keineswegs offensichtlich, wie man den Wert von h bei Anwendung auf einen konkreten Datensatz wählen soll. Ohne Einführung von mathematischen Modellen versteht man an dieser Stelle auch nicht richtig, was man überhaupt macht und kann nur schlecht Verfahren zur Wahl der Bandbreite erzeugen.

Abschließend wird noch ein weiteres Beispiel für den Einsatz eines Dichteschätzers gegeben. In einer im Rahmen einer Diplomarbeit an der Universität Stuttgart durchgeführten kontrollierten Studie mit Randomisierung wurde der Einfluss eines Crash-Kurses auf die Noten in einer Statistik-Prüfung untersucht. Ziel der Diplomarbeit war die Entwicklung eines Verfahrens zur Identifikation von durchfallgefährdeten Studenten. Nach Entwicklung eines solchen Verfahren stellte sich die Frage, ob man durch Abhalten eines Crash-Kurses zur Wiederholung des Stoffes die Noten bzw. die Durchfallquote bei diesen Studenten verbessern kann. Dazu wurden 60 Studenten mit Hilfe des Verfahrens ausgewählt und zufällig in zwei Gruppen (Studien- und Kontrollgruppe) mit jeweils 30 Studenten unterteilt. Die Studenten aus der Studiengruppe wurden vor der Prüfung schriftlich zu einem Crash-Kurs eingeladen, die aus der Kontrollgruppe nicht.

In Abbildung 3.5 ist das Ergebnis der Anwendung eines Kern-Dichteschätzer mit Gauss-Kern und verschiedenen Bandbreiten auf die Noten in Studien- und Kontrollgruppe dargestellt. Wie man sieht, hatte der Crash-Kurs den erfreulichen Effekt, dass Noten im Bereich 5.0 in der Studiengruppe deutlich seltener auftraten als in der Kontrollgruppe. Darüberhinaus variieren aber auch die Noten in der Studiengruppe insgesamt etwas weniger als in der Kontrollgruppe, so dass auch sehr gute Noten in der Studiengruppe etwas seltener auftreten. Dies lässt sich dadurch erklären, dass die Studenten nach Besuch des Crash-Kurses kaum Zeit zum individuellen Lernen auf die Prüfung hatten und sich daher auch nicht überproportional gut auf die Prüfung vorbereiten konnten.

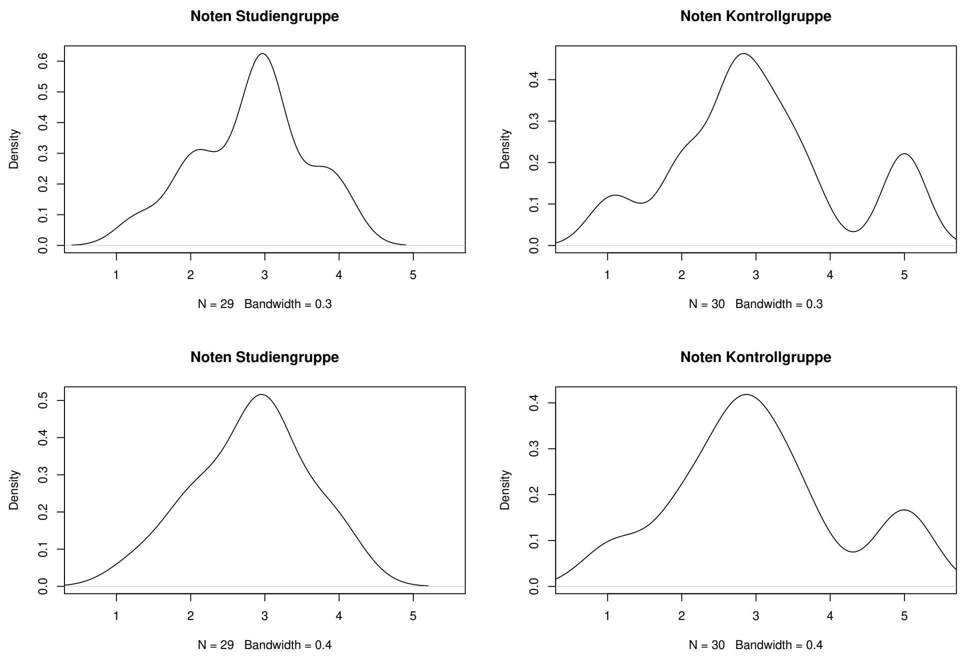

Abbildung 3.5: Einfluss eines Crash-Kurses auf Abschneiden bei einer Prüfung.

3.3 Statistische Maßzahlen

Im Folgenden werden verschiedene statistische Maßzahlen eingeführt. Diese kann man unterteilen in *Lagemaßzahlen* und *Streuungsmaßzahlen*. Lagemaßzahlen geben an, in welchem Bereich der Zahlengeraden die Werte (oder die "Mitte" der Werte) der betrachteten Messreihe liegt. Streuungsmaßzahlen dienen zur Beschreibung des Bereiches, über den sich die Werte im wesentlichen erstrecken, insbesondere kann man aus diesen ablesen, wie stark die Werte um die "Mitte" der Werten schwanken.

Im Folgenden sei

$$x_1, \dots, x_n$$

die Messreihe. Die der Größe nach aufsteigend sortierten Werte seien

$$x_{(1)}, \dots, x_{(n)}.$$

Als Beispiel werden Mathematik-Noten (Note in der letzten Mathematik-Prüfung vor Besuch der Vorlesung, in der Regel handelt es sich dabei um die Abiturprüfung) von 38 zufällig ausgewählten Studenten der Vorlesung Statistik für WirtschaftswissenschaflerInnen betrachtet. Hier sind die $x_1, \dots, x_n$ gegeben durch

1.0, 2.7, 3.0, 2.7, 2.7, 2.0, 1.0, 2.5, 2.0, 1.0, 1.3, 4.0, 1.7, 2.7, 2.0, 4.0, 4.0, 3.5, 2.7, 2.0, 4.0, 4.0, 1.0, 1.7, 2.5, 2.0, 2.0, 2.0, 3.0, 3.0, 1.0, 3.0, 1.0, 2.3, 1.0, 1.0, 3.3, 3.3.

Die der Größe nach aufsteigend geordneten Werte $x_{(1)}, \dots, x_{(n)}$ sind

1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.3, 1.7, 1.7, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.3, 2.5, 2.5, 2.7, 2.7, 2.7, 2.7, 2.7, 3.0, 3.0, 3.0, 3.0, 3.3, 3.3, 3.5, 4.0, 4.0, 4.0, 4.0, 4.0

Beispiele für Lageparameter sind das *(empirische arithmetische) Mittel* und der *(empirische) Median*.

Beim *(empirischen arithmetischen) Mittel* teilt man die Summe aller Messgrößen durch die Anzahl der Messgrößen:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot (x_1 + x_2 + \dots + x_n)$$

Bei den Noten oben erhält man $\bar{x} = 2.358$.

Nachteil des arithmetischen Mittels ist, dass es einerseits nur für reelle Messgrößen berechnet werden kann (das dabei vorgenommene Mitteln von Abständen setzt implizit voraus, dass Abstände definiert sind) und dass es andererseits sehr stark durch sogenannte *Ausreißer* beeinflusst werden kann. Darunter versteht man Werte, die "sehr stark" von den anderen Werten abweichen. Wie man leicht sieht, führt im oben angebenen Beispiel bereits eine (z.B. aufgrund eines Tippfehlers) sehr große Note zu einer starken Änderung des arithmetischen Mittels.

In diesen Fällen ist der sogenannte *(empirische) Median*, definert als

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{falls} \quad n \text{ ungerade,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{falls} \quad n \text{ gerade,} \end{cases}$$

bzw. - sofern die x_i nicht reell sind - definiert gemäß

$$\tilde{x} = x_{(\lceil \frac{n}{2} \rceil)}$$

besser geeignet. Hierbei bezeichnet $\lceil \frac{n}{2} \rceil$ die kleinste ganze Zahl, die größer oder gleich $n/2$ ist (z.B. $\lceil 39/2 \rceil = 20$, $\lceil 40/2 \rceil = 20$ und $\lceil 41/2 \rceil = 21$). Der empirische Median hat die Eigenschaft, dass ungefähr $n/2$ der Datenpunkte kleiner oder gleich und ebenfalls ungefähr $n/2$ der Datenpunkte größer oder gleich wie der empirische Median sind.

Im Beispiel oben erhält man $\tilde{x} = 2.4$ bzw. $\tilde{x} = 2.5$.

Beispiele für Streuungsparameter sind die *(empirische) Spannweite*, die *(empirische) Varianz*, die *(empirische) Standardabweichung*, der *Variationskoeffizient* und der *Interquartilabstand*.

Die *(empirische) Spannweite* oder Variationsbreite ist definiert als

$$r := x_{max} - x_{min} := x_{(n)} - x_{(1)}.$$

Sie gibt die Länge des Bereichs an, über den sich die Datenpunkte erstrecken. Im Beispiel oben erhält man $r = 4 - 1 = 3$.

Die *(empirische) Varianz* beschreibt, wie stark die Datenpunkte um das empirische Mittel schwanken. Sie ist definiert als arithmetisches Mittel der quadratischen Abstände der Datenpunkte vom empirischen Mittel:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \cdot \left((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \right).$$

Die Mittelung durch $n-1$ statt durch n kann dabei folgendermaßen plausibel gemacht werden: Da

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n \cdot \bar{x} = 0$$

gilt, ist z.B. die letzte Abweichung $x_n - \bar{x}$ bereits durch die ersten $n-1$ Abweichungen festgelegt. Somit variieren nur $n-1$ Abweichungen frei und man mittelt indem man die Summe durch die Anzahl $n-1$ der sogenannten Freiheitsgrade teilt. Eine mathematisch exakte Begründung dafür erfolgt in Kapitel 5.

Im Beispiel oben erhält man $s^2 = 0.986\ldots$.

Die *(empirische) Standardabweichung* oder Streuung ist definiert als die Wurzel aus der (empirischen) Varianz:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Im Beispiel oben erhält man $s = 0.993$.

Die Größe der empirischen Standardabweichung relativ zum empirischen Mittel beschreibt der sogenannte *Variationskoeffizient*, definiert durch

$$V = \frac{s}{\bar{x}}.$$

Für nichtnegative Messreihen mit $\bar{x} > 0$ ist der Variationskoeffizient maßstabsunabhängig und kann daher zum Vergleich der Streuung verschiedener Messreihen verwendet werden.

Im Beispiel oben erhält man $V = 0.421$.

Wie das empirische Mittel sind auch alle diese Streuungsparameter bei nicht–reellen Messgrößen oder beim Vorhandensein von Ausreißern nicht sinnvoll. Hier kann man dann aber den sogennanten *Interquartilabstand* verwenden, der definiert ist als Differenz des 25% größten und des 25% kleinsten Datenpunktes:

$$IQR = x_{(\lceil \frac{3}{4}n \rceil)} - x_{(\lceil \frac{1}{4}n \rceil)}$$

Im Beispiel oben erhält man $IQR = 3 - 1.7 = 1.3$.

Einige dieser Lage- und Streuungsparameter werden im sogenannten *Boxplot* graphisch dargestellt (vgl. Abbildung 3.6). Dabei beschreibt die mittlere waagrechte

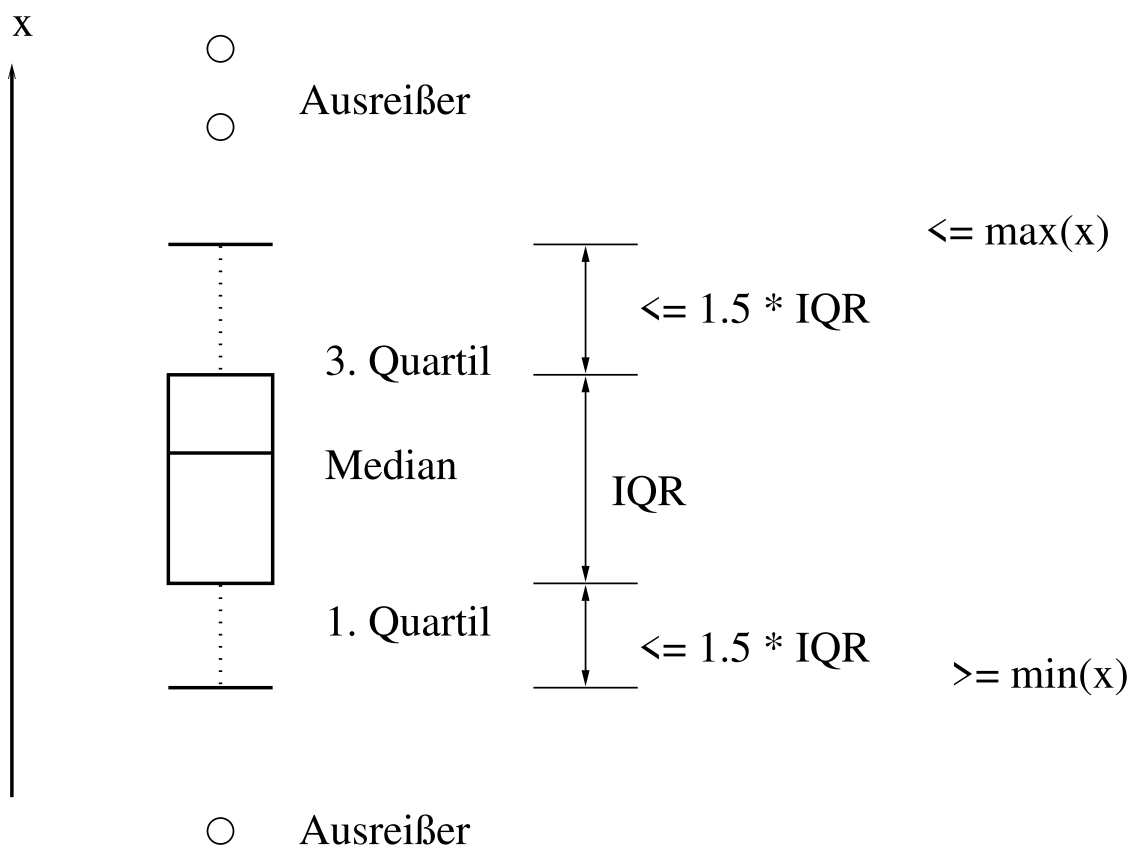

Abbildung 3.6: Darstellung einer Messreihe im Boxplot.

Linie die Lage des Medians, die obere Kante des Rechtecks die Lage des 25% größten Datenpunktes (3. Quartil) und die untere Kante des Rechtecks die Lage des 25% kleinsten Datenpunktes (1. Quartil). Die Länge des Rechtecks ist gleich dem Interquartilabstand. Datenpunkte, deren Abstand nach oben bzw. nach unten vom 3. Quartil bzw. vom 1. Quartil größer als 1.5 mal dem Interquartilabstand ist, werden als Ausreißer betrachtet und durch Kreise gesondert dargestellt. Bezüglich den restlichen Datenpunkten gibt die oberste bzw. die unterste waagrechte Linie die Lage des Maximums bzw. des Minimums an.

Der zum obigen Beispiel gehörende Boxplot ist in Abbildung 3.7 dargestellt.

Mit Hilfe von Boxplots kann man auch sehr schön verschiedene Mengen von Datenpunkten vergleichen. Betrachtet man z.B. die Mathematik-Noten der Studenten, die pünktlich bzw. unpünktlich zur Vortragsübung erschienen sind, so kann man erkennen, dass dabei eine *Verzerrung durch Auswahl* auftritt, so dass eine Umfrage bzgl. der Mathematik-Noten, die zu Beginn der Vortragsübungen durchgeführt worden wäre, ein falsches Resultat geliefert hätte (vgl. Abbildung

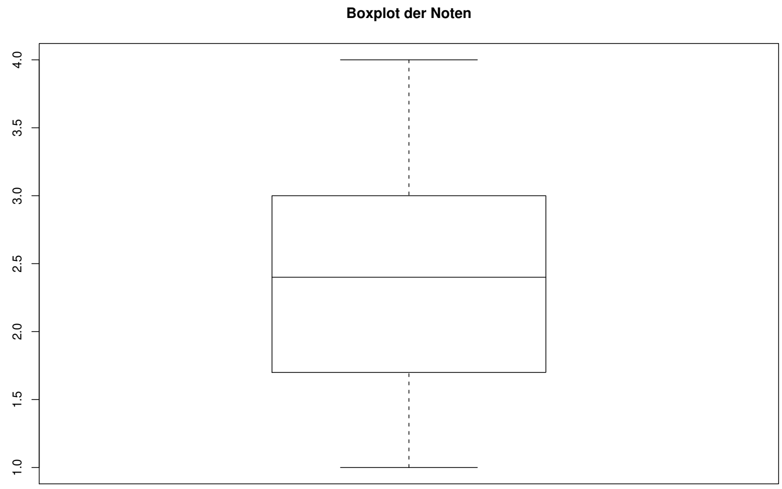

Abbildung 3.7: Boxplot der Mathematik-Noten.

3.8).

Das gleiche Phänomen tritt auch bei der Frage nach dem Interesse am Vorlesungsstoff auf (vgl. Abbildung 3.9, man beachte aber, dass die hier angegebenen Boxplot leicht irreführend sind, da die Wertebereiche an den Achsen verschieden sind).

3.4 Regressionsrechnung

Bei der Regressionsrechnung betrachtet man mehrdimensionale Messreihen (d.h. die betrachtete Messgröße besteht aus mehreren Komponenten) und man interessiert sich für Zusammenhänge zwischen den verschiedenen Komponenten der Messgröße. Um diese zu bestimmen, versucht man, eine der Komponenten durch eine Funktion der anderen Komponenten zu approximieren.

Der Einfacheit halber wird im Folgenden nur eine zweidimensionale Messreihe betrachtet, diese wird mit

$$(x_1, y_1), \ldots, (x_n, y_n)$$

bezeichnet. Hier ist n wieder der Stichprobenumfang. Herausgefunden werden soll, ob ein Zusammenhang zwischen den x– und den y–Koordinaten der Datenpunkte besteht.

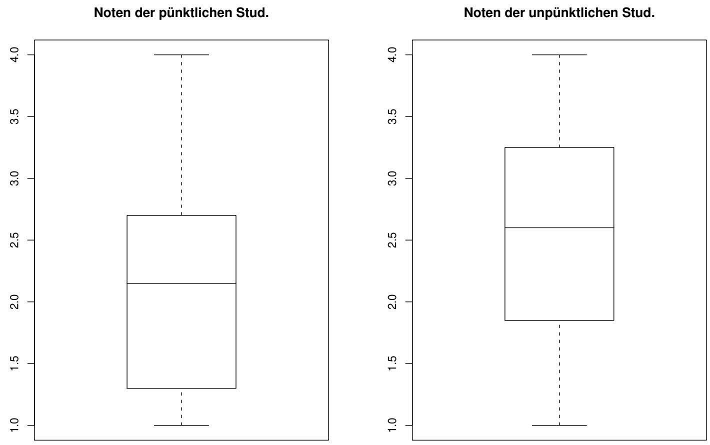

Abbildung 3.8: Vergleich der Noten der pünktlichen und der unpünktlichen Studenten.

Als Beispiel wird das Ergebnis einer Umfrage betrachtet, die in der Vorlesung "Statistik I für WirtschaftswissenschaftlerInnen" am 26.10.01 durchgeführt wurde. Dabei wurden 40 zufällig ausgewählte Studenten unter anderem nach ihrer Ankunftszeit bei der Vorlesung (Angabe in Minuten relativ zum Veranstaltungsbeginn), nach der Note in ihrer letzten Mathematik–Prüfung (Angabe als Zahl zwischen 1 und 6) sowie nach ihrem Interesse an der Vorlesung (Angabe als Zahl zwischen 1 und 5, 1 = sehr geringes Interesse, 5 = sehr großes Interesse) befragt. Wissen möchte man nun, ob hier einerseits ein Zusammenhang zwischen der Ankunftszeit und der Mathematik–Note sowie andererseits ein Zusammenhang zwischen der Ankunftszeit und dem Interesse an der Vorlesung besteht. Dazu könnte man natürlich wieder die Studenten in pünktliche und unpünktliche Studenten unterteilen und die Mathematik–Noten bzw. das Interesse an der Vorlesung getrennt in Boxplots darstellen. Gefragt ist jetzt aber nach einem funktionalem Zusammenhang zwischen Ankunftszeit und Note, der z.B. auch beschreibt wie stark die Note schwankt wenn man die Ankunftszeit von -10 Minuten auf +5 Minuten verändert.

Eine erste Möglichkeit um einen optischen Eindruck davon zu bekommen, ist eine Darstellung der Messreihe im sogenannten *Scatterplot* (bzw. Streudiagramm). Dabei trägt man für jeden Wert (x_i, y_i) der Messreihe den Punkt mit den Koordinaten (x_i, y_i) in ein zweidimensionales Koordinatensystem ein. Für das obige Beispiel sind die Scatterplots in den Abbildungen 3.10 und 3.11 angegeben. Dabei steht ein Punkt im Koordinatensystem unter Umständen für mehrere Datenpunkten mit den gleichen (x_i, y_i)-Werten. In Abbildung 3.10 repräsentieren

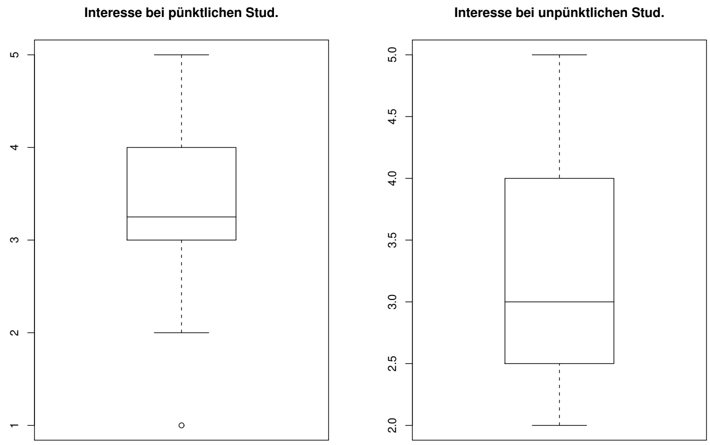

Abbildung 3.9: Pünktlichkeit und Interesse an der Statistik-Vorlesung.

Datenpunkte mit y–Koordinate gleich -1 Studenten, die keine Angabe zur Note in der letzten Mathematik-Prüfung gemacht haben.

Eine Möglichkeit zur Bestimmung einer funktionalen Abhängigkeit ist die sogenannte **lineare Regression**. Bei dieser passt man eine Gerade

$$y = a \cdot x + b$$

an die Daten an.

Eine weit verbreitete (aber keineswegs die einzige) Möglichkeit dafür ist das *Prinzip der Kleinsten-Quadrate*, bei dem $a, b \in \mathbb{R}$ durch Minimierung der Summe der quadratischen Abstände der Datenpunkte zu den zugehörigen Punkten auf der Geraden gewählt werden. Dazu muss man

$$\sum_{i=1}^{n} (y_i - (a \cdot x_i + b))^2 = (y_1 - (a \cdot x_1 + b))^2 + \cdots + (y_n - (a \cdot x_n + b))^2$$

bzgl. $a, b \in \mathbb{R}$ minimieren. Die zugehörige Gerade nennt man **Regressionsgerade**.

Vor der Herleitung einer allgemeinen Formel zur Berechnung der Regressionsgeraden wird zuerst ein Beispiel betrachtet. Sei $n = 3$, $(x_1, y_1) = (0, 0)$, $(x_2, y_2) = (1, 2)$ und $(x_3, y_3) = (2, 2)$. Zur Berechnung der Regressionsgeraden muss man dann diejenigen Zahlen $a, b \in \mathbb{R}$ bestimmen, für die

$$(0 - (a \cdot 0 + b))^2 + (2 - (a \cdot 1 + b))^2 + (2 - (a \cdot 2 + b))^2 \tag{3.4}$$

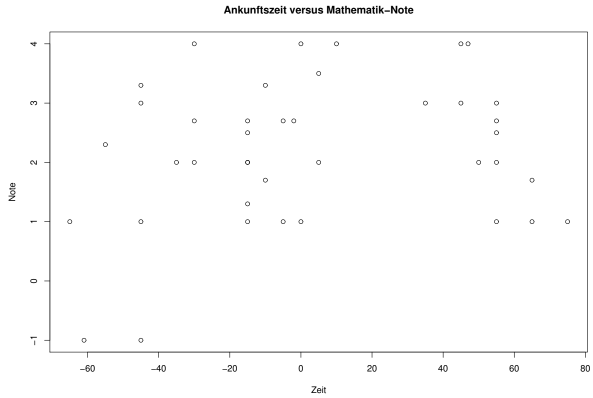

Abbildung 3.10: Zusammenhang zwischen Ankunftszeit und Mathematik-Note.

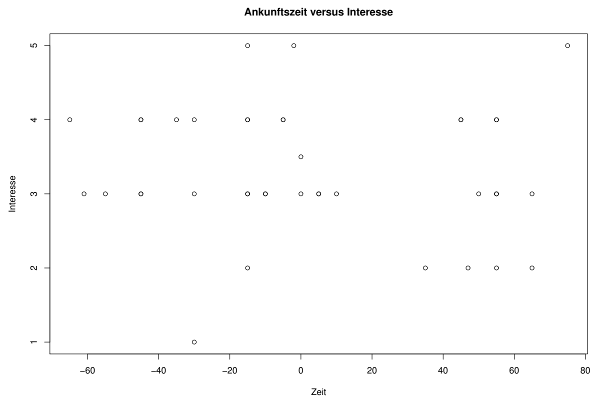

Abbildung 3.11: Zusammenhang zwischen Ankunftszeit und Interesse an Vorlesung.

minimal wird. Für diese Zahlen gilt, dass die Funktionen

$$f(u) = (0 - (u \cdot 0 + b))^2 + (2 - (u \cdot 1 + b))^2 + (2 - (u \cdot 2 + b))^2$$

und

$$g(v) = (0 - (a \cdot 0 + v))^2 + (2 - (a \cdot 1 + v))^2 + (2 - (a \cdot 2 + v))^2$$

Minimalstellen für $u = a$ bzw. $v = b$ haben. Also muss die Ableitung

$$f'(u) = 2 \cdot (0 - (u \cdot 0 + b)) \cdot 0 + 2 \cdot (2 - (u \cdot 1 + b)) \cdot (-1) + 2 \cdot (2 - (u \cdot 2 + b)) \cdot (-2)$$

von f an der Stelle $u = a$ sowie die Ableitung

$$g'(v) = 2 \cdot (0 - (a \cdot 0 + v)) \cdot (-1) + 2 \cdot (2 - (a \cdot 1 + v)) \cdot (-1) + 2 \cdot (2 - (a \cdot 2 + v)) \cdot (-1)$$

von g an der Stelle $v = b$ Null sein.

Damit folgt, dass $a, b \in \mathbb{R}$ Lösungen des linearen Gleichungssystems

$$\begin{aligned}(2 - (a \cdot 1 + b)) + (2 - (a \cdot 2 + b)) \cdot 2 &= 0\\(0 - (a \cdot 0 + b)) + (2 - (a \cdot 1 + b)) + (2 - (a \cdot 2 + b)) &= 0\end{aligned}$$

sein müssen, was äquivalent ist zu

$$\begin{aligned}5a + 3b &= 6\\3a + 3b &= 4.\end{aligned}$$

Durch Subtraktion der zweiten Gleichung von der ersten erhält man $a = 1$, Einsetzen in die erste Gleichung liefert $b = 1/3$, so dass in diesem Beispiel die Regressionsgerade gegeben ist durch

$$y = x + \frac{1}{3}.$$

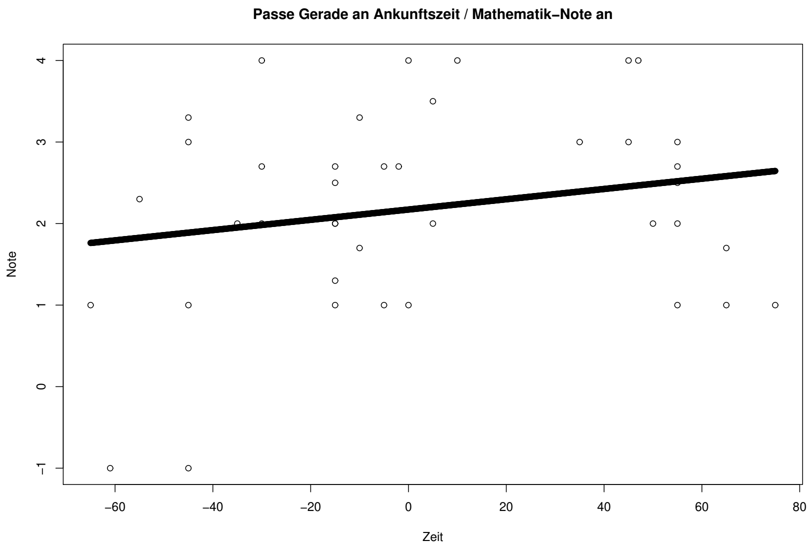

Abbildung 3.12: Lineare Regression angewandt auf Ankunftszeit und Mathematik-Note.

Im Folgenden soll nun für allgemeine $(x_1, y_1), \ldots, (x_n, y_n)$ die zugehörige Regressionsgerade bestimmt werden. Dazu muss man

$$\sum_{i=1}^{n} (y_i - (a \cdot x_i + b))^2 \tag{3.5}$$

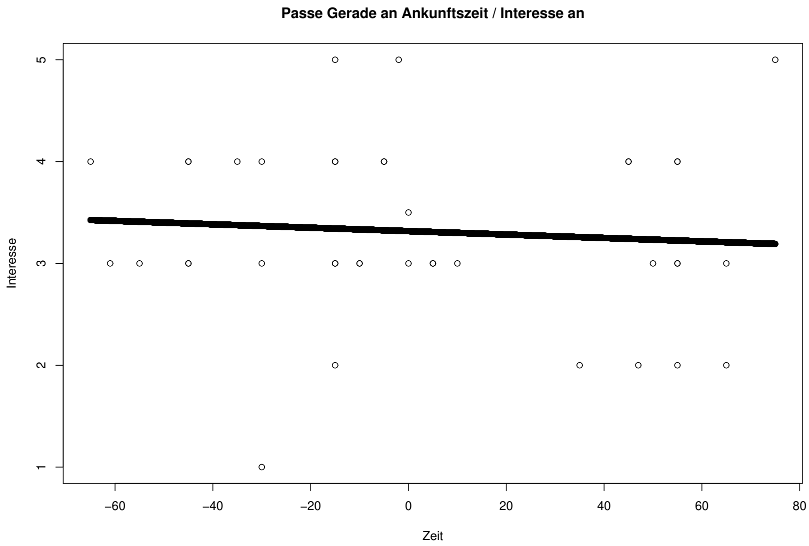

Abbildung 3.13: Lineare Regression angewandt auf Ankunftszeit und Interesse an Vorlesung.

bzgl. $a, b \in \mathbb{R}$ minimieren.

Wird der Ausdruck (3.5) für $a, b \in \mathbb{R}$ minimal, so müssen die Funktionen

$$f(u) = \sum_{i=1}^{n} (y_i - (u \cdot x_i + b))^2 \text{ und } g(v) = \sum_{i=1}^{n} (y_i - (a \cdot x_i + v))^2$$

an den Stellen $u = a$ bzw. $v = b$ Minimalstellen haben. Durch Nullsetzen der Ableitungen erhält man

$$0 = f'(a) = \sum_{i=1}^{n} 2 \cdot (y_i - (a \cdot x_i + b)) \cdot (-x_i) = -2 \cdot \sum_{i=1}^{n} x_i y_i + 2a \cdot \sum_{i=1}^{n} x_i^2 + 2b \cdot \sum_{i=1}^{n} x_i$$

und

$$0 = g'(b) = \sum_{i=1}^{n} 2 \cdot (y_i - (a \cdot x_i + b)) \cdot (-1) = -2 \cdot \sum_{i=1}^{n} y_i + 2a \cdot \sum_{i=1}^{n} x_i + 2b \cdot \sum_{i=1}^{n} 1,$$

was äquivalent ist zum linearen Gleichungssystem

$$\begin{aligned} a \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2 + b \cdot \frac{1}{n} \sum_{i=1}^{n} x_i &= \frac{1}{n} \sum_{i=1}^{n} x_i y_i \\ a \cdot \frac{1}{n} \sum_{i=1}^{n} x_i + b &= \frac{1}{n} \sum_{i=1}^{n} y_i. \end{aligned}$$

Aus der zweiten Gleichung erhält man

$$b = \bar{y} - a \cdot \bar{x},$$

wobei

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ und } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Setzt man dies in die erste Gleichung ein, so folgt

$$a \cdot \frac{1}{n}\sum_{i=1}^{n} x_i^2 + (\bar{y} - a \cdot \bar{x}) \cdot \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i,$$

also

$$a \cdot \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}.$$

Mit

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2 \cdot \bar{x} \cdot \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

und

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}) &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \frac{1}{n}\sum_{i=1}^{n} y_i - \bar{y} \cdot \frac{1}{n}\sum_{i=1}^{n} x_i + \bar{x} \cdot \bar{y} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y} \end{aligned}$$

folgt

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Damit ist gezeigt, dass die Regressionsgerade, d.h. die Gerade, die (3.5) minimiert, gegeben ist durch

$$y = \hat{a} \cdot (x - \bar{x}) + \bar{y},$$

wobei

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

und

$$\hat{a} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{x,y}}{s_x^2}$$

($\frac{0}{0} := 0$).

Hierbei wird

$$s_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

als *empirische Kovarianz* der zweidimensionalen Messreihe bezeichnet.

Da das Vorzeichen der empirischen Kovarianz mit dem der Steigung der Regressionsgeraden übereinstimmt, gilt, dass die empirische Kovarianz genau dann *positiv* (bzw. negativ) ist, wenn die Steigung der Regressionsgeraden *positiv* (bzw. negativ) ist.

Nach Konstruktion gilt darüberhinaus

$$\begin{aligned} 0 \leq \sum_{i=1}^{n} \left(y_i - (\hat{a}(x_i - \bar{x}) + \bar{y})\right)^2 &\leq \sum_{i=1}^{n} \left(y_i - (0 \cdot (x_i - \bar{x}) + \bar{y})\right)^2 \\ &= (n-1) \cdot s_y^2. \end{aligned}$$

Mit

$$\begin{aligned} &\sum_{i=1}^{n} \left(y_i - (\hat{a}(x_i - \bar{x}) + \bar{y})\right)^2 \\ &= \sum_{i=1}^{n} \left((y_i - \bar{y}) - \hat{a} \cdot (x_i - \bar{x})\right)^2 \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\hat{a} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) + \hat{a}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= (n-1) \cdot s_y^2 - 2 \cdot \hat{a} \cdot (n-1) \cdot s_{x,y} + (n-1) \cdot \hat{a}^2 s_x^2 \\ &= (n-1) \cdot s_y^2 \left(1 - 2\hat{a} \cdot \frac{s_{x,y}}{s_y^2} + \hat{a}^2 \frac{s_x^2}{s_y^2}\right) \\ &= (n-1) \cdot s_y^2 \left(1 - 2\frac{s_{x,y}}{s_x^2} \cdot \frac{s_{x,y}}{s_y^2} + \frac{s_{x,y}^2}{s_x^2 s_x^2} \cdot \frac{s_x^2}{s_y^2}\right) \\ &= (n-1) \cdot s_y^2 \cdot \left(1 - \frac{s_{x,y}^2}{s_x^2 \cdot s_y^2}\right) \end{aligned} \tag{3.6}$$

folgt

$$0 \leq (n-1) \cdot s_y^2 \cdot \left(1 - \frac{s_{x,y}^2}{s_x^2 \cdot s_y^2}\right) \leq (n-1) \cdot s_y^2.$$

Daraus wiederum folgt, dass die sogenannte **empirische Korrelation**

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2}}$$

im Intervall $[-1, 1]$ liegt.

Die empirische Korrelation dient zur Beurteilung der Abhängigkeit der x– und der y–Koordinaten. Sie macht Aussagen über die Regressionsgerade und die Lage der Punktwolke im Scatterplot. Die folgenden Aussagen ergeben sich unmittelbar aus der obigen Herleitung:

- Die empirische Korrelation ist im Falle $s_y \neq 0$ genau dann $+1$ oder -1, wenn gilt
$$\sum_{i=1}^n (y_i - (\hat{a} \cdot (x_i - \bar{x}) + \bar{y}))^2 = 0$$
(vgl. (3.6)), was wiederum genau dann der Fall ist wenn die Punkte (x_i, y_i) alle auf der Regressionsgeraden liegen.

- Ist die empirische Korrelation *positiv* (bzw. negativ), so ist auch die Steigung der Regressionsgeraden *positiv* (bzw. negativ).

- Ist die empirische Korrelation Null, so verläuft die Regressionsgerade waagrecht.

Die empirische Korrelation misst die Stärke eines *linearen* Zusammenhangs zwischen den x- und den y-Koordinaten. Da die Regressionsgerade aber auch dann waagrecht verlaufen kann, wenn ein starker nicht-linearer Zusammenhang besteht (z.B. bei badewannenförmigen oder runddachförmigen Punktwolken), und in diesem Fall die empirische Korrelation Null ist, kann durch Betrachtung der empirischen Korrelation nicht geklärt werden, ob überhaupt ein Zusammenhang zwischen den x- und den y-Koordinaten besteht.

Bei der linearen Regression passt man eine lineare Funktion an die Daten an. Dies ist offensichtlich nicht sinnvoll, sofern der Zusammenhang zwischen x und y nicht gut durch eine lineare Funktion approximiert werden kann. Ob dies der Fall ist oder nicht, ist insbesondere für hochdimensionale Messsreihen (Dimension von x > 1) nur schlecht feststellbar.

3.5 Nichtparametrische Regressionsschätzung

Bei der linearen Regression wird eine lineare Funktion an die Daten angepasst. Dies lässt sich sofort verallgemeinern hinsichtlich der Anpassung allgemeinerer Funktionen (z.B. Polynome) an die Daten. Dazu gibt man die gewünschte Bauart der Funktion vor. Sofern diese nur von endlichen vielen Parametern abhängt, kann man Werte dazu analog zur linearen Regression durch Anwendung des Prinzips der Kleinsten-Quadrate bestimmen, was auf ein Minimierungsproblem für die gesuchten Parameter führt. Schätzverfahren, bei denen die Bauart der anzupassenden Funktion vorgegeben wird und nur von endlich vielen Parametern abhängt, bezeichnet man als *parametrische Verfahren*. Im Gegensatz dazu stehen die sogenannten *nichtparametrischen Verfahren*, bei denen man keine Annahme über die Bauart der anzupassenden Funktion macht.

Einfachstes Beispiel für eine nichtparametrische Verallgemeinerung der linearen Regression ist die Regressionsschätzung durch *lokale Mittelung*. Dabei versucht man, den durchschnittlichen Verlauf der y-Koordinaten der Datenpunkte in Abhängigkeit der zugehörigen x-Koordinaten zu beschreiben. Dazu bildet man zu gegebenem Wert von x ein gewichtetes Mittel der Werte der y-Koordinaten aller der Datenpunkte, deren x-Koordinate nahe an diesem Wert liegt. Die Gewichte bei der Mittelung wählt man in Abhängigkeit des Abstands der x-Koordinate von dem vorgegebenen Wert.

Formal lässt sich dies z.B. durch den sogenannten *Kernschätzer* beschreiben, der gegeben ist durch

$$m_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \cdot y_i}{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)}.$$

Hierbei ist $K : \mathbb{R} \to \mathbb{R}_+$ die sogenannte *Kernfunktion*. Für diese fordert man üblicherweise, dass sie nichtnegativ ist, monoton in $|x|$ fällt und für $|x| \to \infty$ gegen Null konvergiert. Beispiele dafür sind der *naive Kern*

$$K(u) = \frac{1}{2} 1_{[-1,1]}(u)$$

oder der *Gauss-Kern*

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2).$$

Als weiteren Parameter hat der Kernschätzer die sogenannte *Bandbreite* $h > 0$. Wie beim Kern-Dichteschätzer bestimmt diese die Glattheit bzw. Rauheit der Schätzung.

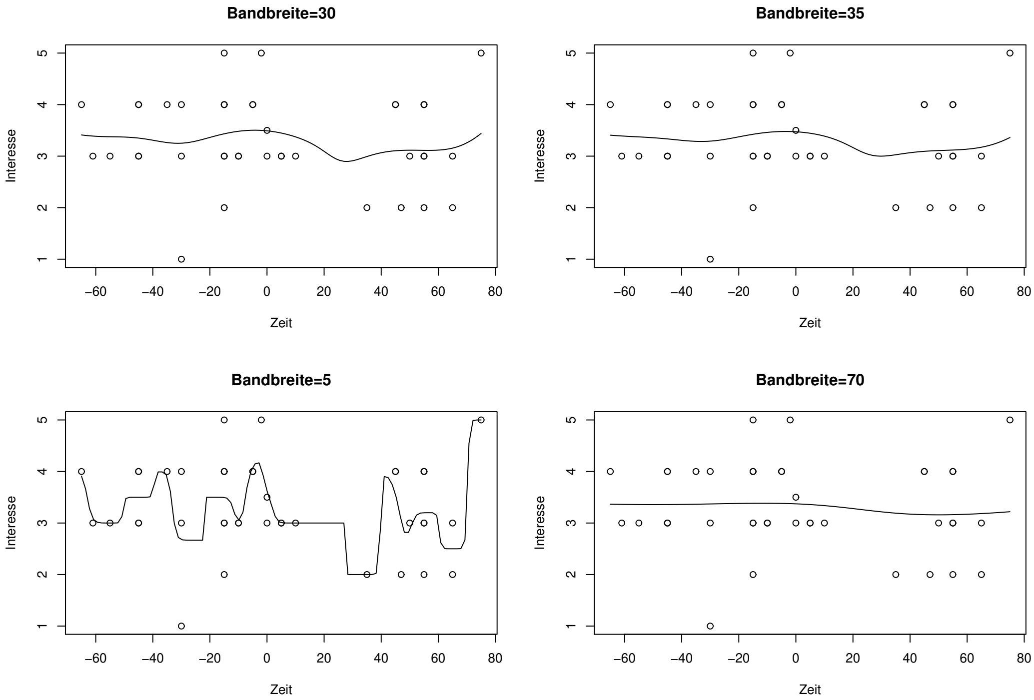

Abbildung 3.14: Kernschätzer angewandt auf Ankunftszeit und Interesse an Vorlesung.

Das Resultat der Anwendung eines Kernschätzers mit Gauss-Kern und verschiedenen Bandbreiten auf die Daten aus Abbildung 3.12 und Abbildung 3.13 ist in den Abbildungen 3.14 und 3.15 dargestellt.

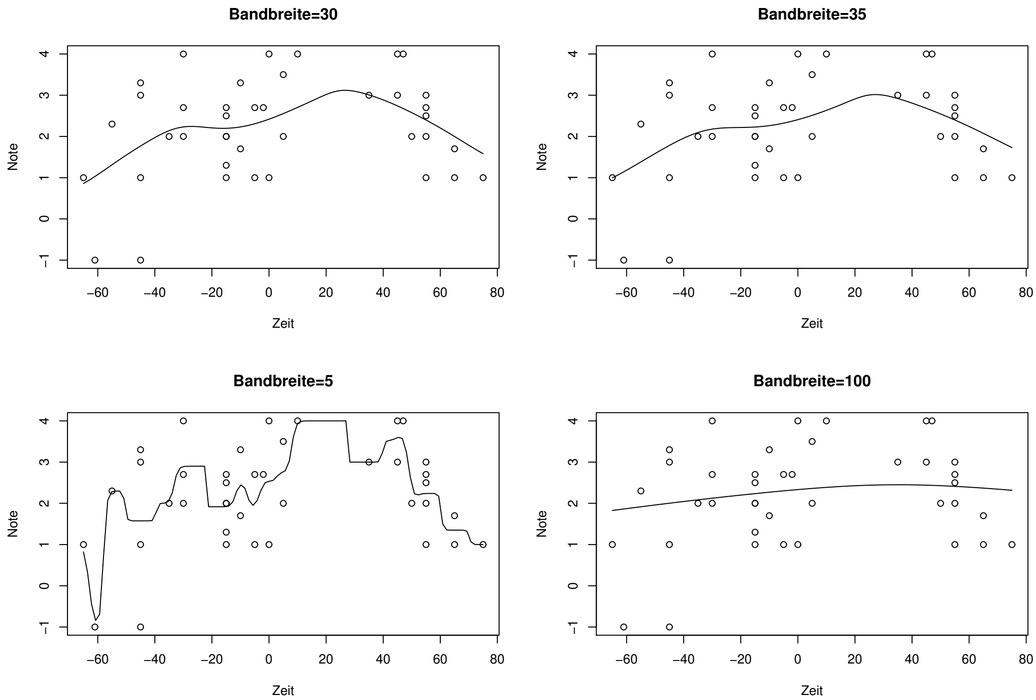

Abbildung 3.15: Kernschätzer angewandt auf Ankunftszeit und Mathematik-Note.

Kapitel 4

Grundlagen der Wahrscheinlichkeitstheorie

In diesem Kapitel beschäftigen wir uns mit der mathematischen Beschreibung *zufälliger* Phänomene. Dabei kann das Auftreten des Zufalls verschiedene Ursachen haben: Zum einen kann es auf unvollständiger Information basieren. Ein Beispiel dafür wäre ein Münzwurf, bei dem man sich vorstellen kann, dass bei exakter Beschreibung der Ausgangslage (Startposition der Münze, Beschleunigung am Anfang) das Resultat (Münze landet mit Kopf oder mit Zahl nach oben) genau berechnet werden kann. Allerdings ist es häufig unmöglich, die Ausgangslage genau zu beschreiben, und es bietet sich daher eine stochastische Modellierung an, bei der man die unbestimmten Größen als zufällig ansieht. Zum anderen kann das Auftreten des Zufalls zur Vereinfachung der Analyse eines deterministischen Vorgangs künstlich eingeführt werden. Beispiele dafür wurden bereits in Kapitel 2 gegeben, wo man statt einer (sehr aufwendigen) Befragung der gesamten Grundmenge bei einer Umfrage nur eine zufällig ausgewählte kleine Teilmenge betrachtet hat.

4.1 Grundaufgaben der Kombinatorik

Manchmal lassen sich Fragestellungen der Wahrscheinlichkeitstheorie durch einfaches Abzählen der "günstigen" bzw. "möglichen" Fälle bestimmen. Dafür sind die in diesem Abschnitt behandelten Formeln der Kombinatorik extrem nützlich.

Betrachtet wird das Ziehen von k Elementen aus einer Grundmenge Ω vom Umfang $|\Omega| = n$. Die Anzahl aller möglichen Stichproben sei N.

Dabei kann man vier verschiedene Vorgehensweisen unterscheiden, und zwar je nachdem, ob man die Elemente unmittelbar nach dem Ziehen wieder zurücklegt oder nicht, und je nachdem, ob man die Reihenfolge, in der die Elemente gezogen werden, beachtet oder nicht.

Zuerst betrachten wir das Ziehen **mit Zurücklegen** und **mit Berücksichtigung der Reihenfolge**. Hierbei wird k mal ein Element aus der Grundmenge gezogen, dabei hat man jeweils n Möglichkeiten, so dass man für die Anzahl der möglichen Stichproben erhält:

$$N = n \cdot n \cdot n \cdot \ldots \cdot n = n^k.$$

Als nächstes wird das Ziehen **ohne Zurücklegen** und **mit Berücksichtigung der Reihenfolge** betrachetet. Hier hat man für das erste Elemente n Möglichkeiten, für das zweite aber nur noch $n-1$, für das dritte $n-2$, u.s.w., und für das k-te noch $(n-k+1)$ Möglichkeiten. Damit erhält man für die Anzahl der möglichen Stichproben:

$$N = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}.$$

Dabei ist $n! = n \cdot (n-1) \cdot \cdots \cdot 1$ (gesprochen: n Fakultät) die sogenannte Fakultät von n.

Nun wird das Ziehen **ohne Zurücklegen** und **ohne Berücksichtigung der Reihenfolge** betrachtet. Ordnet man jede der dabei erhaltenen Stichproben auf alle $k!$ möglichen Weisen um, so erhält man alle Stichproben bzgl. Ziehen ohne Zurücklegen und mit Berücksichtigung der Reihenfolge.

Beispiel: Für $\Omega = \{1, 2, 3\}$, $n = 3$ und $k = 2$ erhält man die Zuordnungen

$$\begin{array}{rcl} (1,2) & \mapsto & (1,2) \text{ oder } (2,1) \\ (1,3) & \mapsto & (1,3) \text{ oder } (3,1) \\ (2,3) & \mapsto & (2,3) \text{ oder } (3,2) \end{array}$$

Daher gilt für die Anzahl der möglichen Stichproben:

$$\begin{aligned} & N \cdot k! \\ = \; & \text{Wert beim Ziehen ohne Zurücklegen und mit Berücksichtigung} \\ & \text{der Reihenfolge} \\ = \; & \frac{n!}{(n-k)!}, \end{aligned}$$

also
$$N = \frac{n!}{(n-k)! \cdot k!} = \binom{n}{k}.$$
Hierbei ist $\binom{n}{k}$ (gesprochen: n über k) der sogenannte Binomialkoeffizient.

Beispiel 4.1 *Binomischer Lehrsatz.*

Zur Illustration der Nützlichkeit der obigen Formel zeigen wir im Folgenden, dass für beliebige $a, b \in \mathbb{R}$, $n \in \mathbb{N}$ gilt:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

(sogenannter Binomischer Lehrsatz).

Beweis: Wir schreiben $(a+b)^n$ in die Form

$$(a+b)^n = (a+b) \cdot (a+b) \cdot \cdots \cdot (a+b),$$

wobei das Produkt aus genau n Faktoren besteht. Beim Ausmultiplizieren kann man sich bei jedem Faktor für a oder b entscheiden. Wählt man k–mal a und $(n-k)$–mal b, so erhält man den Summanden $a^k b^{n-k}$. Da es genau

$$\binom{n}{k}$$

Möglichkeiten gibt, k–mal a und $(n-k)$–mal b zu wählen, taucht nach vollständigem Ausmultiplizieren der Summand $a^k b^{n-k}$ genau $\binom{n}{k}$ mal auf. □

Zum Abschluss wird noch das Ziehen **mit Zurücklegen** und **ohne Berücksichtigung der Reihenfolge** betrachtet. Hierbei gilt für die Anzahl der möglichen Stichproben:
$$N = \binom{n+k-1}{k}.$$

Beweis: Gesucht ist die Anzahl der Elemente der Menge

$$A = \left\{(x_1, \ldots, x_k) \in \mathbb{N}^k \ : \ 1 \leq x_1 \leq \ldots \leq x_k \leq n\right\}.$$

Durch die Zuordnung

$$(x_1, \ldots, x_k) \mapsto (x_1, x_2 + 1, x_3 + 2, \ldots, x_k + k - 1)$$

wird jedem Element aus A genau ein Element (!) aus der Menge

$$B = \left\{(y_1, \ldots, y_k) \in \mathbb{N}^k \ : \ 1 \leq y_1 < y_2 < \ldots < y_k \leq n + k - 1\right\}$$

zugeordnet.

Beispiel: Für $\Omega = \{1, 2, 3\}$, $n = 3$ und $k = 2$ erhält man die Zuordnungen

$$\begin{array}{rcl}
(1,1) & \mapsto & (1,2) \\
(1,2) & \mapsto & (1,3) \\
(1,3) & \mapsto & (1,4) \\
(2,2) & \mapsto & (2,3) \\
(2,3) & \mapsto & (2,4) \\
(3,3) & \mapsto & (3,4)
\end{array}$$

Um dies formal nachzuweisen, betrachten wir die Abbildung

$$f : A \to B,\ f((x_1, \dots, x_k)) = (x_1, x_2 + 1, x_3 + 2, \dots, x_k + k - 1).$$

Für $(x_1, \dots, x_k) \in A$ gilt $1 \le x_1 \le \dots \le x_k \le n$, was impliziert $1 \le x_1 < x_2 + 1 < x_3 + 2 < \dots < x_k + k - 1 \le n + k - 1$, woraus folgt, dass $f((x_1, \dots, x_k))$ in B liegt. Daher ist die Abbildung f *wohldefiniert*.

Als nächstes zeigen wir, dass sie auch *injektiv* ist. Seien dazu $(x_1, \dots, x_k), (y_1, \dots, y_k) \in A$ gegeben mit

$$f((x_1, \dots, x_k)) = f((y_1, \dots, y_k)).$$

Dies bedeutet

$$(x_1, x_2 + 1, x_3 + 2, \dots, x_k + k - 1) = (y_1, y_2 + 1, y_3 + 2, \dots, y_k + k - 1),$$

woraus folgt $x_1 = y_1$, $x_2 = y_2$, $\dots$, $x_k = y_k$, also

$$(x_1, \dots, x_k) = (y_1, \dots, y_k).$$

Abschließend zeigen wir noch, dass f auch *surjektiv* ist. Dazu wählen wir $(y_1, \dots, y_k) \in B$ beliebig. Dann gilt

$$1 \le y_1 < y_2 < y_3 < \dots < y_k \le n + k - 1,$$

woraus folgt

$$1 \le y_1 \le y_2 - 1 \le y_3 - 2 \le \dots \le y_k - (k - 1) \le n,$$

was bedeutet, dass $(y_1, y_2 - 1, \dots, y_k - (k - 1))$ in A liegt. Wegen

$$f((y_1, y_2 - 1, \dots, y_k - (k - 1))) = (y_1, \dots, y_k)$$

folgt die Surjektivität von f.

Da zwei Mengen, zwischen denen eine bijektive (d.h. injektive und surjektive) Abbildung existiert, immer die gleiche Anzahl an Elementen haben, folgt $N = |A| = |B|$ und mit der oben hergeleiteten Formel für das Ziehen *ohne Zurücklegen* und *ohne Berücksichtigung der Reihenfolge* erhält man:

$$N = |A| = |B| = \binom{n+k-1}{k}.$$

□

Die Ergebnisse dieses Abschnitts sind in Tabelle 4.1 zusammengefasst.

Anzahl Möglichkeiten	Ziehen mit Zurücklegen	Ziehen ohne Zurücklegen
Ziehen mit Berücksichtigung der Reihenfolge	n^k	$\frac{n!}{(n-k)!}$
Ziehen ohne Berücksichtigung der Reihenfolge	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Tabelle 4.1: Grundformeln der Kombinatorik.

Eine weitere Illustration der Nützlichkeit der obigen Formeln erfolgt im nächsten Beispiel. In diesem wird gleichzeitig eine grundlegende Schlussweise der Statistik eingeführt.

Beispiel 4.2 *Die an der Universität Stuttgart im Sommer 2002 abgehaltene schriftliche Prüfung "Statistik II für WiWi" wurde von mehreren Prüfern korrigiert. Dabei bewertete Korrektor K von* 98 *Klausuren* 8 *mit der Note* 5,0*, während Korrektor W von* 102 *Klausuren nur* 1 *mit der Note* 5,0 *benotete. Kann man daraus schließen, dass Korrektor K strenger korrigierte als Korrektor W ?*

Offensichtlich hat Korrektor K prozentual deutlich mehr Klausuren mit der Note 5,0 bewertet als Korrektor W. Es stellt sich jedoch die Frage, ob dieser Unterschied vielleicht nur durch das zufällige Aufteilen der Klausuren auf zwei Korrektoren auftrat.

Um dies zu beantworten, gehen wir zunächst einmal von der Annahme aus, dass beide Korrektoren genau gleich korrigiert haben, und betrachten den Fall, dass $98 + 102 = 200$ Klausuren, von denen $8 + 1 = 9$ mit der Note 5,0 zu bewerten sind, rein zufällig auf diese beiden Korrektoren aufgeteilt werden. Wissen möchten wir, wie groß die Wahrscheinlichkeit ist, dass in diesem Fall der Korrektor, der

98 der Klausuren bekommt, mindestens 8 mit der Note $5,0$ bewertet. Sofern diese Wahrscheinlichkeit sich als klein herausstellen wird (und in der Statistik betrachtet man aus historischen Gründen meist Wahrscheinlichkeiten unter $0,05$ als klein), ist es nicht plausibel, dass wir bei Gültigkeit der obigen Annahme ein solches Resultat beobachten würden. Der übliche statistische Schluss ist dann, die obige Annahme zu verwerfen.

Zur Berechnung der gesuchten Wahrscheinlichkeit betrachten wir das folgende *Urnenmodell*. In einer Urne sind 200 Kugeln, und zwar 9 rote und 191 schwarze Kugeln. Aus diesen werden "rein zufällig" 98 Kugeln gezogen. Wie groß ist dann die Wahrscheinlichkeit, dass unter den 98 gezogenen Kugeln mindestens 8 rote Kugeln sind ?

Wir betrachten das Ziehen ohne Zurücklegen und ohne Beachtung der Reihenfolge. Dies ist auf insgesamt

$$\binom{200}{98}$$

verschiedenen Arten möglich. Da die Reihenfolge hierbei nicht beachtet wird, kann man o.B.d.A. davon ausgehen, dass man zuerst die roten Kugeln und dann erst die schwarzen Kugeln zieht. Um genau 8 rote Kugeln dabei zu erhalten, muss man aus den 9 roten Kugeln 8 ziehen und sodann aus den 191 schwarzen Kugeln 90 ziehen, was auf

$$\binom{9}{8} \cdot \binom{191}{90}$$

verschiedene Arten möglich ist. Analog erhält man, dass Ziehen von genau 9 roten Kugeln auf

$$\binom{9}{9} \cdot \binom{191}{89}$$

vielen Arten möglich ist.

Da jede dieser Kombinationen der Kugeln mit der gleichen Wahrscheinlichkeit auftritt, erhält man für die gesuchte Wahrscheinlichkeit

$$\frac{\text{Anzahl günstiger Fälle}}{\text{Anzahl möglicher Fälle}} = \frac{\binom{9}{8} \cdot \binom{191}{90} + \binom{9}{9} \cdot \binom{191}{89}}{\binom{200}{98}} \approx 0,015,$$

und man kommt zu dem Schluss, dass die Annahme des rein zufälligen Verteilens der Noten $5,0$ auf die beiden Korrektoren bei den aufgetretenden Beobachtungen nicht plausibel ist.

Dennoch kann man hier nicht auf Unterschiede bei den beiden Korrektoren schließen. Vielmehr ist es plausibel, dass die Klausuren keineswegs zufällig aufgeteilt wurden. Die Klausuren wurden nämlich in der Reihenfolge der Abgabe der Studenten eingesammelt, und dann in zwei Teile unterteilt. Dabei ist zu vermuten, dass einer der beiden Korrektoren vor allem Abgaben von den Studenten erhalten hat, die auf die Klausur nur sehr schlecht vorbereitet waren, nur eine der vier Aufgaben bearbeiten konnten und daher die Klausur frühzeitig wieder abgegeben haben.

4.2 Der Begriff des Wahrscheinlichkeitsraumes

Ausgangspunkt der folgenden Betrachtungen ist ein *Zufallsexperiment mit unbestimmten Ergebnis* $\omega \in \Omega$. Zur Illustration dienen die folgenden beiden Beispiele.

Beispiel 4.3 *Ein Spieler zahlt zu Beginn 1.50 Euro. Dann werden vier Münzen geworfen, und zwar zwei 1 Euro Münzen und zwei 50 Cent Münzen, und der Spieler bekommt alle die Münzen, die mit Kopf nach oben landen.*

Wie groß ist die Wahrscheinlichkeit, dass der Wert der Münzen, die der Spieler bekommt, höher ist als der Einsatz von 1.50 Euro ?

Beispiel 4.4 *Student S. fährt immer mit dem Auto zur Uni. Dabei passiert er eine Ampelanlage, bei der sich eine zweiminütige Grünphase mit einer dreiminütigen Rotphase abwechselt.*

Wie groß ist die Wahrscheinlichkeit, dass er an der Ampel länger als eine Minute warten muss, vorausgesetzt seine Ankunft an der Ampel erfolgt rein zufällig innerhalb eines fünfminütigen Intervalls, bestehend aus Grün- und Rotphase ?

Zur mathematischen Modellierung der obigen Zufallsexperimente, wird zuerst einmal die **Menge aller möglichen Ergebnisse** (Beobachtungen) festgelegt.

Definition 4.1 *Die Menge* $\Omega \neq \emptyset$ *aller möglichen Ergebnisse* ω *des Zufallsexperiments heißt* **Grundmenge** *(oder* **Ergebnisraum**, **Ergebnismenge** *oder* **Stichprobenraum***). Die Elemente* $\omega \in \Omega$ *heißen* **Elementarereignisse**.

Für die Wahl des Ergebnisses ω des betrachteten Zufallsexperiments (und damit auch für die Grundmenge Ω) gibt es meistens mehrere verschiedene Möglichkeiten. Z.B. kann man in Beispiel 4.3 den Gewinn (d.h. die Differenz zwischen ausgezahltem Betrag und Einsatz) des Spielers als Ergebnis ω des Zufallsexperimentes wählen. In diesem Fall ist

$$\Omega = \{-1.5, -1, -0.5, 0, 0.5, 1, 1.5\},$$

oder auch eine Obermenge davon, z.B. $\Omega = [-1.5, 1.5]$ oder $\Omega = \mathbb{R}$. Die Modellierung wird aber (wie wir später sehen werden) deutlich einfacher, wenn man als Ergebnis des Zufallsexperiments die Lage der vier Münzen nach dem Werfen wählt. In diesem Fall ist

$$\omega = (\omega_1, \omega_2, \omega_3, \omega_4)$$

mit $\omega_i \in \{K, Z\}$. Dabei seien die Münzen von 1 bis 4 durchnummeriert, die Münzen 1 und 2 haben den Wert 1 Euro, die Münzen 3 und 4 den Wert 50 Cent, und $\omega_i = K$ (bzw. $\omega_i = Z$) bedeutet, dass die i-te Münze mit Kopf (bzw. Zahl) nach oben landet. Die Grundmenge ist dann

$$\Omega = \{(\omega_1, \omega_2, \omega_3, \omega_4) \ : \ \omega_i \in \{K, Z\}\}$$

Auch in Beispiel 4.4 gibt es mehrere Möglichkeiten für die Wahl des Ergebnisses des Zufallsexperiments. Betrachtet man die Wartezeit an der Ampel als ω, so ist die Grundmenge gegeben durch

$$\Omega = [0, 3]$$

(bzw. durch eine Obermenge davon, z.B. $\Omega = \mathbb{R}_+$). Wie wir später sehen werden, wird die Berechnung der gesuchten Wahrscheinlichkeit aber einfacher, wenn man den Eintreffzeitpunkt in Minuten relativ zu Beginn der letzten Rotphase als ω wählt. In diesem Fall ist

$$\Omega = [0, 5]$$

(bzw. eine Obermenge davon).

Gesucht ist in beiden Beispielen nach der Wahrscheinlichkeit, dass das Ergebnis ω des Zufallsexperimentes in einer Menge $A \subseteq \Omega$ zu liegen kommt.

Definition 4.2 *Teilmengen A der Grundmenge Ω heißen* **Ereignisse***. Ein Ereigniss* **tritt ein***, falls das Ergebnis ω des Zufallsexperiments in A liegt.*

Wählt man in Beispiel 4.3 den Gewinn des Spielers als Ergebnis ω des Zufallsexperiments (und dann z.B. $\Omega = \{-1.5, -1, -0.5, 0, 0.5, 1, 1.5\}$), so ist dort nach der Wahrscheinlichkeit gefragt, dass ω in

$$A = \{0.5, 1, 1.5\}$$

zu liegen kommt. Wählt man dagegen

$$\Omega = \{(\omega_1, \omega_2, \omega_3, \omega_4) \ : \ \omega_i \in \{K, Z\}\},$$

d.h., ist die Lage der Münzen das Ergebnis des Zufallsexperimentes, so ist A die Menge aller der $(\omega_1, \omega_2, \omega_3, \omega_4)$, bei denen der Wert der Münzen mit Kopf oben größer als 1.50 Euro ist. Diese Menge lässt sich am einfachsten durch Betrachtung aller Möglichkeiten für die Lage der Münzen bestimmen.

Zur Bestimmung von A betrachten wir alle 16 Elemente von Ω und bestimmen jeweils den Wert der Münzen mit Kopf oben.

ω_1 1 Euro	ω_2 1 Euro	ω_3 50 Cent	ω_4 50 Cent	Wert der Münzen mit Kopf oben
K	K	K	K	3
K	K	K	Z	2.5
K	K	Z	K	2.5
K	K	Z	Z	2
K	Z	K	K	2
K	Z	K	Z	1.5
K	Z	Z	K	1.5
K	Z	Z	Z	1
Z	K	K	K	2
Z	K	K	Z	1.5
Z	K	Z	K	1.5
Z	K	Z	Z	1
Z	Z	K	K	1
Z	Z	K	Z	0.5
Z	Z	Z	K	0.5
Z	Z	Z	Z	0

Aus der obigen Tabelle liest man ab:

$$\begin{aligned} A &= \big\{(Z,K,K,K),(K,Z,K,K),(K,K,Z,Z),(K,K,Z,K), \\ &\quad (K,K,K,Z),(K,K,K,K)\big\}. \end{aligned}$$

Als nächstes betrachten wir nochmals Beispiel 4.4. Betrachtet man hier den Eintreffzeitpunkt in Minuten relativ zu Beginn der letzten Rotphase als Ergebnis des Zufallsexperiments (und setzt $\Omega = [0,5]$), so ist die Wartezeit an der Ampel genau dann länger als eine Minute, wenn man weniger als zwei Minuten nach Beginn der letzten Rotphase an der Ampel eintrifft. In diesem Fall ist also nach der Wahrscheinlichkeit gefragt, dass ω in

$$A = [0,2)$$

zu liegen kommt.

Im Folgenden wollen wir nun Teilmengen A der Grundmenge Ω Wahrscheinlichkeiten, d.h. Zahlen aus dem Intervall $[0,1]$, zuweisen. Die intuitive Bedeutung dieser Wahrscheinlichkeiten ist wie folgt: Führt man das Zufallsexperiment viele Male unbeeinflusst voneinander hintereinander durch, so soll die relative Anzahl des Eintretens von A (d.h., des Auftretens eines Ergebnisses ω, welches in A liegt) ungefähr gleich $\mathbf{P}(A)$ sein.

Hier gibt es zuerst einmal eine naive Möglichkeit für die Festlegung der Wahrscheinlichkeiten. Dabei legt man für jedes $\tilde{\omega} \in \Omega$ die Wahrscheinlichkeit $P(\{\tilde{\omega}\})$, dass das Ergebnis des Zufallsexperiments gerade gleich $\tilde{\omega}$ ist, fest, und setzt dann

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

d.h., die Wahrscheinlichkeit, dass A eintritt ist gleich der Summe der Wahrscheinlichkeiten aller Elemente in A.

Dies ist problemlos möglich in Beispiel 4.3. Wählt man hier

$$\Omega = \{(\omega_1, \omega_2, \omega_3, \omega_4) \ : \ \omega_i \in \{K, Z\}\},$$

so ist

$$\begin{aligned} A \quad = \quad & \{(Z,K,K,K),(K,Z,K,K),(K,K,Z,Z),(K,K,Z,K), \\ & (K,K,K,Z),(K,K,K,K)\}. \end{aligned}$$

Jedes Element ω von Ω tritt dann mit gleicher Wahrscheinlichkeit

$$\mathbf{P}(\{\omega\}) = \frac{1}{|\Omega|} = \frac{1}{16}$$

auf. Die Wahrscheinlichkeit, dass ein ω in $A \subseteq \Omega$ auftritt, ist dann

$$\begin{aligned} \mathbf{P}(A) \quad &= \quad \sum_{\omega \in A} \mathbf{P}(\{\omega\}) = \sum_{\omega \in A} \frac{1}{|\Omega|} = \frac{|A|}{|\Omega|} \\ &= \quad \frac{\text{“Anzahl der für } A \text{ günstigen Fälle”}}{\text{“Anzahl der möglichen Fälle”}}. \end{aligned}$$

Mit $|A| = 6$ berechnet sich die gesuchte Wahrscheinlichkeit zu

$$\mathbf{P}(A) = 6/16 = 3/8.$$

Dieser Zugang ist in Beispiel 4.4 aber nicht möglich. Betrachtet man hier den Eintreffzeitpunkt in Minuten relativ zu Beginn der letzten Rotphase als Ergebnis des Zufallsexperiments (und setzt $\Omega = [0, 5]$), so ist die Wahrscheinlichkeit $\mathbf{P}(\{\omega\})$, genau ω Minuten nach der letzten Rotphase einzutreffen, für alle $\omega \in [0, 5]$ gleich

Null. Denn diese ist sicherlich nicht größer als die Wahrscheinlichkeit, dass der Eintreffzeitpunkt im Intervall $[\omega - \epsilon, \omega + \epsilon]$ liegt ($\epsilon > 0$ beliebig), und da letztere proportional zur Intervalllänge ist, liegt sie für ϵ klein beliebig nahe bei Null.

Als alternativen Zugang in Beispiel 4.4 bietet sich an, die Wahrscheinlichkeit für das Eintreffen innerhalb eines Intervalls $[a, b) \subseteq [0, 5)$ proportional zur Intervalllänge zu wählen. Genauer setzt man

$$\mathbf{P}([a,b)) = \frac{\text{Länge von } [a,b)}{\text{Länge von } [0,5)} = \frac{b-a}{5},$$

und erhält die gesuchte Wahrscheinlichkeit zu

$$\mathbf{P}([0,2)) = \frac{2}{5} = 0,4.$$

Nachteil der obigen Ansätze ist, dass sie ziemlich unsystematisch sind. Insbesondere werden hier die beiden Beispiele auf verschiedene Arten gelöst. Möchte man nun gewisse theoretische Aussagen über die zugrunde liegenden stochastischen Strukturen herleiten, so muss man dies für beide Fälle separat machen. Um dies zu vermeiden, verallgemeinern wir beide Fälle im Folgenden. Dabei fordern wir, motiviert von Eigenschaften relativer Häufigkeiten, dass bei der Zuweisung von Wahrscheinlichkeiten zu Mengen gewisse Eigenschaften vorliegen sollen. Anschließend werden wir separat untersuchen, wie man Abbildungen konstruieren kann, die diese Eigenschaften besitzen, und welche Schlussfolgerungen man hinsichtlich des Ausgangs von Zufallsexperimenten, die durch solche Abbildungen beschrieben werden, ziehen kann.

Ziel im Folgenden ist die Festlegung von Eigenschaften, die die Zuweisung von Wahrscheinlichkeiten (d.h. Zahlen aus dem Intervall $[0, 1]$) zu Teilmengen der Grundmenge Ω, haben soll. Diese Zuweisung kann zusammengefasst werden zu einer Abbildung

$$\mathbf{P} : \mathcal{P}(\Omega) \to [0,1].$$

Hierbei ist $\mathcal{P}(\Omega) = \{A | A \subseteq \Omega\}$ die sogenannte *Potenzmenge* von Ω, d.h., die Menge aller Teilmengen von Ω. $\mathbf{P}$ weist jeder Menge $A \subseteq \Omega$ eine Zahl $\mathbf{P}(A) \in [0, 1]$ zu.

Da das Ergebnis unseres Zufallsexperiments niemals in der leeren Menge $\emptyset$ sowie immer in der Grundmenge Ω zu liegen kommt, ist eine naheligende Forderung an $\mathbf{P}$:

$$\mathbf{P}(\emptyset) = 0 \quad \text{und} \quad \mathbf{P}(\Omega) = 1.$$

Ist außerdem A eine beliebige Teilmenge von Ω und $A^c = \Omega \setminus A$ das sogenannte *Komplement* von A bestehend aus allen Elementen von Ω, die nicht in A enthalten sind, so liegt das Ergebnis des Zufallsexperiments genau dann in A^c, wenn es nicht in A liegt. Dies legt die Forderung

$$\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$$

nahe. Sind darüberhinaus A und B zwei *disjunkte* Teilmengen von Ω, d.h. zwei Teilmengen von Ω mit $A \cap B = \emptyset$, so liegt das Ergebnis des Zufallsexperiments genau dann in $A \cup B$, wenn es entweder in A oder in B liegt. Dies motiviert die Forderung

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) \quad \text{falls } A \cap B = \emptyset.$$

Durch wiederholtes Anwenden folgt daraus

$$\begin{aligned} \mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) &= \mathbf{P}(A_1) + \mathbf{P}(A_2 \cup \cdots \cup A_n) \\ &= \ldots \\ &= \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots + \mathbf{P}(A_n) \end{aligned}$$

für *paarweise disjunkte* Mengen $A_1, \ldots, A_n \subseteq \Omega$, d.h. für Mengen mit $A_i \cap A_j = \emptyset$ für alle $i \neq j$. Hinsichtlich der Herleitung von theoretischen Aussagen wird es sich als sehr günstig erweisen, dies auch für Vereinigungen von abzählbar vielen paarweise disjunkten Mengen zu fordern:

$$\mathbf{P}\left(\cup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(A_n) \quad \text{für } A_n \subseteq \Omega \text{ mit } A_i \cap A_j = \emptyset \text{ für } i \neq j.$$

Dies führt auf

Definition 4.3 *(Vorläufige Definition des Wahrscheinlichkeitsmaßes).*

Sei Ω eine nichtleere Menge. Eine Abbildung

$$\mathbf{P} : \mathcal{P}(\Omega) \to [0, 1]$$

heißt **Wahrscheinlichkeitsmaß** *(kurz: W–Maß), falls gilt:*

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

(ii) Für alle $A \subseteq \Omega$:

$$\mathbf{P}(A^c) = 1 - \mathbf{P}(A).$$

(iii) Für alle $A, B \subseteq \Omega$ mit $A \cap B = \emptyset$:

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

(iv) Für alle $A_1, A_2, \dots \subseteq \Omega$ mit $A_i \cap A_j = \emptyset$ für $i \neq j$:

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)$$

(sog. σ–Additivität).

In diesem Falle heißt $(\Omega, \mathcal{P}(\Omega), \mathbf{P})$ **Wahrscheinlichkeitsraum** *(kurz: W–Raum),* $\mathbf{P}(A)$ **Wahrscheinlichkeit** *des Ereignisses $A \subseteq \Omega$.*

Die hier geforderten Eigenschaften sind z.B. im Falle

$$\Omega = \{(\omega_1, \omega_2, \omega_3, \omega_4) \,:\, \omega_i \in \{K, Z\}\}$$

für

$$\mathbf{P} : \mathcal{P}(\Omega) \to [0,1], \quad \mathbf{P}(A) = \frac{|A|}{|\Omega|}$$

erfüllt (vergleiche Beispiel 4.3 und Satz 4.1).

Will man jedoch auch für Beispiel 4.4 einen Wahrscheinlichkeitsraum (mit den Eigenschaften aus der obiger Definition) konstruieren, so stößt man auf das folgende technische Problem: Man kann zeigen, dass keine Abbildung $\mathbf{P} : \mathcal{P}([0,5]) \to [0,1]$ existiert, für die einerseits

$$\mathbf{P}([a,b)) = \frac{b-a}{5} \quad \text{für alle } 0 \leq a < b \leq 5$$

gilt, und die andererseits ein W-Maß ist, d.h. für die die Eigenschaften (i) bis (iv) aus der obigen Definition erfüllt sind.

Um dieses Problem zu umgehen, legt man in solchen Beispielen nicht die Wahrscheinlichkeiten für *alle* Teilmengen von Ω fest, sondern nur für einen möglichst "großen" Teil dieser Mengen. Ohne Probleme kann man die Wahrscheinlichkeiten für die Mengen $\emptyset$ und Ω festlegen. Die leere Menge $\emptyset$ beschreibt das sogenannte unmögliche Ereignis, welches nie eintritt, und dem man daher die Wahrscheinlichkeit Null zuweisen kann. Die gesamte Grundmenge Ω steht für das Ereignis, das immer eintritt, und dem man die Wahrscheinlichkeit Eins zuordnen kann. Außerdem sollte es nach Festlegung der Wahrscheinlichkeiten zweier Ereignisse A und B auch möglich sein, die Wahrscheinlichkeit, dass A oder B (oder beide) eintreten, d.h., dass ein $\omega \in A \cup B$ eintritt, sowie die Wahrscheinlichkeit, dass A und B eintreten, d.h., dass ein $\omega \in A \cap B$ eintritt, und die Wahrscheinlichkeit, dass A nicht eintritt, d.h., dass ein ein $\omega \in A^c = \Omega \setminus A$ eintritt, festzulegen. Hierbei heißt A^c das *komplementäre Ereignis zu A.*

Dies motiviert, dass die Menge aller Ereignisse, für die man die Wahrscheinlichkeiten festlegt, zumindest $\emptyset$ und Ω enthalten sollte, sowie mit zwei Ereignissen A und B auch $A \cup B$, $A \cap B$ und A^c enthalten sollte. Aus technischen Gründen (hinsichtlich asymptotischen Aussagen) ist es darüberhinaus auch sinnvoll zu fordern, dass die sukzessive Anwendung von abzählbar vielen Mengenoperationen wie Vereinigung, Schnitt und Komplementbildung, auf solche Mengen wieder eine Menge ergibt, für die man die Wahrscheinlichkeit festlegen kann. Dies führt auf den Begriff der sogenannten σ-Algebra:

Definition 4.4 *Sei Ω eine nichtleere Menge. Eine Menge $\mathcal{A}$ von Teilmengen von Ω heißt σ–**Algebra** (über Ω), falls gilt:*

(i) $\emptyset \in \mathcal{A}$ und $\Omega \in \mathcal{A}$.

(ii) Aus $A \in \mathcal{A}$ folgt $A^c := \Omega \setminus A \in \mathcal{A}$.

(iii) Aus $A, B \in \mathcal{A}$ folgt $A \cup B \in \mathcal{A}$, $A \cap B \in \mathcal{A}$ und $A \setminus B \in \mathcal{A}$.

(iv) Sind $A_1, A_2, \dots \in \mathcal{A}$, so ist auch $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$ und $\cap_{n=1}^{\infty} A_n \in \mathcal{A}$.

Eine σ–Algebra ist also eine Menge von Teilmengen von Ω, die $\emptyset$ und Ω enthält, und bei der man bei Anwendung von endlich oder abzählbar unendlich vielen der üblichen Mengenoperationen auf Mengen aus der σ-Algebra immer wieder eine Menge erhält, die in der σ-Algebra enthalten ist.

Beispiele:

a) Sei $\Omega \neq \emptyset$ beliebig. Dann sind $\{\emptyset, \Omega\}$ und $\mathcal{P}(\Omega)$ σ-Algebren über Ω.

b) Wir betrachten das Werfen eines Würfels. Als Augenzahl kann dabei eine der Zahlen 1, …, 6 auftreten, so dass man $\Omega = \{1, 2, 3, 4, 5, 6\}$ setzt. Als σ–Algebren kommen dann Teilmengen der Potenzmenge von Ω in Frage, d.h., Mengen, deren Elemente wieder Mengen sind und zwar Teilmengen von Ω. Hier ist $\mathcal{A} = \{\emptyset, \{1\}, \Omega\}$ keine σ–Algebra über Ω, da

$$\{1\} \in \mathcal{A} \quad \text{aber } \{1\}^c = \{2, 3, 4, 5, 6\} \notin \mathcal{A}.$$

Wie man leicht sieht, ist aber $\mathcal{A} = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$ eine σ–Algebra über Ω.

Ist die Grundmenge wie im hier vorliegenden Fall endlich oder abzählbar unendlich, so wird in Anwendungen immer die σ–Algebra $\mathcal{A} = \mathcal{P}(\Omega)$ verwendet.

c) Als nächstes betrachten wir die stochastische Modellierung der Lebensdauer einer Glühbirne. Hier tritt als Resultat des Zufallsexperiments eine Zahl $t \geq$

0 (z.B. Lebensdauer in Sekunden) auf. Der Einfachheit halber wählen wir als Grundmenge sogar die etwas zu große Menge $\Omega = \mathbb{R}$.

Es stellt sich dann die Frage, was eine sinnvolle Wahl für die σ–Algebra über $\mathbb{R}$ ist. $\mathcal{A} = \mathcal{P}(\mathbb{R})$ ist zwar eine σ–Algebra über $\mathbb{R}$, sie ist aber für die Festlegung von Wahrscheinlichkeiten (siehe oben) meist zu groß.

Statt dessen verwendet man:

$\mathcal{A}$= kleinste σ–Algebra, die alle Intervalle der Form $(a,b] := \{x \ : \ a < x \leq b\}$ $(a,b \in \mathbb{R})$ enthält.

Formal kann man diese kleinste σ–Algebra definieren als Menge bestehend aus allen denjenigen Teilmengen von $\mathbb{R}$, die die Eigenschaft haben, dass sie in allen σ–Algebren, die alle Intervalle der Form $(a,b]$ $(a,b \in \mathbb{R})$ enthalten, enthalten sind. Nach Definition sind Mengen aus dieser σ–Algebra in jeder σ–Algebra enthalten, die alle Intervalle der Form $(a,b]$ $(a,b \in \mathbb{R})$ enthält. Darüberhinaus kann man leicht zeigen, dass es sich bei dieser Menge von Mengen um eine σ–Algebra handelt (z.B. enthält sie die leere Menge, da diese ja nach Definition in jeder der σ–Algebren, die alle Intervalle enthalten, enthalten ist).

Man bezeichnet diese σ–Algebra als **Borelsche σ–Algebra** über $\mathbb{R}$ und verwendet dafür häufig die Abkürzung $\mathcal{B}$. Man kann zeigen, dass sie alle in der Praxis vorkommenden Teilmengen von $\mathbb{R}$ (wie z.B. Einpunktmengen, abzählbare Mengen, Intervalle, offene Mengen, abgeschlossene Mengen, ...) enthält.

Wir erweitern nun den Begriff des Wahrscheinlichkeitsraums aus Definition 4.3, indem wir die Wahrscheinlichkeiten nicht mehr für alle Teilmengen von Ω festlegen, sondern nur für diejenigen, die in einer vorgegebenen σ-Algebra enthalten sind.

Definition 4.5 *(Endgültige Definition des Wahrscheinlichkeitsmaßes).*

Sei Ω eine nichtleere Menge und $\mathcal{A}$ eine σ-Algebra über Ω. Eine Abbildung

$$\mathbf{P} : \mathcal{A} \to [0,1]$$

heißt **Wahrscheinlichkeitsmaß** *(kurz: W–Maß), falls gilt:*

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

(ii) Für alle $A \in \mathcal{A}$:

$$\mathbf{P}(A^c) = 1 - \mathbf{P}(A).$$

(iii) Für alle $A, B \in \mathcal{A}$ mit $A \cap B = \emptyset$:

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

(iv) Für alle $A_1, A_2, \dots \in \mathcal{A}$ mit $A_i \cap A_j = \emptyset$ für $i \neq j$:

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)$$

(sog. σ–Additivität).

In diesem Falle heißt $(\Omega, \mathcal{A}, \mathbf{P})$ **Wahrscheinlichkeitsraum** *(kurz: W-Raum),* $\mathbf{P}(A)$ **Wahrscheinlichkeit** *des Ereignisses $A \in \mathcal{A}$.*

Für die Wahl der σ-Algebra ist es im Falle einer endlichen oder abzählbar unendlichen Grundmenge Ω üblich, $\mathcal{A} = \mathcal{P}(\Omega)$ zu setzen. Im Falle von $\Omega = \mathbb{R}$ wählt man meistens $\mathcal{A} = \mathcal{B}$, d.h., man wählt die oben eingeführte Borelsche σ-Algebra. Dies hat den Vorteil, dass man z.B. ein W-Maß $\mathbf{P} : \mathcal{B} \to [0, 1]$ konstruieren kann mit

$$\mathbf{P}([a, b)) = \frac{b-a}{5} \quad \text{für alle } 0 \leq a < b \leq 5.$$

Dieses kann dann zur Beschreibung der Situation in Beispiel 4.4 verwendet werden.

Zum Nachweis, dass eine Abbildung $\mathbf{P} : \mathcal{A} \to \mathbb{R}$ ein W-Maß ist, muss man nicht alle Forderungen aus Definition 4.5 nachrechnen. Es gilt nämlich

Lemma 4.1 *Sei Ω eine nichtleere Menge und $\mathcal{A}$ eine σ-Algebra über Ω. Dann ist eine Abbildung*

$$\mathbf{P} : \mathcal{A} \to \mathbb{R}$$

genau dann ein W-Maß, wenn sie die drei folgenden Eigenschaften hat:

1. $\mathbf{P}(A) \geq 0$ *für alle $A \in \mathcal{A}$.*
2. $\mathbf{P}(\Omega) = 1$.
3. *Für alle $A_1, A_2, \dots \in \mathcal{A}$ mit $A_i \cap A_j = \emptyset$ für $i \neq j$ gilt*

$$\mathbf{P}\left(\cup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(A_n).$$

Beweis. Es ist klar, dass ein W-Maß die Eigenschaften 1. bis 3. aus Lemma 4.1 hat. Also genügt es im Folgenden zu zeigen, dass bei Gültigkeit von 1. bis 3. die Bedingungen (i) bis (iv) aus Definition 4.5 sowie $\mathbf{P}(A) \leq 1$ für alle $A \in \mathcal{A}$ erfüllt sind.

Aus 3. folgt

$$\mathbf{P}(\emptyset) = \mathbf{P}(\emptyset \cup \emptyset \cup \emptyset \cup \dots) = \mathbf{P}(\emptyset) + \mathbf{P}(\emptyset) + \mathbf{P}(\emptyset) + \dots$$

Mit $\mathbf{P}(\emptyset) \in \mathbb{R}$ folgt daraus $\mathbf{P}(\emptyset) = 0$.

Damit folgt unter erneuter Verwendung von 3., dass für $A, B \in \mathcal{A}$ mit $A \cap B = \emptyset$ gilt:

$\mathbf{P}(A \cup B) = \mathbf{P}(A \cup B \cup \emptyset \cup \emptyset \cup \dots) = \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(\emptyset) + \mathbf{P}(\emptyset) + \dots = \mathbf{P}(A) + \mathbf{P}(B) + 0 + 0 + \dots = \mathbf{P}(A) + \mathbf{P}(B)$.

Mit $A \cup A^c = \Omega$, $A \cap A^c = \emptyset$ und 2. folgt weiter

$$\mathbf{P}(A) + \mathbf{P}(A^c) = \mathbf{P}(A \cup A^c) = \mathbf{P}(\Omega) = 1,$$

also gilt für $A \in \mathcal{A}$: $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$. Letzteres impliziert insbesondere

$$\mathbf{P}(A) = 1 - \mathbf{P}(A^c) \leq 1 - 0 = 1.$$

□

Einige weitere nützliche Eigenschaften von W-Maßen sind zusammengefasst in

Lemma 4.2 *Sei* $(\Omega, \mathcal{A}, \mathbf{P})$ *ein W-Raum.*

a) *Sind* $A, B \in \mathcal{A}$ *mit* $A \subseteq B$, *so gilt:*

$$\mathbf{P}(A) \leq \mathbf{P}(B) \quad \text{und} \quad \mathbf{P}(B \setminus A) = \mathbf{P}(B) - \mathbf{P}(A).$$

b) *Sind* $A_1, A_2, \dots \in \mathcal{A}$ *so gilt für jedes* $n \in \mathbb{N}$

$$\mathbf{P}\left(\cup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbf{P}(A_i)$$

sowie

$$\mathbf{P}\left(\cup_{i=1}^\infty A_i\right) \leq \sum_{i=1}^\infty \mathbf{P}(A_i)$$

c) *Sind* $A, B \in \mathcal{A}$, *so gilt*

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B).$$

d) *Sind $A_1, \ldots, A_n \in \mathcal{A}$, so gilt*

$$\begin{aligned}
&\mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \\
&= \sum_{i=1}^{n} \mathbf{P}(A_i) - \sum_{1 \le i < j \le n} \mathbf{P}(A_i \cap A_j) + \sum_{1 \le i < j < k \le n} \mathbf{P}(A_i \cap A_j \cap A_k) - + \ldots \\
&\quad + (-1)^{n-1} \mathbf{P}(A_1 \cap A_2 \cap \cdots \cap A_n).
\end{aligned}$$

Beweis: a) Aus $A \subseteq B$ folgt $B = (B \setminus A) \cup A$, wobei die beiden Mengen auf der rechten Seite leeren Schnitt haben. Dies impliziert

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

bzw. $0 \le \mathbf{P}(B \setminus A) = \mathbf{P}(B) - \mathbf{P}(A)$.

b) Für $A, B \in \mathcal{A}$ gilt

$$\mathbf{P}(A \cup B) = \mathbf{P}(A \cup (B \setminus A)) = \mathbf{P}(A) + \mathbf{P}(B \setminus A) \le \mathbf{P}(A) + \mathbf{P}(B),$$

wobei die letzte Ungleichung aus a) folgt. Mit Induktion ergibt sich der erste Teil von b).

Für den zweiten Teil von b) schließt man analog:

$$\begin{aligned}
\mathbf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) &= \mathbf{P}\left(A_1 \cup \bigcup_{i=2}^{\infty} A_i \setminus (A_1 \cup \cdots \cup A_{i-1})\right) \\
&= \mathbf{P}(A_1) + \sum_{i=2}^{\infty} \mathbf{P}\left(A_i \setminus (A_1 \cup \cdots \cup A_{i-1})\right) \\
&\le \sum_{i=1}^{\infty} \mathbf{P}(A_i).
\end{aligned}$$

c) folgt aus

$$\begin{aligned}
&\mathbf{P}(A \cup B) \\
&= \mathbf{P}\left((A \setminus (A \cap B)) \cup (B \setminus (A \cap B)) \cup (A \cap B)\right) \\
&= \mathbf{P}(A \setminus (A \cap B)) + \mathbf{P}(B \setminus (A \cap B)) + \mathbf{P}(A \cap B) \\
&\overset{a)}{=} \mathbf{P}(A) - \mathbf{P}(A \cap B) + \mathbf{P}(B) - \mathbf{P}(B \cap A) + \mathbf{P}(A \cap B) \\
&= \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B).
\end{aligned}$$

Mit (schreibtechnisch etwas aufwendiger) Induktion folgt d) aus c). □

Lemma 4.3 *(Erstes Lemma von Borel und Cantelli).*

Sei $(\Omega, \mathcal{A}, \mathbf{P})$ *ein W-Raum und sei* $(A_n)_n$ *eine Folge von Ereignissen mit*

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty.$$

Dann gilt

$$\mathbf{P}\left(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k\right) = 0.$$

Beweis. Für beliebiges $N \in \mathbb{N}$ gilt

$$\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k \subseteq \cup_{k=N}^{\infty} A_k,$$

woraus folgt

$$\mathbf{P}\left(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k\right) \leq \mathbf{P}\left(\cup_{k=N}^{\infty} A_k\right) \leq \sum_{k=N}^{\infty} \mathbf{P}\left(A_k\right) \to 0 \quad (N \to \infty),$$

da $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$. □

4.3 Konstruktion von W–Räumen

4.3.1 Laplacesche W–Räume

Als nächstes betrachteten wir Zufallsexperimente, bei denen zum einen nur endlich viele Werte auftreten, und bei denen zum anderen jeder einzelne Wert mit der gleichen Wahrscheinlichkeit auftritt. Solche Zufallsexperimente modelliert man durch die im nächsten Satz beschriebenen Laplaceschen Wahrscheinlichkeitsräume.

Satz 4.1 *Sei* Ω *eine (nichtleere) endliche Menge,* $\mathcal{A} = \mathcal{P}(\Omega)$ *und* $\mathbf{P} : \mathcal{A} \to [0, 1]$ *definiert durch*

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|} \quad (A \in \mathcal{A}).$$

Dann ist $(\Omega, \mathcal{A}, \mathbf{P})$ *ein W–Raum. In diesem gilt*

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

für alle $\omega \in \Omega$.

Beweis. Offensichtlich ist Ω eine nichtleere Menge und $\mathcal{A}$ eine σ–Algebra über Ω, also genügt es zu zeigen, dass $\mathbf{P} : \mathcal{A} \to [0,1]$ ein W–Maß ist. Es gilt $\mathbf{P}(A) \geq 0$ für alle $A \subseteq \Omega$ und

$$\mathbf{P}(\Omega) = \frac{|\Omega|}{|\Omega|} = 1.$$

Da darüberhinaus die Anzahl der Elemente einer Vereinigung von nicht überlappenden Mengen gleich der Summe der Anzahlen der Elemente in den einzelnen Mengen ist, ist $\mathbf{P}$ auch σ–additiv. Mit Lemma 4.1 folgt daraus die Behauptung. □

Definition 4.6 *Der W–Raum aus Satz 4.1 heißt* **Laplacescher W–Raum**.

Bemerkung. In einem Laplaceschen W–Raum gilt für beliebiges $A \subseteq \Omega$:

$$\mathbf{P}(A) \quad = \quad \frac{|A|}{|\Omega|} = \frac{\text{“Anzahl der für } A \text{ günstigen Fälle”}}{\text{“Anzahl der möglichen Fälle”}}.$$

Im Folgenden werden drei (einfache) Beispiele für Laplacesche W–Räume betrachtet.

Beispiel 4.5 *Viermaliges Werfen einer “echten” Münze.*

Dies läßt sich beschreiben durch einen Laplaceschen W–Raum mit Grundmenge

$$\Omega = \{(\omega_1, \dots, \omega_4) \quad : \quad \omega_i \in \{0,1\} \quad (i = 1, \dots, 4)\}.$$

Hierbei steht $\omega_i = 0$ für “i-te Münze landet mit Kopf nach oben” und $\omega_i = 1$ für “i-te Münze landet mit Zahl nach oben”. Da hierbei jeder Wert $(\omega_1, \dots, \omega_4)$ mit der gleichen Wahrscheinlichkeit $1/|\Omega|$ auftritt, verwendet man zur stochastischen Modellierung einen Laplaceschen W-Raum, d.h. man setzt $\mathcal{A} = \mathcal{P}(\Omega)$ und

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{2^4} \quad (A \subseteq \Omega).$$

Sei A das Ereignis, dass mindestens einmal Kopf auftritt. Dann gilt:

$$\mathbf{P}(A) = 1 - \mathbf{P}(A^c) = 1 - \mathbf{P}(\{(1,1,1,1)\}) = 1 - \frac{1}{2^4} = \frac{15}{16}.$$

Beispiel 4.6 *In einer Fernsehshow wird folgendes Glücksspiel angeboten: Versteckt hinter drei Türen befinden sich ein Auto und zwei Ziegen. Im ersten Schritt deutet der Spieler (in zufälliger Weise) auf eine der drei Türen, die aber geschlossen bleibt. Dann öffnet der Spielleiter eine der beiden anderen Türen, hinter der sich eine Ziege befindet. Im zweiten Schritt wählt der Spieler eine der beiden noch geschlossenen Türen. Befindet sich dahinter das Auto, so hat er dieses gewonnen.*

Im Folgenden soll die Wahrscheinlichkeit für den Spieler, das Auto zu gewinnen, bestimmt werden, wenn er im zweiten Schritt

a) seine im ersten Schritt getroffene Wahl beibehält,

b) seine im ersten Schritt getroffene Wahl aufgibt und die andere geschlossene Türe wählt.

Dazu werden die Türen von 1 bis 3 durchnummeriert. Der Einfachheit halber wird davon ausgegangen, dass der Spielleiter die Tür mit dem kleineren Index öffnet, sofern er zwei Möglichkeiten zum Öffnen hat.

Zur Bestimmung der beiden Wahrscheinlichkeiten wird das obige Zufallsexperiment beschrieben durch einen W–Raum mit Grundmenge

$$\Omega = \{(\omega_1, \omega_2) \ : \ \omega_1, \omega_2 \in \{1, 2, 3\}\}.$$

Hierbei ist ω_1 die Nummer der Tür, hinter der sich das Auto befindet, und ω_2 die Nummer der Tür, auf die der Spieler tippt. Da jeder Wert (ω_1, ω_2) mit der gleichen Wahrscheinlichkeit $1/|\Omega|$ auftritt, wird zur stochastischen Modellierung wieder ein Laplacescher W–Raum verwendet, d.h. es wird gesetzt

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

und

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{9} \quad \text{für } A \in \mathcal{A}.$$

Seien nun A bzw. B die Ereignisse, dass der Spieler bei Strategie a) bzw. b) das Auto gewinnt. Zur Bestimmung von $|A|$ bzw. $|B|$ betrachtet man alle 9 Elemente von Ω und bestimmt jeweils, ob der Spieler das Auto bei Strategie a) bzw. b) gewinnt oder nicht:

ω_1	ω_2	Spielleiter öffnet	Spieler tippt bei a) auf	Gewinn bei a)	Spieler tippt bei b) auf	Gewinn bei b)
1	1	2	1	Ja	3	Nein
1	2	3	2	Nein	1	Ja
1	3	2	3	Nein	1	Ja
2	1	3	1	Nein	2	Ja
2	2	1	2	Ja	3	Nein
2	3	1	3	Nein	2	Ja
3	1	2	1	Nein	3	Ja
3	2	1	2	Nein	3	Ja
3	3	1	3	Ja	2	Nein

Aus der Tabelle liest man ab:

$$A = \{(1,1),(2,2),(3,3)\} \text{ und } B = \{(1,2),(1,3),(2,1),(2,3),(3,1),(3,2)\}$$

und damit erhält man

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|} = \frac{3}{9} = \frac{1}{3} \text{ und } \mathbf{P}(B) = \frac{|B|}{|\Omega|} = \frac{6}{9} = \frac{2}{3}.$$

Beispiel 4.7 *In einer Stadt mit m Längs– und n Querstraßen sollen k Verkehrspolizisten ($k \leq \min\{m,n\}$) auf die $m \cdot n$ Straßenkreuzungen aufgeteilt werden. Aufgrund des Ausbildungsstandes der Polizisten ist klar, dass eine Kreuzung von höchstens einem Polizisten gesichert wird. Wie groß ist bei rein zufälliger Verteilung der Polizisten auf die Kreuzungen die Wahrscheinlichkeit, dass auf jeder Straße höchstens ein Polizist steht ?*

1. Lösung: *Reihenfolge bei der Auswahl der Kreuzungen wird beachtet.*

Anzahl möglicher Fälle:

$$(m \cdot n) \cdot (m \cdot n - 1) \cdot \ldots \cdot (m \cdot n - k + 1)$$

(Aus $m \cdot n$ Kreuzungen k auswählen ohne Zurücklegen und mit Beachten der Reihenfolge.)

Anzahl günstiger Fälle:

$$m \cdot n \cdot (m-1) \cdot (n-1) \cdot \ldots \cdot (m-k+1) \cdot (n-k+1)$$

(Zweite Kreuzung darf nicht in der gleichen Längs– oder Querstraße liegen wie 1. Kreuzung, etc.)

Damit ist die gesuchte Wahrscheinlichkeit gleich

$$\frac{m \cdot n \cdot (m-1) \cdot (n-1) \cdot \ldots \cdot (m-k+1) \cdot (n-k+1)}{(m \cdot n) \cdot (m \cdot n - 1) \cdot \ldots \cdot (m \cdot n - k + 1)} = \frac{\binom{m}{k} \cdot \binom{n}{k} \cdot k!}{\binom{m \cdot n}{k}}.$$

2. Lösung: *Reihenfolge bei der Auswahl der Kreuzungen wird nicht beachtet.*

Anzahl möglicher Fälle:

$$\binom{m \cdot n}{k}$$

(Aus $m \cdot n$ Kreuzungen k auswählen ohne Zurücklegen und ohne Beachten der Reihenfolge.)

Anzahl günstiger Fälle:

$$\binom{m}{k} \cdot n \cdot (n-1) \cdot \ldots \cdot (n-k+1).$$

(Zuerst aus m Längsstraßen k ohne Zurücklegen und ohne Beachtung der Reihenfolge auswählen. Dann noch aus n Querstraßen k ohne Zurücklegen und *mit* (!) Beachtung der Reihenfolge auswählen.)

Damit ist die gesuchte Wahrscheinlichkeit gleich

$$\frac{\binom{m}{k} \cdot n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{\binom{m \cdot n}{k}} = \frac{\binom{m}{k} \cdot \binom{n}{k} \cdot k!}{\binom{m \cdot n}{k}}.$$

4.3.2 W–Räume mit Zähldichten

Zur Motivierung dient das folgende

Beispiel 4.8 *Mit einem (echten) Würfel wird so lange gewürfelt, bis zum ersten Mal eine 6 erscheint.*

Wie groß ist die Wahrscheinlichkeit, dass die zufällige Anzahl der Würfe bis (einschließlich) zum ersten Wurf mit 6 oben eine gerade Zahl ist ?

Wir wählen

$$\Omega = \mathbb{N} = \{1, 2, \ldots\},$$

wobei $\omega = k$ bedeutet, dass beim k–ten Wurf der Würfel zum ersten Mal mit 6 oben landet. Gefragt ist dann nach der Wahrscheinlichkeit $\mathbf{P}(A)$, wobei

$$A = \{2, 4, 6, 8, \ldots\}.$$

Zur Festlegung der Wahrscheinlichkeit einer Menge legen wir zuerst die Wahrscheinlichkeiten aller Einpunktmengen fest und setzen dann

$$\mathbf{P}(A) = \sum_{\omega \in A} \mathbf{P}(\{\omega\}) \overset{hier}{=} \sum_{k \in \{2,4,6,8,\ldots\}} \mathbf{P}(\{k\}).$$

Um festzustellen, mit welcher Wahrscheinlichkeit der Würfel beim k–ten Wurf zum ersten Mal mit 6 oben landet, beschreiben wir die ersten k Würfe durch einen Laplaceschen W–Raum mit Grundmenge

$$\{(\omega_1, \ldots, \omega_k) \ : \ \omega_i \in \{1, \ldots, 6\}\}.$$

Diese besteht aus insgesamt 6^k Elementen, davon sind $5^{k-1} \cdot 1$ günstig, so dass folgt

$$\mathbf{P}(\{k\}) = \frac{5^{k-1}}{6^k} = \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{k-1}.$$

Die gesuchte Wahrscheinlichkeit ist dann

$$\begin{aligned}
\mathbf{P}(A) &= \sum_{k \in \{2,4,6,8,\ldots\}} \mathbf{P}(\{k\}) \\
&= \frac{1}{6} \cdot \left(\frac{5}{6}\right)^1 + \frac{1}{6} \cdot \left(\frac{5}{6}\right)^3 + \frac{1}{6} \cdot \left(\frac{5}{6}\right)^5 + \frac{1}{6} \cdot \left(\frac{5}{6}\right)^7 + \ldots \\
&= \frac{5}{36} \cdot \left(\left(\frac{5}{6}\right)^0 + \left(\frac{5}{6}\right)^2 + \left(\frac{5}{6}\right)^4 + \left(\frac{5}{6}\right)^6 + \ldots\right) \\
&= \frac{5}{36} \cdot \left(\left(\frac{25}{36}\right)^0 + \left(\frac{25}{36}\right)^1 + \left(\frac{25}{36}\right)^2 + \left(\frac{25}{36}\right)^3 + \ldots\right) \\
&= \frac{5}{36} \cdot \frac{1}{1 - \frac{25}{36}} \\
&\approx 0.455
\end{aligned}$$

Als nächstes betrachten wir eine allgemeine Definitionsmöglichkeit für W–Räume mit endlicher oder abzählbar unendlicher Grundmenge Ω. Hierbei wird sinnvollerweise $\mathcal{A} = \mathcal{P}(\Omega)$ gewählt. Jede beliebige Menge $A \subseteq \Omega$ lässt sich als endliche oder abzählbar unendliche Vereinigung von Einpunktmengen schreiben:

$$A = \bigcup_{\omega \in A} \{\omega\}.$$

Ist $\mathbf{P} : \mathcal{A} \to \mathbb{R}$ ein W–Maß, so folgt daraus aufgrund der σ–Additivität:

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

d.h., $\mathbf{P} : \mathcal{A} \to \mathbb{R}$ ist bereits durch die Werte $\mathbf{P}(\{\omega\})$ $(\omega \in \Omega)$ festgelegt. Wir zeigen in dem folgenden Satz 4.2, dass die obige Beziehung auch zur Definition von W–Maßen ausgehend von den Werten $\mathbf{P}(\{\omega\})$ $(\omega \in \Omega)$ verwendet werden kann.

Satz 4.2 *Sei $\Omega = \{x_1, x_2, \ldots\}$ eine abzählbar unendliche Menge und $(p_k)_{k \in \mathbb{N}}$ eine Folge reeller Zahlen mit*

$$0 \leq p_k \leq 1 \quad (k \in \mathbb{N}) \quad \textit{und} \quad \sum_{k=1}^{\infty} p_k = 1.$$

Dann wird durch $\mathcal{A} := \mathcal{P}(\Omega)$ *und*

$$\mathbf{P}(A) := \sum_{k:x_k \in A} p_k \quad (A \subseteq \Omega)$$

ein W–Raum definiert. Hierbei gilt

$$\mathbf{P}(\{x_k\}) = p_k \quad (k \in \mathbb{N}),$$

d.h. p_k *gibt die Wahrscheinlichkeit an, dass* x_k *das Ergebnis des Zufallsexperiments ist.*

Beweis: Offensichtlich ist Ω eine nichtleere Menge und $\mathcal{A}$ eine σ–Algebra über Ω, also genügt es zu zeigen, dass $\mathbf{P} : \mathcal{A} \to \mathbb{R}$ ein W–Maß ist. Dazu beachtet man zuerst, dass für $|A| = \infty$ die Reihe

$$\sum_{k:x_k \in A} p_k$$

wohldefiniert ist, da die Reihenfolge der Summation bei Reihen mit nichtnegativen Summanden keine Rolle spielt. Dann bleibt noch zu zeigen:

(i) $\mathbf{P}(A) \geq 0$ für alle $A \subseteq \Omega$.

(ii) $\mathbf{P}(\Omega) = 1$.

(iii) $\mathbf{P}$ ist σ–additiv.

Unter Beachtung von $p_k \geq 0$ und $\sum_{k=1}^{\infty} p_k = 1$ folgen (i) und (ii) unmittelbar aus der Definition von $\mathbf{P}$.

Zum Nachweis von (iii) betrachten wir Mengen $A_1, A_2, \dots \subseteq \Omega$ mit $A_i \cap A_j = \emptyset$ für alle $i \neq j$. Zu zeigen ist

$$\mathbf{P}\left(\cup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathbf{P}(A_j).$$

Mit der Definition von $\mathbf{P}$ folgt

$$\text{linke Seite} = \sum_{k:x_k \in \cup_{j=1}^{\infty} A_j} p_k$$

und

$$\text{rechte Seite} = \sum_{j=1}^{\infty} \sum_{k:x_k \in A_j} p_k.$$

Bei beiden Summen summiert man alle p_k auf, für die x_k in einer der Mengen A_j ist. Unterscheiden tun sich die beiden Summen nur hinsichtlich der Reihenfolge, in der die p_k's aufsummiert werden. Da aber (wie oben bereits erwähnt) bei endlichen oder abzählbar unendlichen Summen mit nichtnegativen Summanden die Reihenfolge der Summation keine Rolle spielt, stimmen beide Werte überein. □

Gemäß obigem Satz kann also ein W–Raum bereits durch Vorgabe einer Folge von nichtnegativen Zahlen, die zu Eins summieren, eindeutig bestimmt werden. Aus dem Beweis des Satzes ist unmittelbar klar, dass er analog auch für endliche Grundmengen $\Omega = \{x_1, \ldots, x_N\}$ und $0 \leq p_k$ $(k = 1, \ldots, N)$ mit $\sum_{k=1}^N p_k = 1$ gilt.

Definition 4.7 *Die Folge $(p_k)_{k\in\mathbb{N}}$ (bzw. $(p_k)_{k=1,\ldots,N}$ im Falle einer N-elementigen Grundmenge) heißt* **Zähldichte** *des W–Maßes* **P** *in Satz 4.2.*

Zur Illustration betrachten wir das folgende

Beispiel 4.9 *Sonntagsfrage*

Bei einer telefonischen Umfrage (mit rein zufällig gewählten Telefonnummern) werden n Personen gefragt, welche Partei sie wählen würden, wenn nächsten Sonntag Bundestagswahl wäre. Es sei $p \in [0, 1]$ der prozentuale Anteil desjenigen Teils der gesamten Bevölkerung, der SPD wählen würde. Wie groß ist dann die Wahrscheinlichkeit, dass genau k der Befragten ($k \in \{0, \ldots, n\}$ fest) SPD wählen würden ?

Wir Betrachtung zunächst den Spezialfall $n = k = 1$. Sei N die Anzahl aller Wahlberechtigten. Dann sind davon $N \cdot p$ SPD Wähler, und die Wahrscheinlichkeit, bei rein zufälligem Herausgreifen einer Person aus den N Personen einen der $N \cdot p$ SPD Wähler zu erhalten ist

$$\frac{\text{Anzahl günstiger Fälle}}{\text{Anzahl möglicher Fälle}} = \frac{N \cdot p}{N} = p.$$

Analog ist die Wahrscheinlichkeit, bei rein zufälligem Herausgreifen einer Person aus den N Personen keinen der $N \cdot p$ SPD Wähler zu erhalten, gegeben durch

$$\frac{N - N \cdot p}{N} = 1 - p.$$

Nun betrachten wir den allgemeinen Fall. Zwecks Vereinfachung der Rechnung gehen wir davon aus, dass sich der prozentuale Anteil der SPD Wähler nach

Herausgreifen eines Wählers nicht (bzw. nur unwesentlich) verändert. Dann ist die Wahrscheinlichkeit, dass genau die ersten k Befragten SPD Wähler sind und die restlichen $n-k$ nicht, gegeben durch

$$\frac{(N \cdot p)^k (N \cdot (1-p))^{n-k}}{N^n} = p^k (1-p)^{n-k}.$$

Das gleiche Resultat erhält man auch, wenn man beliebige Positionen für die k SPD Wähler unter den n Wählern vorgibt und danach fragt, mit welcher Wahrscheinlichkeit man auf genau so eine Sequenz von Wählern trifft. Da es für die Wahl der k Positionen der SPD Wähler unter den n Positionen genau $\binom{n}{k}$ Möglichkeiten gibt, erhält man für die Wahrscheinlichkeit, dass unter den n Befragten genau k SPD Wähler sind:

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

Das dadurch festgelegte W–Maß heißt Binomialverteilung.

Definition 4.8 *Das gemäß Satz 4.2 durch* $\Omega = \mathbb{N}_0$ *und die Zähldichte* $(b(n,p,k))_{k \in \mathbb{N}_0}$ *mit*

$$b(n,p,k) := \begin{cases} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} & \text{für } \; 0 \le k \le n, \\ 0 & \text{für } \; k > n \end{cases}$$

festgelegte W–Maß heißt **Binomialverteilung** *mit Parametern* $n \in \mathbb{N}$ *und* $p \in [0,1]$.

Gemäß dem binomischen Lehrsatz gilt

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot a^k \cdot b^{n-k}.$$

Wendet man diese Formel mit $a = p$ und $b = 1-p$ an, so erhält man

$$\sum_{k=0}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = (p + (1-p))^n = 1,$$

d.h. es handelt es sich hierbei in der Tat um eine Zähldichte.

Beispiel 4.10 *Bei der Umfrage im Beispiel 4.9 interessiert man sich nun für die Wahrscheinlichkeit, dass der relative Anteil* k/n *der SPD Wähler unter den Befragten um nicht mehr als* 1% *vom Wert* p *in der gesamten Bevölkerung abweicht. Wegen*

$$\left| \frac{k}{n} - p \right| \le 0.01 \Leftrightarrow n \cdot p - 0.01 \cdot n \le k \le n \cdot p + 0.01 \cdot n$$

erhält man dafür

$$\begin{aligned}&\mathbf{P}\left(\{k \in \mathbb{N}_0 \ : \ n \cdot p - 0.01 \cdot n \leq k \leq n \cdot p + 0.01 \cdot n\}\right)\\ &= \sum_{n \cdot p - 0.01 \cdot n \leq k \leq n \cdot p + 0.01 \cdot n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.\end{aligned}$$

Beispiel 4.11 *In einer großen Teigmenge seien $n = 1000$ Rosinen rein zufällig verteilt. Ein Bäcker formt daraus $m = 100$ gleichgroße Brötchen.*

Wie groß ist die Wahrscheinlichkeit, dass ein zufällig herausgegriffenes Brötchen weniger als 8 Rosinen enthält ?

Wir wählen

$$\Omega = \mathbb{N}_0 = \{0, 1, 2, \dots\},$$

wobei $\omega = k$ bedeutet, dass das Brötchen genau k Rosinen enthält. Gefragt ist dann nach der Wahrscheinlichkeit $\mathbf{P}(A)$, wobei

$$A = \{0, 1, 2, 3, 4, 5, 6, 7\}.$$

Zur Festlegung der Wahrscheinlichkeit einer Menge legen wir wieder die Wahrscheinlichkeiten aller Einpunktmengen fest und setzen dann

$$\mathbf{P}(A) = \sum_{\omega \in A} \mathbf{P}(\{\omega\}) \stackrel{hier}{=} \sum_{k=0}^{7} \mathbf{P}(\{k\}).$$

Dazu bestimmen wir zuerst für festes $k \in \{0, 1, \dots, n\}$ die Wahrscheinlichkeit, dass das Brötchen genau k Rosinen enthält. Wir denken uns die Rosinen von 1 bis n und die Brötchen von 1 bis m durchnummeriert. Das zufällig herausgegriffene Brötchen sei das Brötchen mit Nummer 1. Jede der Rosinen landet in einem der m Brötchen. Die Zuordnung der Rosinen zu den Brötchen kann daher durch ein n-Tupel mit Einträgen in $\{1, \dots, m\}$ beschrieben werden, wobei die i-te Komonente die Nummer des Brötchens angibt, in das die i-te Rosine kommt. Dabei gibt es m^n Möglichkeiten, von denen jede mit der gleichen Wahrscheinlichkeit $1/m^n$ auftritt. Damit genau k Rosinen in dem Brötchen mit Nummer 1 landen, müssen in dem n-Tupel genau k Komponenten gleich 1 sein, und alle anderen müssen ungleich 1 sein. Für die Wahl der Positionen dieser k Komponenten mit Eintrag 1 gibt es $\binom{n}{k}$ Möglichkeiten. Damit gibt es insgesamt

$$\binom{n}{k} 1^k (m-1)^{n-k}$$

n-Tupel, bei denen genau k Komponenten 1 sind, und die Wahrscheinlichkeit, dass das Brötchen genau k Rosinen enthält, berechnet sich zu

$$\mathbf{P}(\{k\}) = \frac{\binom{n}{k}\, 1^k (m-1)^{n-k}}{m^n} = \binom{n}{k} \left(\frac{1}{m}\right)^k \left(\frac{m-1}{m}\right)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}$$

mit $p = 1/m = 0.01$.

Die gesuchte Wahrscheinlichkeit ist dann

$$\begin{aligned}
\mathbf{P}(\{0,1,2,3,4,5,6,7\}) &= \sum_{k=0}^{7} \mathbf{P}(\{k\}) \\
&= \sum_{k=0}^{7} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{7} \binom{1000}{k} 0.01^k 0.99^{1000-k}.
\end{aligned}$$

Zur konkreten Berechnung der obigen Summe erweist sich die folgende Approximation als nützlich:

Lemma 4.4 *Seien $\lambda \in \mathbb{R}_+$ und $p_n \in [0,1]$ $(n \in \mathbb{N})$ derart, dass $n \cdot p_n \to \lambda$ $(n \to \infty)$. Dann gilt für jedes feste $k \in \mathbb{N}_0$:*

$$b(n, p_n, k) = \binom{n}{k} \cdot p_n^k \cdot (1-p_n)^{n-k} \to \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad (n \to \infty).$$

Beweis: Wegen $n \cdot p_n \to \lambda$ gilt insbesondere $p_n \to 0$ $(n \to \infty)$. Damit erhält man

$$\begin{aligned}
&b(n, p_n, k) \\
&= \frac{1}{k!} n \cdot (n-1) \cdot \ldots \cdot (n-k+1) \cdot p_n^k \cdot (1-p_n)^{n-k} \\
&= \frac{1}{k!} \cdot np_n \cdot (np_n - p_n) \cdot \ldots \cdot (np_n - (k-1)p_n) \cdot (1-p_n)^{-k} \cdot \left((1-p_n)^{\frac{1}{p_n}}\right)^{n \cdot p_n}.
\end{aligned}$$

Mit

$$n \cdot p_n \to \lambda, (n \cdot p_n - p_n) \to \lambda, \ldots, (n \cdot p_n - (k-1) \cdot p_n) \to \lambda \quad (n \to \infty),$$

$$(1-p_n)^{-k} \to 1 \quad (n \to \infty)$$

und

$$(1-p_n)^{\frac{1}{p_n}} \to e^{-1} \quad (n \to \infty)$$

folgt

$$b(n, p_n, k) \to \frac{1}{k!} \cdot \lambda^k \cdot 1 \cdot (e^{-1})^\lambda \quad (n \to \infty).$$

$\square$

Mit Hilfe von Lemma 4.4 lässt sich motivieren, die Wahrscheinlichkeit in Beispiel 4.11 approximativ folgendermaßen zu berechnen:

$$\begin{aligned}
\mathbf{P}(\{0,1,2,3,4,5,6,7\}) &= \sum_{k=0}^{7} \binom{1000}{k} 0.01^k 0.99^{1000-k} \\
&\approx \sum_{k=0}^{7} \frac{\lambda^k}{k!} \cdot e^{-\lambda} \text{ mit } \lambda = 1000 \cdot 0.01 = 10 \\
&= \sum_{k=0}^{7} \frac{10^k}{k!} \cdot e^{-10} \\
&\approx 0.22
\end{aligned}$$

Definition 4.9 *Das gemäß Satz 4.2 durch* $\Omega = \mathbb{N}_0$ *und die Zähldichte* $(\pi(\lambda, k))_{k \in \mathbb{N}_0}$ *mit*

$$\pi(\lambda, k) := \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad (k \in \mathbb{N}_0)$$

festgelegte W–Maß heißt **Poisson–Verteilung** *mit Parameter* $\lambda \in \mathbb{R}_+$.

Wegen

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \cdot e^{-\lambda} = e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{+\lambda} = 1$$

handelt es sich hierbei in der Tat um eine Zähldichte.

Eine weitere Approximation der Binomialverteilung wird am Ende dieses Kapitels vorgestellt.

4.3.3 W–Räume mit Dichten

Zur Motivation betrachten wir

Beispiel 4.12 *Eine Zahl wird rein zufällig aus dem Intervall* $[0, 1]$ *ausgewählt.*

Wie groß ist die Wahrscheinlichkeit, dass die Zahl zwischen $\frac{1}{3}$ *und* $\frac{1}{2}$ *liegt ?*

Wir wählen
$$\Omega = \mathbb{R},$$
wobei $\omega \in \Omega$ die rein zufällig aus $[0,1]$ gezogene Zahl ist (hierbei treten Zahlen außerhalb von $[0,1]$ nur mit Wahrscheinlichkeit Null auf). Gefragt ist dann nach der Wahrscheinlichkeit $\mathbf{P}(A)$, wobei
$$A = \left[\frac{1}{3}, \frac{1}{2}\right].$$

Diesmal ist die Definition
$$\mathbf{P}(A) := \sum_{\omega \in A} \mathbf{P}(\{\omega\})$$
nicht sinnvoll, da hier gilt:
$$\mathbf{P}(\{\omega\}) = 0 \text{ für alle } \omega \in \Omega.$$

Eine naheliegende Idee ist jedoch, die Summe oben durch ein Integral anzunähern, d.h. zu setzen
$$\mathbf{P}(A) := \int_A f(x)\,dx,$$
mit $f : \mathbb{R} \to \mathbb{R}$.

Damit die obigen Wahrscheinlichkeiten nichtnegativ sind, fordern wir
$$f(x) \geq 0 \text{ für alle } x \in \mathbb{R}.$$

Da $\mathbf{P}(\mathbb{R})$ darüberhinaus Eins sein soll, fordern wir auch
$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Berücksichtigt man, dass Zahlen außerhalb von $[0,1]$ nur mit Wahrscheinlichkeit Null auftreten sollen, sowie jede Zahl aus $[0,1]$ mit der "gleichen Wahrscheinlichkeit" auftreten soll, so ist es naheliegend, im obigen Beispiel zu wählen:
$$f(x) = \begin{cases} 1 & \text{für} \quad 0 \leq x \leq 1, \\ 0 & \text{für} \quad x < 0 \text{ oder } x > 1. \end{cases}$$

Damit erhält man für die gesuchte Wahrscheinlichkeit:
$$\mathbf{P}\left(\left[\frac{1}{3}, \frac{1}{2}\right]\right) = \int_{\left[\frac{1}{3}, \frac{1}{2}\right]} f(x)\,dx = \int_{1/3}^{1/2} 1\,dx = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}.$$

Im Folgenden wird eine allgemeine Definitionsmöglichkeit für W–Räume mit Grundmenge $\Omega = \mathbb{R}$ vorgestellt. Hierbei ist zwar $\mathcal{P}(\mathbb{R})$ eine σ–Algebra über Ω, diese ist für die Festlegung von Wahrscheinlichkeiten aber meist zu groß (z.B. kann die Existenz der im unten stehenden Satz verwendeten Integrale nicht für alle Mengen $A \subseteq \mathbb{R}$ nachgewiesen werden). Daher wählen wir als σ–Algebra die Borelsche σ–Algebra $\mathcal{B}$.

Wie in Beipiel 4.12 ist die Festlegung eines W–Maßes durch

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

nicht möglich, da hier meist $\mathbf{P}(\{\omega\}) = 0$ für alle $\omega \in \Omega$ gilt. Eine naheliegende Idee ist jedoch, die Summe oben durch ein Integral anzunähern.

Satz 4.3 *Ist $f : \mathbb{R} \to \mathbb{R}$ eine Funktion, für die gilt*

$$f(x) \geq 0 \text{ für alle } x \in \mathbb{R} \quad \text{und} \quad \int_{\mathbb{R}} f(x)\, dx = 1$$

(insbesondere sei hier die Existenz des Integrals vorausgesetzt), so wird durch $\Omega := \mathbb{R}$, $\mathcal{A} := \mathcal{B}$ und

$$\mathbf{P}(A) = \int_A f(x)\, dx \quad (A \in \mathcal{B})$$

ein W–Raum definiert.

Beweis: Wieder genügt es zu zeigen, dass $\mathbf{P}$ ein W–Maß ist. Wegen $f(x) \geq 0$ für alle x gilt $\mathbf{P}(A) \geq 0$ $(A \in \mathcal{A})$. Weiter ist

$$\mathbf{P}(\mathbb{R}) = \int_{\mathbb{R}} f(x) dx = 1.$$

Bei geeigneter Definition der auftretenden Integrale kann man auch zeigen, dass $\mathbf{P}$ σ–additiv ist. Mit Lemma 4.1 folgt die Behauptung. □

Definition 4.10 *f heißt* **Dichte** *(bzgl. des LB–Maßes) von dem in Satz 4.3 definierten W–Maß* $\mathbf{P}$.

Bemerkung: Ist $(\Omega, \mathcal{A}, \mathbf{P})$ der W–Raum aus Satz 4.3 und sind $a, b \in \mathbb{R}$ mit $a < b$, so gilt für die Wahrscheinlichkeit, dass beim zugrundeliegenden Zufallsexperiment ein Wert zwischen a und b auftritt:

$$\mathbf{P}((a,b)) = \int_{(a,b)} f(x)\ dx = \int_a^b f(x)\ dx.$$

Das folgende W–Maß haben wir bereits in Beispiel 4.12 kennengelernt.

Definition 4.11 *Die Gleichverteilung* $U(a,b)$ *mit Parametern* $-\infty < a < b < \infty$ *ist das durch die Dichte*

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für} \quad a \leq x \leq b, \\ 0 & \text{für} \quad x < a \text{ oder } x > b \end{cases}$$

gemäß Satz 4.3 festgelegte W–Maß.

Wegen

$$\int_{\mathbb{R}} f(x)\, dx = \frac{1}{b-a} \int_a^b 1\, dx = 1$$

sind hierbei die Voraussetzungen von Satz 4.3 erfüllt.

Ein weiteres W–Maß mit Dichte führen wir ein in

Beispiel 4.13 *Die Lebensdauer einer Glühbirne betrage im Schnitt 24 Monate. Wie groß ist die Wahrscheinlichkeit, dass die Glühbirne bereits innerhalb von drei Monaten ausfällt ?*

Wir wählen

$$\Omega = \mathbb{R}_+,$$

wobei ω die Lebensdauer der Glühbirne in Monaten ist. Gefragt ist dann nach der Wahrscheinlichkeit $\mathbf{P}(A)$, wobei

$$A = [0, 3]\,.$$

Diese Wahrscheinlichkeit lässt sich ohne Zusatzvoraussetzungen an den zugrunde liegenden Zufallsmechanismus nicht berechnen.

Lebensdauern modelliert man häufig mit der sogenannten Exponentialverteilung:

Definition 4.12 *Die Exponentialverteilung* $\exp(\lambda)$ *mit Parameter* $\lambda > 0$ *ist das durch die Dichte*

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & \text{für} \quad x \geq 0, \\ 0 & \text{für} \quad x < 0 \end{cases}$$

gemäß Satz 4.3 festgelegte W–Maß.

Wegen

$$\int_{\mathbb{R}} f(x)\, dx = \int_0^\infty \lambda \cdot e^{-\lambda \cdot x}\, dx = -e^{-\lambda \cdot x}\Big|_{x=0}^{\infty} = 1$$

sind hierbei die Voraussetzungen von Satz 4.3 erfüllt.

Bei der Exponentialverteilung ist $1/\lambda$ die “mittlere Lebensdauer” (wird später noch bewiesen). Daher gehen wir im Beispiel 4.13 davon aus, dass gilt

$$\mathbf{P}(A) := \int_A f(x)\, dx$$

mit

$$f(x) = \begin{cases} \frac{1}{24} \cdot e^{-x/24} & \text{für} \quad x \geq 0, \\ 0 & \text{für} \quad x < 0 \end{cases}$$

und berechnen die gesuchte Wahrscheinlichkeit zu

$$\begin{aligned} \mathbf{P}\left([0,3]\right) &= \int_0^3 \frac{1}{24} \cdot e^{-x/24} dx \\ &= -e^{-x/24}\Big|_{x=0}^{3} \\ &= -e^{-3/24} + e^0 \\ &\approx 0.118 \end{aligned}$$

Ein weiteres Beispiel für ein W–Maß mit Dichte ist gegeben in

Definition 4.13 *Die Normalverteilung $N(a, \sigma^2)$ mit Parametern $a \in \mathbb{R}, \sigma > 0$ ist das durch die Dichte*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}} \quad (x \in \mathbb{R})$$

gemäß Satz 4.3 festgelegte W–Maß.

Wegen

$$\int_{\mathbb{R}} f(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sigma} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = 1$$

sind hierbei wieder die Voraussetzungen von Satz 4.3 erfüllt.

4.3.4 Verallgemeinerung der Begriffe Dichte und Zähldichte

Die Begriffe Dichte und Zähldichte lassen sich verallgemeinern. Dazu dient die folgende Definition:

Definition 4.14 *Ω Grundmenge, $\mathcal{A}$ σ–Algebra. Eine Abbildung*

$$\mu : \mathcal{A} \to \bar{\mathbb{R}}_+$$

(mit $\bar{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$) heißt **Maß**, *wenn gilt:*

(i) $\mu(\emptyset) = 0$,

(ii) $\mu(A \cup B) = \mu(A) + \mu(B)$ *für alle* $A, B \in \mathcal{A}$ *mit* $A \cap B = \emptyset$.

(iii)

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$$

für alle $A_n \in \mathcal{A}$ $(n \in \mathbb{N})$ *mit* $A_i \cap A_j = \emptyset$ *für* $i \neq j$.

In diesem Fall heißt $(\Omega, \mathcal{A}, \mu)$ **Maßraum**.

Unmittelbar aus obiger Definition folgt, dass μ genau dann ein W–Maß ist, wenn μ ein Maß ist und $\mu(\Omega) = 1$ gilt.

Beispiele für Maße:

a) $\Omega = \mathbb{N}_0$, $\mathcal{A} = \mathcal{P}(\Omega)$ und

$$\mu(A) = |A| \quad (A \subseteq \mathbb{N}_0).$$

μ heißt **abzählendes Maß**.

b) $\Omega = \mathbb{R}$, $\mathcal{A} = \mathcal{B}$ und

$$\mu : \mathcal{B} \to \bar{\mathbb{R}}_+$$

dasjenige Maß mit

$$\mu((a, b]) = b - a$$

für alle $-\infty < a \leq b < \infty$.

μ heißt **Lebesgue–Borel–Maß** (kurz: LB–Maß).

Sei nun $(\Omega, \mathcal{A}, \mu)$ ein Maßraum. Dann kann man durch Vorgabe einer Funktion

$$f : \Omega \to \mathbb{R}_+ \text{ mit } \int_\Omega f(x)\, \mu(dx) = 1$$

ein W–Maß $\mathbf{P} : \mathcal{A} \to \mathbb{R}_+$ definieren durch

$$\mathbf{P}(A) := \int_A f(x)\, \mu(dx) \quad (A \in \mathcal{A}). \tag{4.1}$$

Hierbei wird der in Abschnitt 4.6 definierte Integralbegriff verwendet.

Man sagt dann, dass f *Dichte von* $\mathbf{P}$ *bzgl.* μ ist.

Man kann nun zeigen: Ist μ das abzählende Maß, so erhält man mittels (4.1) W–Maße mit Zähldichten. Ist dagegen μ das LB–Maß, so besitzt das durch (4.1) definierte Maß eine Dichte bezüglich dem LB–Maß.

4.4 Bedingte Wahrscheinlichkeit und Unabhängigkeit

Im Folgenden untersuchen wir, wie sich das wahrscheinlichkeitstheoretische Verhalten eines Zufallsexperiments ändert, falls Zusatzinformation über den Ausgang bekannt wird. Zur Motivierung betrachten wir

Beispiel 4.14 *Beim sogenannten Down–Syndrom (Mongolismus) ist das Chromosom 21 dreifach – statt wie sonst zweifach – vorhanden, was zu meist schwerer geistiger Behinderung führt. Im Rahmen einer Fruchtwasseruntersuchung kann festgestellt werden, ob ein ungeborenes Kind diesen Defekt hat oder nicht. Dazu wird unter Ultraschallsicht durch die Bauchdecke der Schwangeren etwas Fruchtwasser abgenommen. Dieses enthält kindliche Zellen, die im Labor vermehrt und auf Fehler beim Chromosomensatz des Kindes hin untersucht werden können. Nachteil dieser Untersuchung ist allerdings, dass es in ca.* 0.5% *der Fälle zu Komplikationen wie Fehlgeburt und Missbildungen beim Kind kommen kann.*

Eine deutlich weniger aufwendige Untersuchung ist der sogenannte Triple Test, bei dem im Rahmen einer Blutuntersuchung in der 15. Schwangerschaftswoche drei Laborwerte des Blutes der Mutter bestimmt werden. Sind zwei dieser Werte erhöht, der dritte hingegen nicht, so sagt man, dass der Triple Test positiv ausfällt.

Im Folgenden soll die Frage untersucht werden, wie sich die Wahrscheinlichkeit, ein Kind mit Down–Syndrom zu bekommen, ändert, falls der Triple Test positiv ausfällt.

Zur Beantwortung obiger Frage wird zuerst einmal die bedingte Wahrscheinlichkeit eines Ereignisses A unter einer Bedingung B definiert. Zur Motivation der Definition betrachten wir die n–malige Durchführung eines Zufallsexperiments. n_A bzw. n_B bzw. $n_{A\cap B}$ seien die Anzahlen des Eintretens des Ereignisses A bzw. B bzw. $A \cap B$. Eine naheliegende Approximation der bedingten Wahrscheinlichkeit von A unter der Bedingung B ist dann die relative Häufigkeit des Auftretens von A unter den Ausgängen des Zufallsexperimentes, bei denen auch B eingetreten ist, d.h.,

$$\frac{n_{A\cap B}}{n_B} = \frac{\frac{n_{A\cap B}}{n}}{\frac{n_B}{n}}.$$

Dies motiviert

Definition 4.15 *Sei* $(\Omega, \mathcal{A}, \mathbf{P})$ *ein W–Raum und seien* $A, B \in \mathcal{A}$ *mit* $\mathbf{P}(B) > 0$. *Dann heißt*

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

bedingte Wahrscheinlichkeit *von* A *unter der Bedingung* B.

Lemma 4.5 *Sei* $(\Omega, \mathcal{A}, \mathbf{P})$ *ein W-Raum und* $B \in \mathcal{A}$ *mit* $\mathbf{P}(B) > 0$. *Dann wird durch*

$$\tilde{\mathbf{P}}(A) = \mathbf{P}(A|B) \quad (A \in \mathcal{A})$$

ein W-Raum $(\Omega, \mathcal{A}, \tilde{\mathbf{P}})$ *definiert. In diesem gilt:*

$$\tilde{\mathbf{P}}(B) = \mathbf{P}(B|B) = \frac{\mathbf{P}(B \cap B)}{\mathbf{P}(B)} = 1.$$

(Sprechweise: " Das W-Maß $\tilde{\mathbf{P}}$ *ist auf* B *konzentriert").*

Beweis. Offensichtlich gilt $\tilde{\mathbf{P}}(A) \geq 0$ für alle $A \in \mathcal{A}$ und

$$\tilde{\mathbf{P}}(\Omega) = \mathbf{P}(\Omega \cap B)/\mathbf{P}(B) = 1.$$

Sind darüberhinaus $A_1, A_2, \dots \in \mathcal{A}$ mit $A_i \cap A_j = \emptyset$ für $i \neq j$, so folgt aus $\mathbf{P}$ W-Maß:

$$\begin{aligned}
\tilde{\mathbf{P}}\left(\cup_{n=1}^{\infty} A_n\right) &= \mathbf{P}\left(\cup_{n=1}^{\infty} A_n | B\right) \\
&= \frac{\mathbf{P}\left(\left(\cup_{n=1}^{\infty} A_n\right) \cap B\right)}{\mathbf{P}(B)} \\
&= \frac{\mathbf{P}\left(\cup_{n=1}^{\infty} (A_n \cap B)\right)}{\mathbf{P}(B)} \\
&= \frac{\sum_{n=1}^{\infty} \mathbf{P}\left(A_n \cap B\right)}{\mathbf{P}(B)} \\
&= \sum_{n=1}^{\infty} \frac{\mathbf{P}\left(A_n \cap B\right)}{\mathbf{P}(B)} \\
&= \sum_{n=1}^{\infty} \tilde{\mathbf{P}}(A_n).
\end{aligned}$$

Mit Lemma 4.1 folgt die Behauptung. □

Aus obigem Lemma folgt, dass für

$$A \mapsto \mathbf{P}(A|B)$$

die üblichen Rechenregeln für Wahrscheinlichkeiten gelten, z.B. ist

$$\mathbf{P}(A^c|B) = 1 - \mathbf{P}(A|B)$$

und

$$\mathbf{P}(A_1 \cup A_2|B) = \mathbf{P}(A_1|B) + \mathbf{P}(A_2|B) \quad \text{falls } A_1 \cap A_2 = \emptyset.$$

Im Beispiel 4.14 interessieren wir uns für die bedingte Wahrscheinlichkeit

$$\mathbf{P}(A|B),$$

wobei

- $A =$ "Kind mit Down–Syndrom",
- $B =$ "Triple Test positiv".

Bekannt sind die folgenden Näherungswerte:

- $\mathbf{P}(A) = 0.0014$ (ohne Berücksichtigung des Alters der Mutter)
- $\mathbf{P}(B|A) = 0.65$
- $\mathbf{P}(B|A^c) = 0.075$

Der folgende Satz zeigt, wie man daraus $\mathbf{P}(A|B)$ berechnen kann.

Satz 4.4 *Sei* $(\Omega, \mathcal{A}, \mathbf{P})$ *ein W–Raum und seien* $B_1, \dots, B_N \in \mathcal{A}$ *mit*

$$\Omega = \cup_{n=1}^N B_n,$$

$$B_i \cap B_j = \emptyset \textit{ für alle } i \neq j$$

und

$$\mathbf{P}(B_n) > 0 \;\; (n = 1, \dots, N).$$

Dann gilt:

a)

$$\mathbf{P}(A) = \sum_{n=1}^{N} \mathbf{P}(A|B_n) \cdot \mathbf{P}(B_n) \quad \textit{für alle } A \in \mathcal{A}.$$

(Formel von der totalen Wahrscheinlichkeit)

b)

$$\mathbf{P}(B_k|A) = \frac{\mathbf{P}(A|B_k) \cdot \mathbf{P}(B_k)}{\sum_{n=1}^{N} \mathbf{P}(A|B_n) \cdot \mathbf{P}(B_n)}$$

für alle $k \in \{1, \dots, N\}$ *und alle* $A \in \mathcal{A}$ *mit* $\mathbf{P}(A) > 0$.

(Formel von Bayes, 1793)

Beweis: a) Es gilt

$$A = A \cap \Omega = A \cap \left(\cup_{n=1}^{N} B_n\right) = \cup_{n=1}^{N} A \cap B_n,$$

wobei die letzte Vereinigung eine endliche Vereinigung von Mengen mit paarweise leerem Schnitt ist. Mit $\mathbf{P}$ W–Maß folgt:

$$\mathbf{P}(A) = \sum_{n=1}^{N} \mathbf{P}(A \cap B_n) = \sum_{n=1}^{N} \frac{\mathbf{P}(A \cap B_n)}{\mathbf{P}(B_n)} \cdot \mathbf{P}(B_n) = \sum_{n=1}^{N} \mathbf{P}(A|B_n) \cdot \mathbf{P}(B_n).$$

b) Nach Definition der bedingten Wk. gilt:

$$\mathbf{P}(B_k|A) = \frac{\mathbf{P}(B_k \cap A)}{\mathbf{P}(A)} = \frac{\frac{\mathbf{P}(B_k \cap A)}{\mathbf{P}(B_k)} \cdot \mathbf{P}(B_k)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A|B_k) \cdot \mathbf{P}(B_k)}{\mathbf{P}(A)}.$$

Mit a) folgt die Behauptung. □

Mit Satz 4.4 erhält man im Beispiel 4.14

$$\begin{aligned}
\mathbf{P}(A|B) &= \frac{\mathbf{P}(B|A) \cdot \mathbf{P}(A)}{\mathbf{P}(B|A) \cdot \mathbf{P}(A) + \mathbf{P}(B|A^c) \cdot \mathbf{P}(A^c)} \\
&= \frac{0.65 \cdot 0.0014}{0.65 \cdot 0.0014 + 0.075 \cdot 0.9986} \\
&\approx 0.012,
\end{aligned}$$

d.h. selbst wenn der Triple Test positiv ausfällt, so beträgt die Wahrscheinlichkeit, ein Kind mit Down–Syndrom zu bekommen, gerade mal 1.2% (oder anders ausgedrückt, mit Wahrscheinlichkeit 98.8% hat das Kind kein Down–Syndrom). Dagegen führt die üblicherweise nach positivem Triple Test empfohlene Fruchtwasseruntersuchung in ca. 0.5% der Fälle zu Komplikationen (Fehlgeburt, Missbildungen, etc.)

Ist der Triple Test dagegen negativ, so sinkt die Wahrscheinlichkeit, ein Kind mit Down–Syndrom zu bekommen, es gilt nämlich:

$$\begin{aligned}\mathbf{P}(A|B^c) &= \frac{\mathbf{P}(B^c|A)\cdot\mathbf{P}(A)}{\mathbf{P}(B^c|A)\cdot\mathbf{P}(A)+\mathbf{P}(B^c|A^c)\cdot\mathbf{P}(A^c)}\\ &= \frac{0.35\cdot 0.0014}{0.35\cdot 0.0014+0.925\cdot 0.9986}\\ &\approx 0.0005.\end{aligned}$$

Allerdings ist auch dieser Wert nicht allzu viel kleiner als $\mathbf{P}(A)$.

Andere Resultate bekommt man bei Frauen über 35, da bei diesen der Wert von $\mathbf{P}(A)$ höher ausfällt und damit auch die durch $\mathbf{P}(A|B)$ gegebene Aussagekraft des positiven Testergebnisses steigt.

Bemerkung: Im Beweis von Satz 4.4 wurde verwendet, dass die Wahrscheinlichkeit einer Vereinigung nicht überlappender Mengen gleich der Summe der Wahrscheinlichkeiten ist. Da dies nicht nur für endliche, sondern auch für abzählbar unendliche Vereinigungen gilt, gelten analoge Aussagen auch für Mengen $B_n \in \mathcal{A}$ $(n \in \mathbb{N})$ mit

$$B_i \cap B_j = \emptyset \text{ für alle } i \neq j, \quad \Omega = \cup_{n=1}^{\infty} B_n \quad \text{und} \quad \mathbf{P}(B_n) > 0 \ \ (n \in \mathbb{N}).$$

Z.B. erhält man in diesem Fall für die Formel von Bayes:

$$\mathbf{P}(B_k|A) = \frac{\mathbf{P}(A|B_k)\cdot\mathbf{P}(B_k)}{\sum_{n=1}^{\infty}\mathbf{P}(A|B_n)\cdot\mathbf{P}(B_n)}$$

für alle $k \in \mathbb{N}$ und beliebige $A \in \mathcal{A}$ mit $\mathbf{P}(A) > 0$.

Im Folgenden möchten wir definieren, wann sich zwei Ereignisse gegenseitig nicht beeinflussen. Naheliegende Forderung dafür ist

$$\mathbf{P}(A|B) = \mathbf{P}(A) \quad \text{und} \quad \mathbf{P}(B|A) = \mathbf{P}(B).$$

Für $\mathbf{P}(B) > 0$ erhält man

$$\mathbf{P}(A|B) = \mathbf{P}(A) \Leftrightarrow \frac{\mathbf{P}(A\cap B)}{\mathbf{P}(B)} = \mathbf{P}(A) \Leftrightarrow \mathbf{P}(A\cap B) = \mathbf{P}(A)\cdot\mathbf{P}(B).$$

Die letzte Bedingung kann auch für $\mathbf{P}(A) = 0$ oder $\mathbf{P}(B) = 0$ betrachtet werden und man definiert:

Definition 4.16 *W–Raum* $(\Omega, \mathcal{A}, \mathbf{P})$. *Zwei Ereignisse* $A, B \in \mathcal{A}$ *heißen* **unabhängig**, *falls gilt:*

$$\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B).$$

Bemerkung: Gemäß obiger Herleitung gilt im Falle $\mathbf{P}(A) > 0$ und $\mathbf{P}(B) > 0$:

$$A, B \text{ unabhängig} \quad \Leftrightarrow \quad \mathbf{P}(A|B) = \mathbf{P}(A) \quad \text{und} \quad \mathbf{P}(B|A) = \mathbf{P}(B).$$

Bei unabhängigen Ereignissen beeinflusst also das Eintreten eines der Ereignisse nicht die Wahrscheinlichkeit des Eintretens des anderen.

Beispiel 4.15 *Wir Betrachten das Werfen zweier echter Würfel. Sei A das Ereignis, dass der erste Würfel mit* 6 *oben landet und sei B das Ereignis, dass der zweite Würfel mit* 3 *oben landet. Beschreibt man dieses Zufallsexperiment durch einen Laplaceschen W–Raum mit Grundmenge*

$$\Omega = \{(i,j) \ : \ i, j \in \{1, \dots, 6\}\},$$

so sieht man

$$\mathbf{P}(A \cap B) = \frac{1}{36} = \frac{6}{36} \cdot \frac{6}{36} = \mathbf{P}(A) \cdot \mathbf{P}(B),$$

also sind A und B unabhängig.

Ist C das Ereignis, dass die Summe der Augenzahlen 12 *ist, so gilt*

$$\mathbf{P}(B \cap C) = \mathbf{P}(\emptyset) = 0 \neq \frac{6}{36} \cdot \frac{1}{36} = \mathbf{P}(B) \cdot \mathbf{P}(C),$$

also sind B und C nicht unabhängig.

Allgemeiner definiert man:

Definition 4.17 *W–Raum* $(\Omega, \mathcal{A}, \mathbf{P})$. *Eine Familie* $\{A_i \ : \ i \in I\}$ *von Ereignissen* $A_i \in \mathcal{A}$ *heißt* **unabhängig**, *falls für jede endliche Teilmenge J von I gilt:*

$$\mathbf{P}\left(\cap_{j \in J} A_j\right) = \prod_{j \in J} \mathbf{P}(A_j).$$

Bemerkung: Ist eine Familie $\{A_i \ : \ i \in I\}$ von Ereignissen unabhängig, so sind für alle $i, j \in I$, $i \neq j$ auch die Ereignisse A_i und A_j unabhängig (folgt mit $J = \{i, j\}$). Die Umkehrung gilt aber im allgemeinen nicht:

Sei $(\Omega, \mathcal{A}, \mathbf{P})$ ein Laplacescher W–Raum mit $\Omega = \{1, 2, 3, 4\}$ und seien

$$A_1 = \{1, 2\}, A_2 = \{1, 3\} \text{ und } A_3 = \{2, 3\}.$$

Dann besteht für alle $i \neq j$ die Menge $A_i \cap A_j$ aus genau einem Element. Daraus folgt:

$$\mathbf{P}(A_i \cap A_j) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbf{P}(A_i) \cdot \mathbf{P}(A_j).$$

Darüberhinaus gilt aber:

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(\emptyset) = 0 \neq \mathbf{P}(A_1) \cdot \mathbf{P}(A_2) \cdot \mathbf{P}(A_3).$$

4.5 Zufallsvariablen

Oft interessieren nur Teilaspekte des Ergebnisses eines Zufallsexperimentes. Dies kann man dadurch modellieren, dass man eine Menge Ω' und eine Abbildung $X : \Omega \to \Omega'$ wählt und $X(\omega)$ anstelle des Ergebnisses ω des Zufallsexperimentes betrachtet.

Beispiel 4.16 *Zufällige Auswahl von Wohnungen zur Erstellung eines Mietspiegels.*

Hier interessiert anstelle einer zufällig ausgewählten Wohnung ω nur Teilaspekte dieser Wohnung wie z.B.

- $X(\omega)$ = *Nettomiete pro Quadratmeter,*
- $Y(\omega)$ = *(Nettomiete, Größe in Quadratmetern),*
- $Z(\omega)$= *Anzahl der Zimmer.*

Wir untersuchen im Folgenden, wie man einen W–Raum $(\Omega', \mathcal{A}', \mathbf{P}_X)$ konstruieren kann, der das Zufallsexperiment mit Ergebnis $X(\omega)$ beschreibt.

$X(\omega)$ liegt genau dann in A', wenn das zufällige Ergebnis ω des Zufallsexperiments in der Menge

$$\{\bar{\omega} \in \Omega \ : \ X(\bar{\omega}) \in A'\}$$

liegt. Daher ist es naheliegend zu definieren

$$\mathbf{P}_X(A') := \mathbf{P}[X \in A'] := \mathbf{P}\left(\{\omega \in \Omega \ : \ X(\omega) \in A'\}\right). \tag{4.2}$$

Damit diese Wahrscheinlichkeit wohldefiniert ist, muss

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

erfüllt sein. Abbildungen X, für die das für alle betrachteten Mengen gilt, heißen *Zufallsvariablen.*

Definition 4.18 Ω' *Grundmenge,* $\mathcal{A}'$ σ*–Algebra über* Ω'*. Dann heißt* $(\Omega', \mathcal{A}')$ **Messraum***.*

Definition 4.19 $(\Omega, \mathcal{A}, \mathbf{P})$ *W–Raum,* $(\Omega', \mathcal{A}')$ *Messraum. Dann heißt jede Abbildung*

$$X : \Omega \to \Omega'$$

mit

$$X^{-1}(A') := \{\omega \in \Omega \,:\, X(\omega) \in A'\} \in \mathcal{A} \quad \textit{für alle } A' \in \mathcal{A}'$$

Zufallsvariable *(kurz: ZV). Im Fall* $\Omega' = \mathbb{R}$ *und* $\mathcal{A} = \mathcal{B}$ *heißt* X **reelle Zufallsvariable***.*

Der Begriff Zufallsvariable ist zunächst einmal nur eine Bezeichnung. Obwohl sich diese sicherlich mit einiger Mühe rechtfertigen lässt, sei darauf hingewiesen, dass es sich bei einer Zufallsvariablen keineswegs um eine Variable, sondern um eine Abbildung handelt. Es ist daher nicht sinnvoll, den Begriff Zufallsvariable zu intensiv zu interpretieren.

Beispiel 4.17 *n Personen stimmen bei einer Abstimmung über zwei Vorschläge A und B ab. Dabei entscheidet sich jede Person unabhängig von den anderen mit Wahrscheinlichkeit $p \in [0,1]$ für Vorschlag A und mit Wahrscheinlichkeit $1-p$ für B. Gesucht ist eine Möglichkeit zur stochastischen Modellierung des Abstimmungsverhaltens der n Personen.*

Als Ergebnis des Zufallsexperiments betrachten wir

$$\omega = (x_1, \ldots, x_n) \quad \text{mit } x_1, \ldots, x_n \in \{0, 1\},$$

wobei $x_i = 1$ bzw. $x_i = 0$ bedeutet, dass die i–te Person für Vorschlag A bzw. Vorschlag B stimmt.

Der zugehörige Wahrscheinlichkeitsraum ist dann $(\Omega, \mathcal{A}, \mathbf{P})$ mit

$$\Omega = \{(x_1, \ldots, x_n) : x_i \in \{0,1\}\}, \quad \mathcal{A} = \mathcal{P}(\Omega) \quad \text{und} \quad \mathbf{P} : \mathcal{A} \to [0,1]$$

festgelegt durch

$$\mathbf{P}(\{(x_1, \ldots, x_n)\}) = \prod_{i=1}^{n} (p^{x_i} \cdot (1-p)^{1-x_i}) = p^{\sum_{i=1}^n x_i} \cdot (1-p)^{n - \sum_{i=1}^n x_i}$$

für $x_1, \ldots, x_n \in \{0,1\}$.

Interessiert man sich aber nur für die Anzahl der Stimmen für Vorschlag A (und nicht für die Reihenfolge), so ist es naheliegend statt

$$(x_1, \ldots, x_n)$$

nur

$$X((x_1, \ldots, x_n)) = x_1 + \cdots + x_n$$

zu betrachten.

Da hier $\mathcal{A} = \mathcal{P}(\Omega)$ gewählt wurde, ist die Bedingung

$$X^{-1}(A') \in \mathcal{A}$$

trivialerweise für alle A' erfüllt. Also ist X (unabhängig von der Wahl von $\mathcal{A}'$) eine Zufallsvariable gemäß obiger Definition. Im Folgenden bestimmen wir einen Wahrscheinlichkeitsraum $(\Omega', \mathcal{A}', \mathbf{P}_X)$, der das Zufallsexperiment mit Ausgang $X((x_1, \ldots, x_n))$ beschreibt.

Dabei setzen wir $\Omega' = \mathbb{N}_0$ und $\mathcal{A}' = \mathcal{P}(\mathbb{N}_0)$ und bestimmen die Wahrscheinlichkeit $\mathbf{P}_X(\{k\})$, dass $X(\omega)$ den Wert k annimmt, gemäß

$$\mathbf{P}_X(\{k\}) = \mathbf{P}(\{\omega \in \Omega : X(\omega) = k\}).$$

Für $k > n$ gilt dann

$$\mathbf{P}_X(\{k\}) = \mathbf{P}(\{(x_1, \ldots, x_n) \in \{0,1\}^n \ : \ x_1 + \cdots + x_n = k\}) = \mathbf{P}(\emptyset) = 0,$$

während für $0 \leq k \leq n$ gilt:

$$\mathbf{P}_X(\{k\}) = \mathbf{P}\left(\{(x_1, \ldots, x_n) \in \{0,1\}^n \ : \ x_1 + \cdots + x_n = k\}\right).$$

Es gibt $\binom{n}{k}$ n–Tupel $(x_1, \ldots, x_n) \in \{0,1\}^n$ mit $x_1 + \cdots + x_n = k$, für jedes dieser n–Tupel gilt

$$\mathbf{P}(\{(x_1, \ldots, x_n)\}) = p^{\sum_{i=1}^n x_i} \cdot (1-p)^{n - \sum_{i=1}^n x_i} = p^k \cdot (1-p)^{n-k},$$

womit folgt

$$\mathbf{P}_X(\{k\}) = \sum_{\substack{(x_1, \ldots, x_n) \in \{0,1\}^n, \\ x_1 + \cdots + x_n = k}} \mathbf{P}(\{(x_1, \ldots, x_n)\}) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

$\mathbf{P}_X$ ist also eine Binomialverteilung mit den Parametern n und p.

Beispiel 4.18 *Um seinen aufwendigen Lebensstil zu finanzieren, beschließt Student S., seinen Lebensunterhalt durch Betreiben eines Glücksrads auf dem Cannstatter Volksfest aufzubessern.*

Nach Drehen bleibt dieses rein zufällig auf einem von 64 Feldern stehen. Bleibt es auf einem der fünf braun gefärbten Felder stehen, so erhält der Spieler einen Mohrenkopf (Wert 20 Cent). Bleibt es auf einem der beiden rot gefärbten Felder stehen, so erhält der Spieler eine rote Rose (Wert 3 Euro). Und bleibt es auf dem einzigen schwarzen Feld stehen, so erhält der Spieler das Buch Statistik - Der Weg zur Datenanalyse von Fahrmeir, Künstler, Pigeot und Tutz, Springer 2001 (Wert ca. 25 Euro). Auf den 56 übrigen weißen Feldern wird kein Gewinn ausgegeben.

Gesucht ist eine Möglichkeit zur stochastischen Modellierung des (zufälligen) Wertes des Gewinns.

Der zufällige Gewinn X nimmt nur endlich viele Werte an, nämlich nur die Werte $0, 20, 300$ und 2500 (in Cent). Wir bestimmen

$$\mathbf{P}[X = x]$$

für jeden dieser Werte:

Für $x = 300$: X nimmt den Wert 300 an, wenn das Glücksrad auf einem der 2 roten Felder stehenbleibt. Dass genau eines der beiden roten Felder von den insgesamt 64 Feldern auftritt, kommt mit Wk. 2/64 vor. Daher gilt:

$$\mathbf{P}[X = 300] = \frac{2}{64}.$$

Analog bestimmt man

$$\mathbf{P}[X = 0] = \frac{56}{64}, \mathbf{P}[X = 20] = \frac{5}{64} \text{ und } \mathbf{P}[X = 2500] = \frac{1}{64}.$$

Für alle anderen Werte $x \in \mathbb{R}$ gilt $\mathbf{P}[X = x] = 0$. Wir setzen dann

$$\mathbf{P}[X \in B] = \sum_{k \in \{0,20,300,2500\} \cap B} \mathbf{P}[X = k]$$

für $B \subseteq \mathbb{R}$.

Formal ist X eine *Zufallsvariable*, die definiert werden kann wie folgt:

Wir beschreiben das Drehen am Glücksrad durch einen Laplaceschen W-Raum $(\Omega, \mathcal{A}, \mathbf{P})$ mit

$$\Omega = \{1, 2, \ldots, 64\},$$
$$\mathcal{A} = \mathcal{P}(\Omega)$$

und

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

Hierbei ist $\omega \in \Omega$ die Nummer des Feldes, auf dem das Glücksrad stehenbleibt. Die Felder 1 bis 5 seien braun, die Felder 6 und 7 seien rot, Feld 8 sei schwarz und die Felder 9 bis 64 seien weiß.

Der bei Auftreten von Feld ω ausgezahlte Gewinn ist gegeben durch

$$X(\omega) = \begin{cases} 20 & \text{für} \quad \omega \in \{1, \ldots, 5\} \\ 300 & \text{für} \quad \omega \in \{6, 7\} \\ 2500 & \text{für} \quad \omega = 8 \\ 0 & \text{für} \quad \omega \in \{9, 10, \ldots, 64\}. \end{cases}$$

Die Wahrscheinlichkeit, dass $X(\omega)$ in einer Menge $B \subseteq \mathbb{R}$ landet, wird dann festgelegt gemäß

$$\mathbf{P}[X \in B] := \mathbf{P}(\{\omega \in \Omega \ : \ X(\omega) \in B\})$$

Speziell gilt:

$$\mathbf{P}[X = 20] := \mathbf{P}[X \in \{20\}] = \mathbf{P}(\{\omega \in \Omega \ : \ X(\omega) \in \{20\}\}) = \mathbf{P}(\{1, 2, 3, 4, 5\}) = \frac{5}{64},$$

Analog erhält man

$$\begin{aligned} \mathbf{P}[X = 300] &= \mathbf{P}(\{6, 7\}) = \frac{2}{64} \\ \mathbf{P}[X = 2500] &= \mathbf{P}(\{8\}) = \frac{1}{64} \\ \mathbf{P}[X = 0] &= \mathbf{P}(\{9, 10, \ldots, 64\}) = \frac{56}{64}. \end{aligned}$$

Darüberhinaus gilt für $B \subseteq \mathbb{R}$:

$$\begin{aligned} \mathbf{P}[X \in B] &= \mathbf{P}(\{\omega \in \Omega : X(\omega) \in B\}) \\ &= \mathbf{P}\left(\{\omega \in \Omega : X(\omega) \in \{0, 20, 300, 2500\} \cap B\}\right) \\ &= \sum_{k \in \{0,20,300,2500\} \cap B} \mathbf{P}\left(\{\omega \in \Omega : X(\omega) = k\}\right) \\ &= \sum_{k \in \{0,20,300,2500\} \cap B} \mathbf{P}[X = k]. \end{aligned}$$

Wie der folgende Satz zeigt, hat die Zuweisung (4.2) von Wahrscheinlichkeiten zu Mengen immer die Eigenschaften, die wir für Wahrscheinlichkeitsmaße gefordert haben.

Satz 4.5 *Sei* $(\Omega, \mathcal{A}, \mathbf{P})$ *ein W–Raum,* $(\Omega', \mathcal{A}')$ *ein Messraum und* $X : \Omega \to \Omega'$ *eine Zufallsvariable. Dann wird durch*

$$\mathbf{P}_X(A') := \mathbf{P}(X^{-1}(A')) = \mathbf{P}\left(\{\omega \in \Omega \ : \ X(\omega) \in A'\}\right) \quad (A' \in \mathcal{A}')$$

ein W–Raum $(\Omega', \mathcal{A}', \mathbf{P}_X)$ *definiert.*

Beweis: Da X Zufallsvariable ist, gilt $X^{-1}(A') \in \mathcal{A}$ für alle $A' \in \mathcal{A}'$, und daher ist $\mathbf{P}_X$ wohldefiniert. Weiter gilt wegen $\mathbf{P}$ W-Maß

$$\mathbf{P}_X(A') = \mathbf{P}(X^{-1}(A')) \geq 0$$

für alle $A' \in \mathcal{A}'$, sowie

$$\mathbf{P}_X(\Omega') = \mathbf{P}(X^{-1}(\Omega')) = \mathbf{P}(\{\omega \in \Omega : X(\omega) \in \Omega'\}) = \mathbf{P}(\Omega) = 1.$$

Sind darüberhinaus $A'_1, A'_2, \ldots \in \mathcal{A}'$ paarweise disjunkt (d.h. gilt $A'_i \cap A'_j = \emptyset$ für $i \neq j$), so sind auch $X^{-1}(A'_1), X^{-1}(A'_2), \ldots \in \mathcal{A}$ paarweise disjunkt, denn aus

$$\begin{aligned} \omega \in X^{-1}(A'_i) \cap X^{-1}(A'_j) &\Leftrightarrow \omega \in X^{-1}(A'_i) \text{ und } \omega \in X^{-1}(A'_j) \\ &\Leftrightarrow X(\omega) \in A'_i \text{ und } X(\omega) \in A'_j \\ &\Leftrightarrow X(\omega) \in A'_i \cap A'_j \end{aligned}$$

folgt $X^{-1}(A'_i) \cap X^{-1}(A'_j) = \emptyset$ für $i \neq j$. Beachtet man darüberhinaus

$$\begin{aligned} \omega \in X^{-1}(\cup_{n=1}^{\infty} A'_n) &\Leftrightarrow X(\omega) \in \cup_{n=1}^{\infty} A'_n \\ &\Leftrightarrow \exists n \in \mathbb{N} : X(\omega) \in A'_n \\ &\Leftrightarrow \exists n \in \mathbb{N} : \omega \in X^{-1}(A'_n) \\ &\Leftrightarrow \omega \in \cup_{n=1}^{\infty} X^{-1}(A'_n), \end{aligned}$$

woraus

$$X^{-1}(\cup_{n=1}^{\infty} A'_n) = \cup_{n=1}^{\infty} X^{-1}(A'_n)$$

folgt, so erhält man aufgrund der σ-Additivität des W-Maßes $\mathbf{P}$:

$$\begin{aligned} &\mathbf{P}_X\left(\cup_{n=1}^{\infty} A'_n\right) = \mathbf{P}\left(X^{-1}\left(\cup_{n=1}^{\infty} A'_n\right)\right) = \mathbf{P}\left(\cup_{n=1}^{\infty} X^{-1}\left(A'_n\right)\right) \\ &= \sum_{n=1}^{\infty} \mathbf{P}\left(X^{-1}\left(A'_n\right)\right) = \sum_{n=1}^{\infty} \mathbf{P}_X\left(A'_n\right). \end{aligned}$$

Mit Lemma 4.1 folgt die Behauptung. □

Für das in Satz 4.5 eingeführte W–Maß ist die folgende Bezeichnung üblich:

Definition 4.20 *Das in in Satz 4.5 eingeführte W–Maß* $\mathbf{P}_X$ *heißt* **Verteilung** *der Zufallsvariablen* X.

Bemerkung: Sei $(\Omega, \mathcal{A}, \mathbf{P})$ W–Raum. Dann ist $\mathbf{P}$ Verteilung der Zufallsvariablen

$$Y : \Omega \to \Omega, \quad Y(\omega) = \omega.$$

Jedes W–Maß kann also als Verteilung einer geeigneten Zufallsvariablen aufgefasst werden. Daher ist es üblich, die Begriffe W–Maß und Verteilung synonym zu verwenden.

Im Folgenden werden die bisher eingeführten Bezeichnungen auf Zufallsvariablen übertragen. Dem Begriff *W–Maß mit Zähldichte* entspricht der Begriff *diskrete Zufallsvariable.*

Definition 4.21 *Sei* X *eine reelle Zufallsvariable. Dann heißt* X **diskrete Zufallsvariable**, *falls für eine endliche oder abzählbar unendliche Menge* $A \subseteq \mathbb{R}$ *gilt:*

$$\mathbf{P}[X \in A] = 1,$$

d.h. falls X *mit Wahrscheinlichkeit Eins nur Werte aus einer endlichen oder abzählbar unendlichen Menge annimmt.*

Definition 4.22 *Sei* X *eine diskrete Zufallsvariable mit Werten* $x_1, x_2, \ldots$ *bzw. mit Werten* $x_1, \ldots, x_N$. *Dann heißt*

$$(\mathbf{P}[X = x_k])_{k \in \mathbb{N}} \quad bzw. \quad (\mathbf{P}[X = x_k])_{k=1,\ldots,N}$$

Zähldichte *von* X.

Beispiele für diskrete Zufallsvariablen:

1. Seien $n \in \mathbb{N}$ und $p \in [0, 1]$. Eine reelle Zufallsvariable X mit

$$\mathbf{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k} \quad (k \in \{0, \ldots, n\})$$

heißt **binomialverteilt mit Parametern n und p** (kurz: $b(n, p)$-verteilt).

Hierbei gilt:

$$\mathbf{P}[X \in \{0, \ldots, n\}] = \sum_{k=0}^{n} \mathbf{P}[X = k] = (p + (1-p))^n = 1$$

und

$$\mathbf{P}[X \in \mathbb{R} \setminus \{0, \ldots, n\}] = 1 - \mathbf{P}[X \in \{0, \ldots, n\}] = 0.$$

2. Sei $\lambda \in \mathbb{R}_+$. Eine reelle Zufallsvariable X mit

$$\mathbf{P}[X = k] = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

heißt **Poisson-verteilt mit Parameter** λ (kurz: $\pi(\lambda)$-verteilt).

Hierbei gilt:

$$\mathbf{P}[X \in \mathbb{N}_0] = \sum_{k=0}^{\infty} \mathbf{P}[X = k] = e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1$$

und

$$\mathbf{P}[X \in \mathbb{R} \setminus \mathbb{N}_0] = 1 - \mathbf{P}[X \in \mathbb{N}_0] = 0.$$

Als nächstes übertragen wir den Begriff *W-Maß mit Dichte* auf Zufallsvariablen.

Definition 4.23 *Sei X eine reelle Zufallsvariable und sei $f : \mathbb{R} \to \mathbb{R}_+$ eine Funktion mit $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Dann heißt X* **stetig verteilte Zufallsvariable mit Dichte f**, *falls gilt*

$$\mathbf{P}[X \in B] = \int_B f(x)\,dx \quad (B \in \mathcal{B}).$$

In diesem Fall heißt f **Dichte** *von X bzw. von $\mathbf{P}_X$.*

Beispiele für stetig verteilte Zufallsvariablen:

1. Seien $a, b \in \mathbb{R}$ mit $a < b$ und sei $f : \mathbb{R} \to \mathbb{R}_+$ definiert durch

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für} \quad a \leq x \leq b, \\ 0 & \text{für} \quad x < a \text{ oder } x > b. \end{cases}$$

Eine reelle Zufallsvariable X mit

$$\mathbf{P}[X \in B] = \int_B f(x)\,dx \quad (B \in \mathcal{B})$$

heißt **gleichverteilt auf** $[a, b]$ (kurz: $U([a, b])$-verteilt).

2. Sei $\lambda \in \mathbb{R}_+$ und sei $f : \mathbb{R} \to \mathbb{R}_+$ definiert durch

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & \text{für} \quad x \geq 0, \\ 0 & \text{für} \quad x < 0. \end{cases}$$

Eine reelle Zufallsvariable X mit

$$\mathbf{P}[X \in B] = \int_B f(x)\,dx \quad (B \in \mathcal{B}).$$

heißt **exponential-verteilt mit Parameter** λ (kurz: $exp(\lambda)$-verteilt).

3. Seien $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$ und sei $f : \mathbb{R} \to \mathbb{R}_+$ definiert durch
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (x \in \mathbb{R}).$$
Eine reelle Zufallsvariable X mit
$$\mathbf{P}[X \in B] = \int_B f(x)\,dx \quad (B \in \mathcal{B})$$
heißt **normalverteilt mit Parametern μ und σ^2** (kurz: $N(\mu, \sigma^2)$-verteilt).

Als nächstes übertragen wir den Begriff der Unabhängigkeit auf Zufallsvariablen.

Definition 4.24 *Sei $(\Omega, \mathcal{A}, \mathbf{P})$ ein W–Raum, seien $(\Omega_i, \mathcal{A}_i)$ $(i = 1, \dots, n)$ Messräume und seien*
$$X_i : \Omega \to \Omega_i \quad (i = 1, \dots, n)$$
Zufallsvariablen. Dann heißen $X_1, \dots, X_n$ **unabhängig**, *falls für alle $A_i \in \mathcal{A}_i$ $(i = 1, \dots, n)$ gilt:*
$$\mathbf{P}\left[X_1 \in A_1, \dots, X_n \in A_n\right] = \mathbf{P}\left[X_1 \in A_1\right] \cdots \mathbf{P}\left[X_n \in A_n\right].$$
Eine Folge $(X_n)_{n \in \mathbb{N}}$ von Zufallsvariablen heißt unabhängig, falls $X_1, \dots, X_n$ unabhängig sind für jedes $n \in \mathbb{N}$.

Bemerkung:

a) In der obigen Definition wurden die Schreibweisen
$$\begin{aligned}&\mathbf{P}\left[X_1 \in A_1, \dots, X_n \in A_n\right] \\ &:= \mathbf{P}\left(\{\omega \in \Omega \,:\, X_1(\omega) \in A_1, \dots, X_n(\omega) \in A_n\}\right) \\ &= \mathbf{P}\left(\{\omega \in \Omega \,:\, X_1(\omega) \in A_1\} \cap \dots \cap \{\omega \in \Omega \,:\, X_n(\omega) \in A_n\}\right)\end{aligned}$$
und
$$\mathbf{P}\left[X_i \in A_i\right] := \mathbf{P}\left(\{\omega \in \Omega \,:\, X_i(\omega) \in A_i\}\right)$$
verwendet. Formal sind Ausdrücke wie $X \in A$, wobei X eine Abbildung und A eine Menge von Zahlen ist, natürlich unsinnig. Da sie aber sowohl üblich als auch sehr suggestiv sind, werden sie im Folgenden vielfach verwendet.

b) Sind $X_1, \dots, X_n$ unabhängig, so ist die Wahrscheinlichkeit, dass alle Zufallsvariablen gleichzeitig gewisse Bedingungen erfüllen, gleich dem Produkt der Einzelwahrscheinlichkeiten.

c) Die obige Definition der Unabhängigkeit ist für n-Tupel von Zufallsvariablen mit $n > 2$ etwas einfacher als die entsprechende Definition für Ereignisse, da oben einige A_i gleich Ω_i gesetzt werden können und damit endliche Teilmengen der Indexmenge automatisch mit erfasst werden.

Definition 4.25 *Sei X eine reelle Zufallsvariable. Dann heißt die durch*

$$F : \mathbb{R} \to \mathbb{R}, \quad F(x) := \mathbf{P}[X \leq x] := \mathbf{P}_X\left((-\infty, x]\right)$$

definierte Funktion die **Verteilungsfunktion** *(kurz: Vf) der Zufallsvariablen X (bzw. des W–Maßes $\mathbf{P}_X$).*

Bemerkung: Durch die Verteilungsfunktion sind die Werte von $\mathbf{P}_X$ für alle Intervalle $(a, b]$ $(a, b \in \mathbb{R},\ a < b)$ festgelegt:

$$\begin{aligned}
\mathbf{P}_X\left((a, b]\right) &= \mathbf{P}_X\left((-\infty, b] \setminus (-\infty, a]\right) \\
&= \mathbf{P}_X\left((-\infty, b]\right) - \mathbf{P}_X\left((-\infty, a]\right) \\
&= F(b) - F(a).
\end{aligned}$$

Man kann zeigen, dass dadurch sogar das gesamte W–Maß $\mathbf{P}_X : \mathcal{B} \to \mathbb{R}$ festgelegt ist (!)

Beispiel 4.19 *Sei X eine* $\exp(\lambda)$*–verteilte ZV, d.h.,*

$$\mathbf{P}_X(A) = \int_A f(x)\, dx \quad \textit{mit} \quad f(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & \textit{für} \quad x \geq 0, \\ 0 & \textit{für} \quad x < 0, \end{cases}$$

wobei $\lambda > 0$. Dann gilt für die Verteilungsfunktion F von X:

$$\begin{aligned}
F(x) &= \mathbf{P}_X((-\infty, x]) = \int_{(-\infty, x]} f(u)\, du \\
&= \begin{cases} \int_0^x \lambda \cdot e^{-\lambda \cdot u} du = 1 - e^{-\lambda \cdot x} & \textit{für} \quad x \geq 0, \\ 0 & \textit{für} \quad x < 0. \end{cases}
\end{aligned}$$

Satz 4.6 *(Eigenschaften der Verteilungsfunktion) Sei F die Verteilungsfunktion einer reellen Zufallsvariablen X auf einem W–Raum $(\Omega, \mathcal{A}, \mathbf{P})$. Dann gilt:*

a) *$F(x) \in [0, 1]$ für alle $x \in \mathbb{R}$,*

b) *F ist monoton nichtfallend, d.h. aus $x_1 \leq x_2$ folgt $F(x_1) \leq F(x_2)$,*

c) $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to-\infty} F(x) = 0$,

d) *F ist rechtsseitig stetig, d.h.*

$$\lim_{\substack{y\to x\\ y>x}} F(y) = F(x)$$

für alle $x \in \mathbb{R}$.

Zum Beweis von Satz 4.6 benötigen wir das folgende Lemma.

Lemma 4.6 *Sei $(\Omega, \mathcal{A}, \mathbf{P})$ ein beliebiger W–Raum.*

a) *Für alle $A, A_n \in \mathcal{A}$ $(n \in \mathbb{N})$ mit*

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots \quad \textit{und} \quad \bigcup_{n=1}^{\infty} A_n = A$$

gilt

$$\lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}(A)$$

(sog. Stetigkeit von unten des W–Maßes $\mathbf{P}$).

b) *Für alle $A, A_n \in \mathcal{A}$ $(n \in \mathbb{N})$ mit*

$$A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots \quad \textit{und} \quad \bigcap_{n=1}^{\infty} A_n = A$$

gilt

$$\lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}(A)$$

(sog. Stetigkeit von oben des W–Maßes $\mathbf{P}$).

Beweis. a) *Nachweis der Stetigkeit von unten:* Wir zeigen,

$$\lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}(A),$$

indem wir beide Seiten separat umformen.

Zur Umformung der linken Seite stellen wir die Menge A_N dar als

$$A_N = A_1 \cup \bigcup_{n=2}^{N} (A_n \setminus A_{n-1}).$$

Dabei haben die Mengen A_1, $A_2 \setminus A_1$, $\dots$, $A_N \setminus A_{N-1}$ paarweise leeren Schnitt.

Mit der σ–Additivität von $\mathbf{P}$ folgt:

$$\mathbf{P}(A_N) = \mathbf{P}\left(A_1 \cup \bigcup_{n=2}^{N}(A_n \setminus A_{n-1})\right) = \mathbf{P}(A_1) + \sum_{n=2}^{N}\mathbf{P}(A_n \setminus A_{n-1})$$

und damit

$$\begin{aligned}\lim_{N\to\infty}\mathbf{P}(A_N) &= \lim_{N\to\infty}\left(\mathbf{P}(A_1) + \sum_{n=2}^{N}\mathbf{P}(A_n \setminus A_{n-1})\right)\\ &= \mathbf{P}(A_1) + \lim_{N\to\infty}\sum_{n=2}^{N}\mathbf{P}(A_n \setminus A_{n-1})\\ &= \mathbf{P}(A_1) + \sum_{n=2}^{\infty}\mathbf{P}(A_n \setminus A_{n-1})\end{aligned}$$

Zur Umformung der rechten Seite stellen wir die Menge $\cup_{n=1}^{\infty}A_n$ dar als

$$\cup_{n=1}^{\infty}A_n = A_1 \cup \bigcup_{n=2}^{\infty}(A_n \setminus A_{n-1}).$$

Dabei haben die Mengen A_1, $A_2 \setminus A_1$, $A_3 \setminus A_2$, ... wieder paarweise leeren Schnitt.

Mit der σ–Additivität von $\mathbf{P}$ folgt:

$$\mathbf{P}\left(\cup_{n=1}^{\infty}A_n\right) = \mathbf{P}\left(A_1 \cup \bigcup_{n=2}^{\infty}(A_n \setminus A_{n-1})\right) = \mathbf{P}(A_1) + \sum_{n=2}^{\infty}\mathbf{P}(A_n \setminus A_{n-1})$$

Dies impliziert die Behauptung.

b) *Nachweis der Stetigkeit von oben:*

Es gilt:

$$\Omega \setminus A_1 \subseteq \Omega \setminus A_2 \subseteq \Omega \setminus A_3 \subseteq \ldots$$

und

$$\cup_{n=1}^{\infty}\Omega \setminus A_n = \Omega \setminus \left(\cap_{n=1}^{\infty}A_n\right) = \Omega \setminus A.$$

Anwendung der Stetigkeit von unten ergibt:

$$\lim_{n\to\infty}\mathbf{P}\left(\Omega \setminus A_n\right) = \mathbf{P}\left(\Omega \setminus A\right).$$

Mit

$$\mathbf{P}\left(\Omega \setminus A_n\right) = 1 - \mathbf{P}\left(A_n\right) \quad \text{und} \quad \mathbf{P}\left(\Omega \setminus A\right) = 1 - \mathbf{P}\left(A\right)$$

folgt

$$\lim_{n\to\infty} (1 - \mathbf{P}(A_n)) = 1 - \mathbf{P}(A),$$

also

$$\lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}(A).$$

□

Beweis von Satz 4.6.

a) Da $\mathbf{P}_X$ W-Maß ist, gilt

$$F(x) = \mathbf{P}[X \leq x] = \mathbf{P}_X((-\infty, x]) \in [0,1].$$

b) Für $x_1 \leq x_2$ gilt $(-\infty, x_1] \subseteq (-\infty, x_2]$, und dies wiederum impliziert

$$F(x_1) = \mathbf{P}_X((-\infty, x_1]) \leq \mathbf{P}_X((-\infty, x_2]) = F(x_2).$$

c$_1$) *Nachweis von* $\lim_{x\to\infty} F(x) = 1$*:*

Sei $(x_n)_n$ eine beliebige monoton wachsende Folge reeller Zahlen mit $x_n \to \infty$ $(n \to \infty)$. Dann gilt

$$(-\infty, x_1] \subseteq (-\infty, x_2] \subseteq \dots \quad \text{und} \quad \cup_{n=1}^{\infty} (-\infty, x_n] = \mathbb{R},$$

und mit der Stetigkeit von unten des W-Maßes $\mathbf{P}_X$ folgt

$$\lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} \mathbf{P}_X((-\infty, x_n]) = \mathbf{P}_X(\mathbb{R}) = 1.$$

Aufgrund der Monotonie von F folgt daraus die Behauptung.

c$_2$) *Nachweis von* $\lim_{x\to-\infty} F(x) = 0$*:*

Sei $(x_n)_n$ eine beliebige monoton fallende Folge reeller Zahlen mit $x_n \to -\infty$ $(n \to \infty)$. Dann gilt

$$(-\infty, x_1] \supseteq (-\infty, x_2] \supseteq \dots \quad \text{und} \quad \cap_{n=1}^{\infty} (-\infty, x_n] = \emptyset$$

und mit der Stetigkeit von oben des W-Maßes $\mathbf{P}_X$ folgt

$$\lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} \mathbf{P}_X((-\infty, x_n]) = \mathbf{P}_X(\emptyset) = 0.$$

Aufgrund der Monotonie von F folgt daraus die Behauptung.

d) *Nachweis von* $\lim_{y\to x, y>x} F(y) = F(x)$*:*

Sei $(x_n)_n$ eine beliebige monoton fallende Folge reeller Zahlen mit $x_n \to x$ $(n \to \infty)$. Dann gilt

$$(-\infty, x_1] \supseteq (-\infty, x_2] \supseteq \dots \quad \text{und} \quad \cap_{n=1}^{\infty} (-\infty, x_n] = (-\infty, x]$$

und mit der Stetigkeit von oben des W-Maßes $\mathbf{P}_X$ folgt

$$\lim_{n\to\infty} F(x_n) \;=\; \lim_{n\to\infty} \mathbf{P}_X\left((-\infty, x_n]\right) = \mathbf{P}_X\left((-\infty, x]\right) = F(x).$$

Aufgrund der Monotonie von F folgt daraus die Behauptung. $\square$

Beispiel 4.20 *Die zufällige Lebensdauer X der Batterie eines Computers sei* $\exp(\lambda)$*–verteilt. Um die Wahrscheinlichkeit eines plötzlichen Ausfalls des Rechners zu verringern wird diese spätestens nach einer festen Zeit* $t > 0$ *ausgetauscht, d.h., für die Betriebszeit Y der Batterie gilt*

$$Y(\omega) = \min\{X(\omega), t\} \quad (\omega \in \Omega).$$

Zu ermitteln ist die Verteilungsfunktion G von Y.

Wegen

$$\min\{X(\omega), t\} \le y \quad \Leftrightarrow \quad X(\omega) \le y \text{ oder } t \le y$$

gilt

$$\begin{aligned}
&G(y)\\
&= \mathbf{P}_Y((-\infty, y]) = \mathbf{P}[\min\{X, t\} \le y]\\
&= \mathbf{P}\left(\{\omega \in \Omega : \min\{X(\omega), t\} \le y\}\right)\\
&= \begin{cases} \mathbf{P}(\Omega) = 1 & \text{für} \quad y \ge t,\\ \mathbf{P}\left(\{\omega \in \Omega : X(\omega) \le y\}\right) = \mathbf{P}[X \le y] = 1 - e^{-\lambda \cdot y} & \text{für} \quad 0 \le y < t,\\ \mathbf{P}(\emptyset) = 0 & \text{für} \quad y < 0. \end{cases}
\end{aligned}$$

4.6 Erwartungswert

Sei X eine reelle Zufallsvariable. Im Folgenden wird festgelegt, was man unter dem "mittleren Wert" des Ergebnisses $X(\omega)$ des zugehörigen Zufallsexperiments versteht.

Dieser Begriff ist in vielen Anwendungen von zentraler Bedeutung. Z.B. wird oft versucht, einen möglichst hohen (zufälligen) Gewinn zu erzielen, indem man den "mittleren Gewinn" (bei Versendung von Werbung, Vergabe von Krediten, Kauf von Aktien, etc.) optimiert.

Das weitere Vorgehen wird mit Hilfe der drei folgenden hypothetischen Beispiele illustriert werden:

Beispiel 4.21 *Ein "echter" Würfel wird so lange geworfen, bis er zum ersten Mal mit* 6 *oben landet.*

Wie oft wird der Würfel dann "im Mittel" geworfen ?

Beispiel 4.21 kann durch eine diskrete Zufallsvariable mit Werten in $\mathbb{N}$ beschrieben werden.

Beispiel 4.22 *Dozent K. fährt nach seiner Statistik Vorlesung immer mit der S-Bahn nach Vaihingen. Diese fährt alle 10 Minuten. Da Dozent K. sich die genauen Abfahrtszeiten nicht merken kann, trifft er rein zufällig innerhalb eines zehnminütigen Intervalls zwischen zwei aufeinanderfolgenden Abfahrtszeiten am Bahnhof ein.*

Wie lange muss Dozent K. dann "im Mittel" warten ?

Die Wartezeit in Beispiel 4.22 ist rein zufällig im Intervall $[0, 10]$ verteilt und wird daher durch eine auf $[0, 10]$ gleichverteilte Zufallsvariable, d.h. durch eine stetig verteilte Zufallsvariable mit Dichte, beschrieben.

Beispiel 4.23 *Student S. fährt immer mit dem Auto zur Uni. Dabei passiert er eine Ampelanlage, bei der sich eine einminütige Grünphase mit einer zweiminütigen Rotphase abwechselt.*

Wie lange wartet er "im Mittel", wenn seine Ankunft an der Ampel rein zufällig innerhalb eines dreiminütigen Intervalls, bestehend aus Grün- und Rotphase, erfolgt ?

Die Zufallsvariable X, die die zufällige Wartezeit an der Ampel beschreibt, ist weder diskret verteilt noch stetig verteilt mit Dichte (denn aus letzterem würde folgen:

$$\mathbf{P}[X = 0] \leq \mathbf{P}[X \in (-\epsilon, \epsilon)] = \int_{-\epsilon}^{\epsilon} f(x)\, dx \to 0 \quad (\epsilon \to 0),$$

was im Widerspruch steht zu $\mathbf{P}[X = 0] = 1/3$).

4.6.1 Diskrete Zufallsvariablen

Sei X eine diskrete Zufallsvariable mit Werten $x_1, x_2, \dots \in \mathbb{R}$. Dann ist es naheliegend, als "mittleren Wert" von X das mit $\mathbf{P}[X = x_k]$ gewichtete Mittel der Zahlen x_k zu wählen:

Definition 4.26 *Sei X eine diskrete Zufallsvariable mit Werten $x_1, x_2, \dots \in \mathbb{R}$. Dann heißt*

$$\mathbf{E}X = \sum_{k=1}^{\infty} x_k \cdot \mathbf{P}[X = x_k]$$

– sofern existent – der **Erwartungswert** *von X.*

Anwendung im Beispiel 4.21: Für die zufällige Anzahl X der Würfe des Würfels in Beispiel 4.21 gilt

$$\begin{aligned} \mathbf{P}[X = k] &= \mathbf{P}\big[1.\,Augenzahl \in \{1,\dots,5\}, \dots, (k-1)te\,Augenzahl \in \{1,\dots,5\}, \\ &\qquad\qquad k-te\,Augenzahl = 6\big] \\ &= \left(\frac{5}{6}\right)^{k-1} \cdot \frac{1}{6}. \end{aligned}$$

Damit erhält man

$$\mathbf{E}X = \sum_{k=1}^{\infty} k \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{k-1} = \frac{1}{6} \cdot \frac{d}{dx} \sum_{k=0}^{\infty} x^k \Big|_{x=5/6} = \frac{1}{6} \cdot \frac{1}{(1-x)^2}\Big|_{x=5/6} = 6.$$

Beispiel 4.24 *Wir betrachten nochmals das Glücksrad aus Beispiel 4.18. Bestimmen möchten wir den "mittlere Gewinn" (Erwartungswert) beim Drehen an diesem Glücksrad.*

Der zufällige Gewinn X nimmt hier nur die Werte 0, 20, 300 und 2500 an, und zwar mit den in Beispiel 4.18 bestimmten Wahrscheinlichkeiten $\mathbf{P}[X = 0] = 56/64$, $\mathbf{P}[X = 20] = 5/64$, $\mathbf{P}[X = 300] = 2/64$ und $\mathbf{P}[X = 2500] = 1/64$.

Damit ergibt sich der mittlerer Wert (*Erwartungswert*) des zufälligen Gewinns X als

$$\begin{aligned} \mathbf{E}X &= 0 \cdot \mathbf{P}[X = 0] + 20 \cdot \mathbf{P}[X = 20] + 300 \cdot \mathbf{P}[X = 300] + 2500 \cdot \mathbf{P}[X = 2500] \\ &= 0 \cdot \frac{56}{64} + 20 \cdot \frac{5}{64} + 300 \cdot \frac{2}{64} + 2500 \cdot \frac{1}{64} \\ &= 50. \end{aligned}$$

Beispiel 4.25 *X sei eine $b(n,p)$–verteilte Zufallsvariable ($n \in \mathbb{N}, p \in [0,1]$), d.h.*

$$\mathbf{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k} \quad (k \in \{0, \ldots, n\}).$$

Dann gilt

$$\begin{aligned}
\mathbf{E}X &= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} k \cdot \frac{n}{k} \binom{n-1}{k-1} p^k (1-p)^{n-k} \\
&= n \cdot p \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= n \cdot p \cdot (p + (1-p))^{n-1} \\
&= n \cdot p.
\end{aligned}$$

Beispiel 4.26 *X sei eine $\pi(\lambda)$–verteilte Zufallsvariable ($\lambda > 0$), d.h.*

$$\mathbf{P}[X = k] = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad (k \in \mathbb{N}_0).$$

Dann gilt

$$\mathbf{E}X = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} = \lambda \cdot \left(\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) \cdot e^{-\lambda} = \lambda \cdot e^{\lambda} \cdot e^{-\lambda} = \lambda.$$

4.6.2 Stetig verteilte Zufallsvariablen

Im Falle einer stetig verteilten Zufallsvariablen X mit Dichte f ersetzt man die Summe in der obigen Definition durch das entsprechende Integral:

Definition 4.27 *Sei X eine stetig verteilte Zufallsvariable mit Dichte f. Dann heißt*

$$\mathbf{E}X = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

– sofern existent – der **Erwartungswert** *von X.*

Anwendung im Beispiel 4.22: Die zufällige Wartezeit auf die S-Bahn in Beispiel 4.22 wird durch eine auf $[0,10]$ gleichverteilte Zufallsvariable X beschrieben, d.h. durch eine stetig verteilte Zufallsvariable mit Dichte

$$f(x) = \begin{cases} \frac{1}{10} & \text{für} \quad 0 \leq x \leq 10, \\ 0 & \text{für} \quad x < 0 \text{ oder } x > 10. \end{cases}$$

Damit folgt für die mittlere Wartezeit:

$$\mathbf{E}X = \int_{\mathbb{R}} x \cdot f(x)\, dx = \int_0^{10} x \cdot \frac{1}{10}\, dx = \left.\frac{x^2}{20}\right|_{x=0}^{10} = 5.$$

Beispiel 4.27 *X sei eine $\exp(\lambda)$–verteilte Zufallsvariable, d.h.*

$$\mathbf{P}_X(A) = \int_A f(x)\, dx \quad \textit{mit} \quad f(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & \textit{für} \quad x \geq 0, \\ 0 & \textit{für} \quad x < 0, \end{cases}$$

wobei $\lambda > 0$. Dann gilt

$$\begin{aligned} \mathbf{E}X &= \int_0^\infty x \cdot \lambda \cdot e^{-\lambda \cdot x} dx = -x \cdot e^{-\lambda \cdot x}\Big|_{x=0}^\infty + \int_0^\infty e^{-\lambda \cdot x} dx = 0 - \frac{1}{\lambda} \cdot e^{-\lambda \cdot x}\Big|_{x=0}^\infty \\ &= \frac{1}{\lambda}. \end{aligned}$$

Beispiel 4.28 *X sei eine $N(a, \sigma^2)$–verteilte Zufallsvariable, d.h.*

$$\mathbf{P}_X(A) = \int_A f(x)\, dx \quad \textit{mit} \quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-(x-a)^2/(2\sigma^2)}.$$

Dann gilt

$$\begin{aligned} \mathbf{E}X &= \int_{-\infty}^\infty x \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-(x-a)^2/(2\sigma^2)}\, dx \\ &= \int_{-\infty}^\infty \frac{x-a}{\sqrt{2\pi}\sigma} \cdot e^{-(x-a)^2/(2\sigma^2)}\, dx + a \cdot \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-(x-a)^2/(2\sigma^2)}\, dx \\ &= 0 + a = a. \end{aligned}$$

Dabei wurde beim dritten Gleichheitszeichen ausgenutzt, dass der erste Integrand punktsymmetrisch bezüglich $x = a$ ist, und dass beim zweiten Integral über eine Dichte integriert wird.

4.6.3 Berechnung allgemeinerer Erwartungswerte

Wie aus Beispiel 4.23 ersichtlich wird, reichen die bisher behandelten Spezialfälle nicht aus. Im nächsten Unterabschnitt wird eine wesentlich allgemeinere (aber auch etwas kompliziertere) Definition des Erwartungswertes gegeben. Die wichtigsten Konsequenzen daraus werden in diesem Unterabschnitt kurz zusammengefasst und an einigen Beispielen illustriert.

Die allgemeine Definition des Erwartungswertes erfolgt durch Definition eines Integrals $\int_\Omega X(\omega) dP(\omega)$.

Ist $h : \mathbb{R} \to \mathbb{R}$ eine (messbare) reelle Funktion, und ist X eine diskrete Zufallsvariable mit Werten $x_1, x_2, \dots \in \mathbb{R}$, so gilt:

$$\mathbf{E}h(X) = \sum_{k=1}^{\infty} h(x_k) \cdot \mathbf{P}[X = x_k].$$

Ist dagegen X stetig verteilt mit Dichte f, so gilt

$$\mathbf{E}h(X) = \int_{-\infty}^{\infty} h(x) \cdot f(x)\, dx.$$

Mit $h(x) = x$ folgen daraus die bisher eingeführten Berechnungsvorschriften für Erwartungswerte.

Anwendung im Beispiel 4.23: Sei X die zufällige Ankunftszeit an der Ampel. Nach Voraussetzung ist diese auf dem Intervall $[0, 3]$ gleichverteilt. Da die Ampel im Intervall $[0, 1)$ grün und im Intervall $[1, 3]$ rot ist, gilt für die zufällige Wartezeit Z an der Ampel:

$$Z = h(X) \quad \text{mit} \quad h(x) = \begin{cases} 0 & \text{für} \quad 0 \le x < 1, \\ 3 - x & \text{für} \quad 1 \le x \le 3, \end{cases}$$

Damit folgt

$$\begin{aligned} \mathbf{E}Z &= \mathbf{E}h(X) = \int_{-\infty}^{\infty} h(x) \cdot f(x)\, dx = \int_0^3 h(x) \cdot \frac{1}{3}\, dx \\ &= \int_1^3 (3 - x) \cdot \frac{1}{3}\, dx = x - \frac{1}{6}x^2 \Big|_{x=1}^{3} = \frac{2}{3}. \end{aligned}$$

Beispiel 4.29 *Die zufällige Zeit, die eine Internet Suchmaschine bis zum Finden der Antwort auf die Anfrage eines Benutzers benötigt, werde durch eine* $\exp(\lambda)$*–verteilte reelle Zufallsvariable* X *angegeben. Um genügend Zeit für die Präsentation von Werbung zu haben, wird dem Benutzer die Antwort aber grundsätzlich nicht vor Ablauf einer festen Zeit* $t > 0$ *gegeben, d.h. für die zufällige Zeit* Y *bis zur Beantwortung der Anfrage des Benutzers gilt*

$$Y(\omega) = \max\{X(\omega), t\} \quad (\omega \in \Omega).$$

Wie lange muss der Benutzer dann im Mittel auf die Antwort auf seine Anfrage warten ?

Man erhält:

$$\begin{aligned}
\mathbf{E}Y &= \int_0^\infty \max\{x,t\}\cdot\lambda\cdot e^{-\lambda\cdot x}dx \\
&= \int_0^t \max\{x,t\}\cdot\lambda\cdot e^{-\lambda\cdot x}dx + \int_t^\infty \max\{x,t\}\cdot\lambda\cdot e^{-\lambda\cdot x}dx \\
&= \int_0^t t\cdot\lambda\cdot e^{-\lambda\cdot x}dx + \int_t^\infty x\cdot\lambda\cdot e^{-\lambda\cdot x}dx \\
&= -t\cdot e^{-\lambda\cdot x}\Big|_{x=0}^{t} + (-x)\cdot e^{-\lambda\cdot x}\Big|_{x=t}^{\infty} + \int_t^\infty e^{-\lambda\cdot x}dx \\
&= -t\cdot e^{-\lambda\cdot t} + t - 0 + t\cdot e^{-\lambda\cdot t} - \frac{1}{\lambda}\cdot e^{-\lambda\cdot x}\Big|_{x=t}^{\infty} \\
&= t + \frac{1}{\lambda}\cdot e^{-\lambda\cdot t}.
\end{aligned}$$

Beispiel 4.30 *Nach erfolgreichen Beenden des Cannstatter Volksfestes beschließt Student S., sich geschäftlich weiterzuentwickeln, und eröffnet einen Weihnachtsbaumgroßhandel. Dazu kauft er von einem Förster 10.000 Weihnachtsbäume, deren Größen rein zufällig zwischen 50cm und 300cm schwanken.*

Zur Festlegung der Preise betrachtet er die beiden folgenden Möglichkeiten:

Bei Möglichkeit 1 verlangt er für jeden Baum pro Zentimeter Länge 10 Cent.

Bei Möglichkeit 2 legt er den Preis (in Euro) eines Baumes in Abhängigkeit von der Länge x des Baumes in Zentimeter fest gemäß

$$h(x) = \begin{cases} 6 & \text{für} \quad x \leq 100, \\ 6 + (x-100)\cdot\frac{24}{150} & \text{für} \quad 100 < x < 250, \\ 30 & \text{für} \quad x \geq 250. \end{cases}$$

Bei welcher der beiden Möglichkeiten ist der mittlere Verkaufserlös höher ?

Erstes Preissystems in Beispiel 4.30:

Als Ergebnis des Zufallsexperiments (Festlegung des Preises für den zufällig ausgewählten Baum gemäß 10 Cent pro Zentimeter) erhält man einen Wert, der rein zufällig zwischen 5 Euro und 30 Euro schwankt. Dabei kann jeder Wert zwischen 5 und 30 auftreten, daher kann dieses Zufallsexperiment nicht durch eine diskrete Zufallsvariable beschrieben werden.

Statt dessen verwendet man hierbei eine *stetig verteilte Zufallsvariable mit Dichte*, d.h. man setzt

$$\mathbf{P}[X \in B] = \int_B f(x)\,dx$$

für $B \subseteq \mathbb{R}$, wobei $f : \mathbb{R} \to \mathbb{R}_+$ eine Funktion mit $\int_{-\infty}^{\infty} f(x)\,dx = 1$ ist (sogenannte *Dichte*).

Da hier der Wert von X rein zufällig zwischen 5 und 30 schwanken soll, setzt man

$$f(x) = \begin{cases} \frac{1}{30-5} = \frac{1}{25} & \text{für} \quad 5 \le x \le 30, \\ 0 & \text{für} \quad x < 5 \text{ oder } x > 30. \end{cases}$$

X ist damit eine auf dem Intervall $[5, 30]$ gleichverteilte Zufallsvariable.

Der mittlere Verkaufserlös ist hier gegeben durch

$$\mathbf{E}X = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_{5}^{30} x \cdot \frac{1}{25}\,dx = \frac{1}{50} x^2 \Big|_{x=5}^{30} = 17.5$$

Beschreibung des zweiten Preissystems in Beispiel 4.30:

Als Ergebnis des Zufallsexperiments (Festlegung des Preises für den zufällig ausgewählten Baum als Funktion h von der Länge x) erhält man einen Wert, der zwischen 6 Euro und 30 Euro liegt. Dabei kann jeder Wert zwischen 6 und 30 auftreten, daher kann dieses Zufallsexperiment nicht durch eine diskrete Zufallsvariable beschrieben werden. Darüberhinaus ist aber die Wahrscheinlichkeit, dass der Wert 6 auftritt, größer Null, woraus folgt, dass das Zufallsexperiment auch nicht durch eine stetig verteilte Zufallsvariable mit Dichte beschrieben werden kann.

Statt dessen beschreiben wir die zufällige Länge X des Baumes durch eine auf eine auf dem Intervall $[50, 300]$ gleichverteilte Zufallsvariable, d.h.

$$\mathbf{P}[X \in B] = \int_B f(x)\,dx$$

für $B \subseteq \mathbb{R}$, wobei $f : \mathbb{R} \to \mathbb{R}_+$ gegeben ist gemäß

$$f(x) = \begin{cases} \frac{1}{300-50} = \frac{1}{250} & \text{für} \quad 50 \le x \le 300, \\ 0 & \text{für} \quad x < 50 \text{ oder } x > 300, \end{cases}$$

und beschreiben dann den Preis eines zufällig ausgewählten Baumes durch $h(X)$, d.h. bei Länge $X(\omega)$ beträgt der Preis $h(X(\omega))$ mit

$$h(x) = \begin{cases} 6 & \text{für} \quad x \le 100, \\ 6 + (x - 100) \cdot \frac{24}{150} & \text{für} \quad 100 < x < 250, \\ 30 & \text{für} \quad x \ge 250. \end{cases}$$

Für den mittleren Verkaufserlös erhält man hier:

$$\begin{aligned}
&\mathbf{E}h(X)\\
&= \int_{-\infty}^{\infty} h(x) \cdot f(x)\, dx\\
&= \int_{50}^{300} h(x) \cdot \frac{1}{250}\, dx\\
&= \int_{50}^{100} 6 \cdot \frac{1}{250}\, dx + \int_{100}^{250} (6 + (x-100) \cdot \frac{24}{150}) \cdot \frac{1}{250}\, dx + \int_{250}^{300} 30 \cdot \frac{1}{250}\, dx\\
&= \frac{6}{250} \cdot x\Big|_{x=50}^{100} + \left(6 \cdot x + \frac{1}{2}(x-100)^2 \frac{24}{150}\right) \cdot \frac{1}{250}\Big|_{x=100}^{250} + \frac{30}{250} \cdot x\Big|_{x=250}^{300}\\
&= 18.
\end{aligned}$$

Also ist der mittlere Verkaufserlös beim zweiten Preissystem höher als beim ersten Preissystem.

Aus der allgemeinen Definition des Erwartungswertes als Integral folgt aufgrund der Linearität des Integrals auch

- $\mathbf{E}(X+Y) = \mathbf{E}X + \mathbf{E}Y$, für beliebige reelle Zufallsvariablen X und Y, sowie
- $\mathbf{E}(\alpha \cdot X) = \alpha \cdot \mathbf{E}X$, für beliebige $\alpha \in \mathbb{R}$ und beliebige reelle Zufallsvariablen X.

Beispiel 4.31 *Zehn perfekten Schützen stehen zehn unschuldige Enten gegenüber. Jeder Schütze wählt zufällig und unbeeinflusst von den anderen Schützen eine Ente aus, auf die er schießt. Wieviele Enten überleben im Mittel ?*

Sei X die zufällige Anzahl der überlebenden Enten. Dann ist X eine diskrete Zufallsvariable die nur Werte in $\{0, \dots, 9\}$ annimmt. Damit erhält man den Erwartungswert von X zu

$$\mathbf{E}X = \sum_{i=0}^{9} i \cdot \mathbf{P}[X = i].$$

Problematisch daran ist, dass die Wahrscheinlichkeiten $\mathbf{P}[X = i]$ schwierig bestimmbar sind. Als Ausweg bietet sich die folgende Darstellung von X an:

$$X = \sum_{i=1}^{10} X_i,$$

wobei

$$X_i = \begin{cases} 1 & \text{falls} \quad \text{Ente } i \text{ überlebt,}\\ 0 & \text{falls} \quad \text{Ente } i \text{ nicht überlebt.} \end{cases}$$

Mit

$$\mathbf{P}[X_i = 1] = \prod_{j=1}^{10} \mathbf{P}[\text{Schütze } j \text{ ziehlt nicht auf Ente } i] = \left(\frac{9}{10}\right)^{10}$$

folgt

$$\mathbf{E}X_i = 1 \cdot \mathbf{P}[X_i = 1] = \left(\frac{9}{10}\right)^{10},$$

und daraus wiederum

$$\mathbf{E}X = \mathbf{E}\left\{\sum_{i=1}^{10} X_i\right\} = \sum_{i=1}^{10} \mathbf{E}\left\{X_i\right\} = 10 \cdot \left(\frac{9}{10}\right)^{10} \approx 3.49.$$

4.6.4 Mathematisch exakte Definition des Erwartungswertes

Im Folgenden wird definiert:

$$\mathbf{E}X = \int_\Omega X(\omega) d\mathbf{P}(\omega) \approx \sum_{\omega\in\Omega} X(\omega) \cdot \mathbf{P}(\{\omega\}),$$

d.h. der Erwartungswert wird als (geeignet definiertes) Integral eingeführt, das anschaulich der mit den Wahrscheinlichkeiten $\mathbf{P}(\{\omega\})$ gewichteten Summe der Ergebnisse $X(\omega)$ des Zufallsexperiments entspricht.

Zur exakten Definition des obigen Integrals werden die folgenden Begriffe benötigt:

Definition 4.28 $(\Omega, \mathcal{A})$ *sei Messraum.*

a) *Eine Funktion* $f : \Omega \to \mathbb{R}$ *heißt* $\mathcal{A} - \mathcal{B}-$**messbar** *(kurz: messbar), falls gilt:*

$$f^{-1}(B) = \{\omega \in \Omega : f(\omega) \in B\} \in \mathcal{A} \quad \textit{für alle } B \in \mathcal{B}.$$

b) *Jede Funktion* $f : \Omega \to \mathbb{R}$ *mit*

$$f = \sum_{i=1}^{n} \alpha_i \cdot 1_{A_i},$$

wobei $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $A_1, \ldots, A_n \in \mathcal{A}$, $\{A_1, \ldots, A_n\}$ *Partition von* Ω, *heißt* **einfache Funktion**.

c) *Eine Folge von Funktionen* $f_n : \Omega \to \mathbb{R}$ **konvergiert von unten** *gegen* $f : \Omega \to \mathbb{R}$*, falls gilt:*

$$f_1(\omega) \leq f_2(\omega) \leq \ldots \quad \textit{und} \quad \lim_{n\to\infty} f_n(\omega) = f(\omega) \quad \textit{für alle } \omega \in \Omega.$$

Schreibweise dafür: $f_n \uparrow f$.

Definition 4.29 *Allgemeine Definition des Maßintegrals.*

$(\Omega, \mathcal{A}, \mu)$ *sei Maßraum,* $f : \Omega \to \mathbb{R}$ *sei messbar.*

a) *Ist* $f = \sum_{i=1}^{n} \alpha_i \cdot 1_{A_i}$ *eine nichtnegative einfache Funktion, so wird definiert:*

$$\int f \, d\mu = \sum_{i=1}^{n} \alpha_i \cdot \mu(A_i).$$

b) *Ist* f *nichtnegativ einfach, so wird definiert:*

$$\int f \, d\mu = \lim_{n\to\infty} \int f_n d\mu,$$

wobei $(f_n)_{n\in\mathbb{N}}$ *eine beliebige Folge nichtnegativer einfacher Funktionen ist mit* $f_n \uparrow f$.

Eine solche Folge existiert immer, z.B. kann man wählen:

$$f_n = n \cdot 1_{\{\omega\in\Omega \,:\, f(\omega)\geq n\}} + \sum_{k=0}^{n\cdot 2^n-1} \frac{k}{n} \cdot 1_{\left\{\omega\in\Omega \,:\, \frac{k}{2^n} \leq f(\omega) < \frac{k+1}{2^n}\right\}}.$$

c) *Nimmt* f *auch negative Werte an, so wird*

$$\begin{aligned} f^+(\omega) &= \max\{f(\omega), 0\}, \\ f^-(\omega) &= \max\{-f(\omega), 0\} \end{aligned}$$

gesetzt (so dass gilt: $f(\omega) = f^+(\omega) - f^-(\omega)$*, wobei* $f^+(\omega) \geq 0$*,* $f^-(\omega) \geq 0$*), und im Falle*

$$\int f^+ d\mu < \infty \quad \textit{oder} \quad \int f^- d\mu < \infty$$

wird definiert:

$$\int f \, d\mu = \int f^+ d\mu - \int f^- d\mu.$$

Schreibweisen:

$$\int f\,d\mu = \int_\Omega f\,d\mu = \int f(\omega)\,d\mu(\omega) = \int_\Omega f(\omega)\,d\mu(\omega) = \int f(\omega)\mu(d\omega) = \int_\Omega f(\omega)\mu(d\omega)$$

Bemerkung: Das obige Integral ist wohldefiniert, da gilt

(i) Ist

$$f = \sum_{i=1}^{n} \alpha_i 1_{A_i} = \sum_{j=1}^{m} \beta_j 1_{B_j}$$

mit $\alpha_i, \beta_j \in \mathbb{R}$, $A_i, B_j \in \mathcal{A}$, $\{A_i \ : \ i = 1, \ldots, n\}$ und $\{B_j \ : \ j = 1, \ldots, m\}$ Partitionen von Ω, so gilt

$$\sum_{i=1}^{n} \alpha_i \mu(A_i) = \sum_{j=1}^{m} \beta_j \mu(B_j).$$

Begründung: Da $\{B_j \ : \ j = 1, \ldots, m\}$ Partition von Ω ist, gilt

$$A_i = A_i \cap \Omega = A_i \cap \left(\cup_{j=1}^{m} B_j\right) = \cup_{j=1}^{m} A_i \cap B_j,$$

wobei die Mengen in der letzten Vereinigung paarweise leeren Schnitt haben. Aufgrund der σ-Additivität von $\mathbf{P}$ folgt daraus

$$l.S. \ = \ \sum_{i=1}^{n} \alpha_i \mu \left(\cup_{j=1}^{m} A_i \cap B_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \mu(A_i \cap B_j).$$

Analog erhält man

$$r.S. = \sum_{j=1}^{m} \beta_j \mu \left(\cup_{i=1}^{n} A_i \cap B_j\right) = \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_j \mu(A_i \cap B_j).$$

Ist nun $A_i \cap B_j \neq \emptyset$, so folgt durch Wahl von $\omega \in A_i \cap B_j$:

$$f(\omega) = \sum_{k=1}^{n} \alpha_k 1_{A_k}(\omega) = \alpha_i \quad \text{sowie} \quad f(\omega) = \sum_{k=1}^{m} \beta_k 1_{B_k}(\omega) = \beta_j,$$

also gilt in diesem Fall $\alpha_i = \beta_j$. Dies impliziert $l.S.{=}r.S.$, w.z.z.w.

(ii) Man kann darüberhinaus (mit Hilfe eines etwas technischen Beweises) zeigen, dass der Grenzwert in b) existiert und unabhängig von der Wahl der f_n mit $f_n \uparrow f$ ist.

Definition 4.30 *W–Raum* $(\Omega, \mathcal{A}, \mathbf{P})$, $X : \Omega \to \mathbb{R}$ *reelle ZV. Dann heißt*

$$\mathbf{E}X := \int_\Omega X(\omega)\, d\mathbf{P}(\omega)$$

– sofern existent – der **Erwartungswert** *der Zufallsvariablen* X.

Einige nützliche Eigenschaften des Integrals werden beschrieben in

Satz 4.7 $(\Omega, \mathcal{A}, \mu)$ *Maßraum,* $f, g : \Omega \to \mathbb{R}$ *messbar,* $\alpha \in \mathbb{R}$. *Dann gilt:*

a)

$$\int (f+g)\, d\mu = \int f\, d\mu + \int g\, d\mu.$$

b)

$$\int (\alpha \cdot f)\, d\mu = \alpha \cdot \int f\, d\mu.$$

c)

$$f(\omega) \le g(\omega) \text{ für alle } \omega \in \Omega \quad \Rightarrow \quad \int f\, d\mu \le \int g\, d\mu.$$

Folgerung: $\mathbf{E}(X_1 + X_2) = \mathbf{E}(X_1) + \mathbf{E}(X_2)$ und $\mathbf{E}(\alpha \cdot X) = \alpha \cdot \mathbf{E}(X)$, wobei $X_1 + X_2$ bzw. $\alpha \cdot X$ die Zufallsvariablen mit Werten $X_1(\omega) + X_2(\omega)$ bzw. $\alpha \cdot X(\omega)$ sind.

Beweis von Satz 4.7:

a) Gemäß der schrittweisen Definition des Integrals erfolgt der Beweis schrittweise für nichtnegative einfache Funktionen, nichtnegative Funktionen und beliebige messbare Funktionen.

Fall 1: f, g *nichtnegativ einfach*

Sei $f = \sum_{i=1}^n \alpha_i 1_{A_i}$ und $g = \sum_{j=1}^m \beta_j 1_{B_j}$, mit $A_i, B_j \in \mathcal{A}$ und $\{A_1, \dots, A_n\}$ bzw. $\{B_1, \dots, B_m\}$ seien Partitionen von Ω. Wegen

$$A_i = A_i \cap \Omega = A_i \cap (\cup_{j=1}^m B_j) = \cup_{j=1}^m A_i \cap B_j$$

und $A_i \cap B_1, \dots, A_i \cap B_m$ paarweise disjunkt gilt dann

$$1_{A_i} = \sum_{j=1}^m 1_{A_i \cap B_j},$$

woraus folgt

$$f = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i 1_{A_i \cap B_j}.$$

Analog erhält man

$$g = \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_j 1_{A_i \cap B_j}.$$

Damit

$$f + g = \sum_{i=1}^{n} \sum_{j=1}^{m} (\alpha_i + \beta_j) \cdot 1_{A_i \cap B_j}, \tag{4.3}$$

und aus der Definition des Integrals folgt

$$\begin{aligned}
\int (f+g)\, d\mu &= \sum_{i=1}^{n} \sum_{j=1}^{m} (\alpha_i + \beta_j) \cdot \mu(A_i \cap B_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \cdot \mu(A_i \cap B_j) + \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_j \cdot \mu(A_i \cap B_j) \\
&= \int f\, d\mu + \int g\, d\mu.
\end{aligned}$$

Fall 2: *f, g nichtnegativ*

Wähle nichtnegative einfache Funktionen f_n und g_n mit $f_n \uparrow f$ und $g_n \uparrow g$. Dann sind $f_n + g_n$ einfache Funktionen (vgl. (4.3)) mit $f_n + g_n \uparrow f + g$, und aus der Definition des Integrals bzw. des ersten Falles folgt

$$\begin{aligned}
\int (f+g)\, d\mu &= \lim_{n\to\infty} \int (f_n + g_n)\, d\mu \\
&= \lim_{n\to\infty} \left(\int f_n\, d\mu + \int g_n\, d\mu \right) \\
&= \lim_{n\to\infty} \int f_n\, d\mu + \lim_{n\to\infty} \int g_n\, d\mu \\
&= \int f\, d\mu + \int g\, d\mu.
\end{aligned}$$

Fall 3: *f, g beliebig*

Aus

$$f + g = (f+g)^+ - (f+g)^-$$

und
$$f + g = (f^+ - f^-) + (g^+ - g^-)$$
folgt
$$(f^+ + g^+) + f^- + g^- = f^+ + g^+ + (f + g)^-.$$
Anwendung des Integrals auf beiden Seiten dieser Gleichung und Verwendung des Resultats von Fall 2 ergibt
$$\int (f+g)^+ \, d\mu + \int f^- \, d\mu + \int g^- \, d\mu = \int f^+ \, d\mu + \int g^+ \, d\mu + \int (f+g)^- \, d\mu,$$
woraus folgt
$$\begin{aligned}\int (f+g)\, d\mu &= \int (f+g)^+ \, d\mu - \int (f+g)^- \, d\mu \\ &= \int f^+ \, d\mu - \int f^- \, d\mu + \int g^+ \, d\mu - \int g^- \, d\mu \\ &= \int f \, d\mu + \int g \, d\mu.\end{aligned}$$

b) Für $\alpha > 0$ folgt die Behauptung analog zu a), für $\alpha = 0$ ist sie trivial. Für $\alpha < 0$ gilt
$$(\alpha \cdot f)^+ = (-\alpha) \cdot f^- \quad \text{und} \quad (\alpha \cdot f)^- = (-\alpha) \cdot f^+.$$
Unter Benutzung des Resultates für den Fall $\alpha > 0$ und der Definition des Integrals folgt daraus
$$\begin{aligned}\int (\alpha \cdot f) d\mu &= \int (\alpha \cdot f)^+ d\mu - \int (\alpha \cdot f)^- d\mu \\ &= (-\alpha) \cdot \left(\int f^- d\mu - \int f^+ d\mu \right) \\ &= \alpha \cdot \int f d\mu.\end{aligned}$$

c) Aus $f(\omega) \le g(\omega)$ für alle $\omega \in \Omega$ folgt $g(\omega) - f(\omega) \ge 0$ für alle $\omega \in \Omega$. Nach Definition des Integrals ist das Integral im Falle nichtnegativer Funktionen nichtnegativ, was impliziert
$$\int (g - f) \, d\mu \ge 0.$$
Mit a) und b) folgt
$$\int g \, d\mu - \int f \, d\mu = \int (g - f) \, d\mu \ge 0.$$

$\square$

Die nächsten beiden Sätze bilden die Grundlage zur Berechnung von Erwartungswerten.

Satz 4.8 *(Transformationssatz für Integrale)*

$(\Omega, \mathcal{A}, \mathbf{P})$ *sei ein W–Raum,* X *sei eine reelle ZV und* $h : \mathbb{R} \to \mathbb{R}$ *sei messbar. Dann gilt*

$$\int_\Omega h(X(\omega))\, d\mathbf{P}(\omega) = \int_\mathbb{R} h(x) d\mathbf{P}_X(x),$$

wobei $\mathbf{P}_X$ *die Verteilung von* X *ist, d.h.,*

$$\mathbf{P}_X(B) = \mathbf{P}(X^{-1}(B)) \quad (B \in \mathcal{B}).$$

Beweis: Gemäß der schrittweisen Definition des Integrals erfolgt der Beweis wieder schrittweise für nichtnegative einfache Funktionen, nichtnegative Funktionen und beliebige messbare Funktionen. Im Folgenden wird die Behauptung nur im Falle h nichtnegativ einfach gezeigt. Der allgemeine Fall folgt daraus analog zum Beweis von Satz 4.7, Teil a).

Sei also $h = \sum_{i=1}^n \alpha_i \cdot 1_{A_i}$ nichtnegativ und einfach. Dann gilt

$$\begin{aligned} h(X(\omega)) &= \sum_{i=1}^n \alpha_i \cdot 1_{A_i}(X(\omega)) \\ &= \sum_{i=1}^n \alpha_i \cdot 1_{X^{-1}(A_i)}(\omega), \end{aligned}$$

und aus der Definition des Integrals und der Verteilung von X folgt:

$$\begin{aligned} \int_\Omega h(X(\omega))\, d\mathbf{P}(\omega) &= \sum_{i=1}^n \alpha_i \cdot \mathbf{P}\left(X^{-1}(A_i)\right) \\ &= \sum_{i=1}^n \alpha_i \cdot \mathbf{P}_X\left(A_i\right) \\ &= \int_\mathbb{R} h(x)\, d\mathbf{P}_X(x). \end{aligned}$$

$\square$

Satz 4.9 $(\Omega, \mathcal{A}, \mathbf{P})$ *sei W–Raum, X sei reelle Zufallsvariable und $g : \mathbb{R} \to \mathbb{R}$ sei messbar.*

a) *Ist X eine diskrete Zufallsvariable mit Werten $x_1, x_2, \dots$, so gilt*

$$\int_{\mathbb{R}} g(\omega)\, d\mathbf{P}_X(\omega) = \sum_{k=1}^{\infty} g(x_k) \cdot \mathbf{P}[X = k].$$

b) *Ist X eine stetig verteilte Zufallsvariable mit Dichte f, so gilt*

$$\int_{\mathbb{R}} g(\omega)\, d\mathbf{P}_X(\omega) = \int_{\mathbb{R}} g(x) \cdot f(x)\, dx.$$

Beweis: Gemäß der schrittweisen Definition des Integrals erfolgt der Beweis wieder schrittweise für nichtnegative einfache Funktionen, nichtnegative Funktionen und beliebige messbare Funktionen. Im Folgenden wird die Behauptung nur im Falle $g = \sum_{i=1}^{n} \alpha_i \cdot 1_{A_i}$ gezeigt, der allgemeine Fall folgt daraus analog zum Beweis von Satz 4.7, Teil a).

a) Aus der Definition des Integrals und der Wahl von X folgt:

$$\begin{aligned}
\int_{\mathbb{R}} g(\omega)\, d\mathbf{P}_X(\omega) &= \sum_{i=1}^{n} \alpha_i \cdot \mathbf{P}_X(A_i) \\
&= \sum_{i=1}^{n} \alpha_i \cdot \sum_{k: x_k \in A_i} \mathbf{P}[X = x_k] \\
&= \sum_{i=1}^{n} \sum_{k: x_k \in A_i} \alpha_i \cdot \mathbf{P}[X = x_k] \\
&= \sum_{i=1}^{n} \sum_{k: x_k \in A_i} g(x_k) \cdot \mathbf{P}[X = x_k] \\
&= \sum_{k=1}^{\infty} g(x_k) \cdot \mathbf{P}[X = x_k],
\end{aligned}$$

wobei für die letzte Gleichheit benutzt wurde, dass $\{A_1, \dots, A_n\}$ eine Partition von $\mathbb{R}$ ist.

b) Aus der Definition des Integrals bzw. der Wahl von X folgt:

$$\begin{aligned}
\int_{\mathbb{R}} g(\omega)\, d\mathbf{P}_X(\omega) &= \sum_{i=1}^{n} \alpha_i \cdot \mathbf{P}_X(A_i) \\
&= \sum_{i=1}^{n} \alpha_i \cdot \int_{A_i} f(x)\, dx \\
&= \sum_{i=1}^{n} \int_{A_i} \alpha_i \cdot f(x)\, dx \\
&= \sum_{i=1}^{n} \int_{A_i} g(x) \cdot f(x)\, dx \\
&= \int_{\mathbb{R}} g(x) \cdot f(x)\, dx
\end{aligned}$$

wobei für die letzte Gleichheit benutzt wurde, dass $\{A_1, \ldots, A_n\}$ eine Partition von $\mathbb{R}$ ist. □

Korollar 4.1 *Sei X eine reelle Zufallsvariable und $h : \mathbb{R} \to \mathbb{R}$ messbar.*

a) *Ist X diskrete Zufallsvariable mit Werten $x_1, x_2, \ldots$, so gilt:*

$$\mathbf{E}h(X) = \sum_{k=1}^{\infty} h(x_k) \cdot \mathbf{P}[X = x_k],$$

insbesondere (mit $h(x) = x$)

$$\mathbf{E}X = \sum_{k=1}^{\infty} x_k \cdot \mathbf{P}[X = x_k].$$

b) *Ist X stetig verteilte Zufallsvariable mit Dichte f, so gilt*

$$\mathbf{E}h(X) = \int_{\mathbb{R}} h(x) \cdot f(x)\, dx,$$

insbesondere (mit $h(x) = x$)

$$\mathbf{E}X = \int_{\mathbb{R}} x \cdot f(x)\, dx.$$

Beweis: Gemäß der Definition des Erwartungswertes und Satz 4.8 gilt

$$\mathbf{E}h(X) = \int_{\Omega} h(X(\omega))\, d\mathbf{P}(\omega) = \int_{\mathbb{R}} h(x)\, d\mathbf{P}_X(x).$$

Mit Satz 4.9 folgt daraus die Behauptung. □

4.7 Varianz

Der Ewartungswert beschreibt den Wert, den man "im Mittel" bei Durchführung eines Zufallsexperiments erhält. In vielen Anwendungen reicht diese Information aber keineswegs aus. Interessiert man sich z.B. für den Kauf einer Aktie, so möchte man nicht nur wissen, was man im Mittel daran verdient. Vielmehr möchte man im Hinblick auf die Beurteilung des Risikos, das man eingeht, unter anderem auch wissen, wie stark der zukünftige Erlös um diesen mittleren Wert schwankt. Ein Kriterium zur Beurteilung der zufälligen Schwankung des Resultats eines Zufallsexperiments ist die sogenannte Varianz, die die mittlere quadratische Abweichung zwischen einem zufälligen Wert und seinem Mittelwert beschreibt:

Definition 4.31 *Sei X eine reelle ZV für die $\mathbf{E}X$ existiert. Dann heißt*

$$V(X) = \mathbf{E}(|X - \mathbf{E}X|^2)$$

die **Varianz** *von X.*

Wir illustrieren diesen neu eingeführten Begriff zunächst anhand zweier Beispiele.

Beispiel 4.32 *Wir betrachten nochmals das Glücksrad aus Beispiel 4.18. Wie wir in Beispiel 4.32 gezeigt haben, beträgt der Wert des Gewinnes dabei im Mittel 50 Cent. Im Folgenden möchten wir wissen, wie stark der zufällige Gewinn um diesen Wert schwankt. Dazu bestimmen wir die mittlere quadratische Abweichnung zwischen zufälligem Gewinn und dem mittleren Gewinn von 50 Cent.*

Der mittlere Wert ist hier $\mathbf{E}X = 50$, die tatsächlich auftretenden Werte sind $0, 20, 300$ und 2500, damit sind die quadratischen Abweichungen gleich

$$(0-50)^2 = 50^2, (20-50)^2 = 30^2, (300-50)^2 = 250^2, (2500-50)^2 = 2450^2.$$

Diese treten mit den Wahrscheinlichkeiten

$$\mathbf{P}[X=0] = \frac{56}{64}, \mathbf{P}[X=20] = \frac{5}{64}, \mathbf{P}[X=300] = \frac{2}{64} \text{ und } \mathbf{P}[X=2500] = \frac{1}{64}.$$

Als mittlere quadratische Abweichung erhält man damit

$$V(X) = 50^2 \cdot \frac{56}{64} + 30^2 \cdot \frac{5}{64} + 250^2 \cdot \frac{2}{64} + 2450^2 \cdot \frac{1}{64} \approx 98.000$$

(Rechnung in Euro ergibt $\approx 9,8$ Euro^2).

Beispiel 4.33 *Sei X $N(a,\sigma^2)$ verteilt. Dann gilt $\mathbf{E}X = a$ (vgl. Beispiel 4.28) und*

$$V(X) = \mathbf{E}(|X-a|^2) = \int_{-\infty}^{\infty} (x-a)^2 \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}}\, dx.$$

Mit der Substitution $u = (x-a)/\sigma$ und partieller Integration folgt

$$\begin{aligned} V(X) &= \sigma^2 \int_{-\infty}^{\infty} u^2 \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}}\, du \\ &= \sigma^2 \int_{-\infty}^{\infty} u \cdot \left(u \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}} \right) du \\ &= \sigma^2 \left(u \cdot \frac{-1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}} \Big|_{u=-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}}\, du \right) \\ &= \sigma^2(0+1) = \sigma^2. \end{aligned}$$

Als nächstes leiten wir einige nützliche Rechenregeln für die Berechnung von Varianzen her:

Satz 4.10 *Sei X eine reelle ZV für die $\mathbf{E}X$ existiert. Dann gilt:*

a)

$$V(X) = \mathbf{E}(X^2) - (\mathbf{E}X)^2.$$

b) *Für alle $\alpha \in \mathbb{R}$:*

$$V(\alpha \cdot X) = \alpha^2 \cdot V(X).$$

c) *Für alle $\beta \in \mathbb{R}$:*

$$V(X+\beta) = V(X).$$

Beweis:

a) Aufgrund der Linearität des Erwartungswertes gilt:

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

b)

$$V(\alpha \cdot X) = \mathbf{E}\left(|\alpha\cdot X - \mathbf{E}(\alpha\cdot X)|^2\right) = \mathbf{E}\left(\alpha^2 \cdot |X - \mathbf{E}(X)|^2\right) = \alpha^2 \cdot V(X).$$

c)

$$\begin{aligned} V(X+\beta) &= \mathbf{E}\left(|(X+\beta) - \mathbf{E}(X+\beta)|^2\right) \\ &= \mathbf{E}\left(|X+\beta - (\mathbf{E}(X)+\beta)|^2\right) \\ &= \mathbf{E}\left(|X - \mathbf{E}(X)|^2\right) = V(X). \end{aligned}$$

□

Beispiel 4.34 *Sei X $\pi(\lambda)$–verteilt, d.h.*

$$\mathbf{P}[X = k] = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad (k \in \mathbb{N}_0).$$

Dann gilt $\mathbf{E}X = \lambda$ (siehe Beispiel 4.26) und

$$\begin{aligned} \mathbf{E}(X^2) &= \sum_{k=0}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} \\ &= \sum_{k=1}^{\infty} k \cdot (k-1) \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} + \sum_{k=1}^{\infty} k \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} \\ &= \lambda^2 \cdot \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \cdot e^{-\lambda} + \lambda \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \cdot e^{-\lambda} \\ &= \lambda^2 + \lambda \end{aligned}$$

und damit

$$V(X) = \mathbf{E}(X^2) - (\mathbf{E}X)^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda.$$

Der folgende Satz zeigt, dass die Varianz wirklich zur Abschätzung der Abweichung zwischen $X(\omega)$ und $\mathbf{E}X$ verwendet werden kann:

Satz 4.11 *Sei X eine reelle ZV für die $\mathbf{E}X$ existiert und sei $\epsilon > 0$ beliebig. Dann gilt:*

a)

$$\mathbf{P}[|X| > \epsilon] \leq \frac{\mathbf{E}(|X|^r)}{\epsilon^r} \quad \textit{für alle } r \geq 0.$$

(Markovsche Ungleichung)

b)

$$\mathbf{P}[|X - \mathbf{E}X| > \epsilon] \leq \frac{V(X)}{\epsilon^2}.$$

(Tschebyscheffsche Ungleichung)

Beweis:

a) Wir definieren zusätzliche Zufallsvariablen Y und Z wie folgt: $Y(\omega)$ sei 1 falls $|X(\omega)| > \epsilon$ und andernfalls 0,

$$Z(\omega) = \frac{|X(\omega)|^r}{\epsilon^r}.$$

Ist dann $Y(\omega) = 1$, so folgt $Z(\omega) \geq 1 = Y(\omega)$, und ist $Y(\omega) = 0$ so ist $Z(\omega) \geq 0 = Y(\omega)$. Also gilt $Y(\omega) \leq Z(\omega)$ für alle ω, was impliziert $\mathbf{E}Y \leq \mathbf{E}Z$. Mit der Definition des Erwartungswertes folgt:

$$\mathbf{P}[|X| > \epsilon] = \mathbf{E}Y \leq \mathbf{E}Z = \frac{\mathbf{E}(|X|^r)}{\epsilon^r}.$$

b) Setze $Y = (X - \mathbf{E}X)$. Dann folgt aus a) mit $r = 2$:

$$\mathbf{P}[|X - \mathbf{E}X| > \epsilon] = \mathbf{P}[|Y| > \epsilon] \leq \frac{\mathbf{E}(Y^2)}{\epsilon^2} = \frac{V(X)}{\epsilon^2}.$$

□

Als nächstes überlegen wir uns, wie die Varianz einer Summe von Zufallsvariablen mit den Varianzen der einzelnen Zufallsvariablen zusammenhängt. Im Falle von Unabhängigkeit zeigen wir, dass die Varianz der Summe gleich der Summe der Varianzen ist. Ein wichtiges Hilfsmittel dazu ist:

Satz 4.12 *Sind X_1, X_2 unabhängige reelle ZVen für die $\mathbf{E}(X_1)$, $\mathbf{E}(X_2)$ und $\mathbf{E}(X_1 \cdot X_2)$ existieren, so gilt:*

$$\mathbf{E}(X_1 \cdot X_2) = \mathbf{E}(X_1) \cdot \mathbf{E}(X_2)$$

(ohne Beweis)

Damit können wir zeigen:

Satz 4.13 *Sind X_1, X_2 unabhängige reelle ZVen für die $\mathbf{E}(X_1)$, $\mathbf{E}(X_2)$ und $\mathbf{E}(X_1 \cdot X_2)$ existieren, so gilt:*

$$V(X_1 + X_2) = V(X_1) + V(X_2)$$

Beweis:

$$\begin{aligned}
&V(X_1 + X_2) \\
&= \mathbf{E}\left(|(X_1 - \mathbf{E}X_1) + (X_2 - \mathbf{E}X_2)|^2\right) \\
&= \mathbf{E}\left(|X_1 - \mathbf{E}X_1|^2 + |X_2 - \mathbf{E}X_2|^2 + 2 \cdot (X_1 - \mathbf{E}X_1) \cdot (X_2 - \mathbf{E}X_2)\right) \\
&= \mathbf{E}\left(|X_1 - \mathbf{E}X_1|^2\right) + \mathbf{E}\left(|X_2 - \mathbf{E}X_2)\,|^2 + 2 \cdot \mathbf{E}\left((X_1 - \mathbf{E}X_1) \cdot (X_2 - \mathbf{E}X_2)\right).
\end{aligned}$$

Die Behauptung folgt aus

$$\begin{aligned}
&\mathbf{E}\left((X_1 - \mathbf{E}X_1) \cdot (X_2 - \mathbf{E}X_2)\right) \\
&= \mathbf{E}\left(X_1X_2 - X_1\mathbf{E}(X_2) - X_2\mathbf{E}(X_1) + \mathbf{E}(X_1) \cdot \mathbf{E}(X_2)\right) \\
&= \mathbf{E}(X_1 \cdot X_2) - \mathbf{E}(X_1) \cdot \mathbf{E}(X_2) - \mathbf{E}(X_2)\mathbf{E}(X_1) + \mathbf{E}(X_1) \cdot \mathbf{E}(X_2) \\
&= \mathbf{E}(X_1 \cdot X_2) - \mathbf{E}(X_1) \cdot \mathbf{E}(X_2) \\
&= 0,
\end{aligned}$$

wobei bei der letzten Gleichheit Satz 4.12 verwendet wurde. □

Bemerkung: Der letzte Satz gilt analog auch für beliebige endliche Summen unabhängiger Zufallsvariablen. Sind nämlich $X_1, \ldots, X_n$ unabhängige reelle Zufallsvariablen, für die $\mathbf{E}X_i$ und $\mathbf{E}(X_i \cdot X_j)$ existieren, so gilt:

$$\begin{aligned}
V\left(\sum_{i=1}^n X_i\right) &= \mathbf{E}\left(\left|\sum_{i=1}^n (X_i - \mathbf{E}X_i)\right|^2\right) \\
&= \mathbf{E}\left(\sum_{i=1}^n (X_i - \mathbf{E}X_i)^2 + \sum_{\substack{1 \le i,j \le n \\ i \ne j}} (X_i - \mathbf{E}X_i) \cdot (X_j - \mathbf{E}X_j)\right) \\
&= \sum_{i=1}^n \mathbf{E}\left((X_i - \mathbf{E}X_i)^2\right) + \sum_{\substack{1 \le i,j \le n \\ i \ne j}} \mathbf{E}\left((X_i - \mathbf{E}X_i) \cdot (X_j - \mathbf{E}X_j)\right) \\
&= \sum_{i=1}^n V(X_i) + \sum_{\substack{1 \le i,j \le n \\ i \ne j}} 0 \\
&= \sum_{i=1}^n V(X_i).
\end{aligned}$$

4.8 Gesetze der großen Zahlen

Beispiel 4.35 *Wir betrachten das wiederholte Drehen am Glücksrad aus Beispiel 4.18.*

Ist der im Durchschnitt ausgezahlte Gewinn wirklich durch $\mathbf{E}X$ *gegeben ?*

Es sei X_i der beim i-ten mal ausgezahlte Gewinn. Dann wird bei n-maligem Drehen im Durchschnitt der Gewinn

$$\frac{1}{n} \cdot (X_1 + \cdots + X_n)$$

ausgezahlt. Dieser beträgt im Mittel

$$\mathbf{E}\left\{\frac{1}{n} \cdot (X_1 + \cdots + X_n)\right\} = \frac{1}{n} \cdot (\mathbf{E}X_1 + \cdots + \mathbf{E}X_n) = \mathbf{E}X_1 = 50,$$

die mittlere quadratische Abweichung ist (da die einzelnen Auszahlungen unbeeinflusst voneinander erfolgen) gegeben durch

$$\begin{aligned} V\left(\frac{1}{n} \cdot (X_1 + \cdots + X_n)\right) &= \frac{1}{n^2} V\left(X_1 + \cdots + X_n\right) \\ &= \frac{1}{n^2}\left(V(X_1) + \cdots + V(X_n)\right) \\ &= \frac{1}{n} V(X_1). \end{aligned}$$

Mit der Ungleichung von Tschebyscheff folgt daraus:

$$\begin{aligned} \mathbf{P}\left\{\left|\frac{1}{n} \cdot (X_1 + \cdots + X_n) - 50\right| > \epsilon\right\} &\leq \frac{V\left(\frac{1}{n} \cdot (X_1 + \cdots + X_n)\right)}{\epsilon^2} \\ &= \frac{\frac{1}{n} V(X_1)}{\epsilon^2} \to 0 \quad (n \to \infty). \end{aligned}$$

Es folgt, dass die Wahrscheinlichkeit, dass im Durchschnitt etwas mehr oder etwas weniger als 50 Cent ausgezahlt wird, für n (=Anzahl Drehen) groß beliebig klein wird.

Im Folgenden interessieren wir uns für asymptotische Aussagen über Summen $\sum_{i=1}^n X_i$ unabhängiger Zufallsvariablen für große Werte von n. Zur Abkürzung der Schreibweise ist dabei die folgende Definition nützlich:

Definition 4.32 *Zufallsvariablen $X_1, \ldots, X_n$ heißen* **identisch verteilt**, *falls gilt:*

$$\mathbf{P}_{X_1} = \cdots = \mathbf{P}_{X_n}.$$

Eine Folge $(X_i)_{i\in\mathbb{N}}$ von Zufallsvariablen heißt identisch verteilt, falls gilt: $\mathbf{P}_{X_1} = \mathbf{P}_{X_2} = \ldots$

Der nächste Satz verallgemeinert Beispiel 4.35.

Satz 4.14 (Schwaches Gesetz der großen Zahlen).
$X_1, X_2, \ldots$ seien unabhängige identisch verteilte reelle Zufallsvariablen mit existierendem Erwartungswert $\mu = \mathbf{E}X_1$. Dann gilt für jedes $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right] = 0,$$

d.h.

$$\lim_{n\to\infty} \mathbf{P}\left(\left\{\omega \in \Omega \ : \ \left|\frac{1}{n}\sum_{i=1}^{n} X_i(\omega) - \mu\right| > \epsilon\right\}\right) = 0.$$

Beweis im Spezialfall $V(X_1) < \infty$:

Mit der Ungleichung von Tschebyscheff folgt:

$$\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right] \leq \frac{\mathbf{E}\left(X^2\right)}{\epsilon^2},$$

wobei

$$X = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu.$$

Wegen $\mathbf{E}X = 0$, der Unabhängigkeit von $X_1, \ldots, X_n$ und der identischen Verteiltheit von $X_1, \ldots, X_n$ gilt:

$$\mathbf{E}\left(X^2\right) = V\left(X\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(X_i\right) = \frac{V(X_1)}{n}.$$

Damit erhält man

$$\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right] \leq \frac{V(X_1)}{n \cdot \epsilon^2} \to 0 \quad (n \to \infty).$$

□

Darüberhinaus gilt:

Satz 4.15 (Starkes Gesetz der großen Zahlen von Kolmogoroff).
$X_1, X_2, \ldots$ seien unabhängige identisch verteilte reelle Zufallsvariablen mit existierendem Erwartungswert $\mu = \mathbf{E}X_1$. Dann gilt

$$\mathbf{P}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right] = 1,$$

d.h.

$$\mathbf{P}\left(\left\{\omega \in \Omega \;:\; \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i(\omega) = \mu\right\}\right) = 1.$$

(**ohne Beweis**)

Bemerkung: Die Behauptung des obigen Satzes läßt sich umformulieren zu

$$\mathbf{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \nrightarrow \mu \; (n \to \infty)\right] := \mathbf{P}\left(\left\{\omega \in \Omega \;:\; \frac{1}{n}\sum_{i=1}^{n} X_i(\omega) \nrightarrow \mu \; (n \to \infty)\right\}\right) = 0.$$

Man sagt, dass eine Folge von Zufallsvariablen Y_n *fast sicher gegen eine ZV* Y *konvergiert* (Schreibweise: $Y_n \to Y$ f.s.), falls gilt:

$$\mathbf{P}\left[Y_n \nrightarrow Y \; (n \to \infty)\right] := \mathbf{P}\left(\{\omega \in \Omega \;:\; Y_n(\omega) \nrightarrow Y(\omega) \; (n \to \infty)\}\right) = 0.$$

In diesem Sinne konvergiert also im obigen Satz $\frac{1}{n}\sum_{i=1}^{n} X_i$ fast sicher gegen μ.

Mit der Konvergenz fast sicher kann man rechnen wie mit der Konvergenz von Zahlenfolgen. Z.B. folgt aus $X_n \to X$ f.s. und $Y_n \to Y$ f.s., dass für beliebige $\alpha, \beta \in \mathbb{R}$ gilt

$$\alpha \cdot X_n + \beta \cdot Y_n \to \alpha \cdot X + \beta \cdot Y \quad f.s.$$

Zum Beweis beachte man

$$\begin{aligned}
&\mathbf{P}\left[\alpha \cdot X_n + \beta \cdot Y_n \nrightarrow \alpha \cdot X + \beta \cdot Y \; (n \to \infty)\right] \\
&\leq \mathbf{P}\left[X_n \nrightarrow X \; (n \to \infty)\right] + \mathbf{P}\left[Y_n \nrightarrow Y \; (n \to \infty)\right] \\
&= 0 + 0 = 0.
\end{aligned}$$

4.9 Der zentrale Grenzwertsatz

Beispiel 4.36 *Bei einer Abstimmung über zwei Vorschläge A und B stimmt eine resolute Gruppe von $r = 3.000$ Personen für A, während sich weitere $n = 1.000.000$ Personen unabhängig voneinander rein zufällig entscheiden. Wie groß ist die Wahrscheinlichkeit p, dass A angenommen wird ?*

Zur Modellierung des Abstimmungsverhalten im obigen Beispiel betrachten wir unabhängige Zufallsvariablen $X_1, \ldots, X_n$ mit

$$\mathbf{P}[X_i = 0] = \mathbf{P}[X_i = 1] = \frac{1}{2} \quad (i = 1, \ldots, n).$$

Hierbei bedeutet $X_i = 1$, dass die i–te Person für A stimmt, während $X_i = 0$ bedeutet, dass die i–te Person für B stimmt. Dann ist die Anzahl der Stimmen für A gleich

$$\sum_{i=1}^{n} X_i + r,$$

während die Anzahl der Stimmen für B gegeben ist durch

$$\sum_{i=1}^{n} (1 - X_i) = n - \sum_{i=1}^{n} X_i,$$

und gefragt ist nach der Wahrscheinlichkeit

$$p = \mathbf{P}\left[\sum_{i=1}^{n} X_i + r > n - \sum_{i=1}^{n} X_i\right].$$

Diese lässt sich wie folgt umformen:

$$\begin{aligned} p &= \mathbf{P}\left[\sum_{i=1}^{n} X_i + r > n - \sum_{i=1}^{n} X_i\right] \\ &= \mathbf{P}\left[2\sum_{i=1}^{n} X_i > n - r\right] \\ &= \mathbf{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{2} > -\frac{r}{2n}\right]. \end{aligned}$$

Die obige Wahrscheinlichkeit kann approximativ bestimmt werden mit

Satz 4.16 (Zentraler Grenzwertsatz von Lindeberg–Lévy).

Seien $X_1, X_2, \ldots$ unabhängige indentisch verteilte reelle Zufallsvariablen mit $\mathbf{E}(X_1^2) < \infty$. Setze

$$\mu = \mathbf{E}X_1 \quad \textit{und} \quad \sigma^2 = V(X_1).$$

Dann gilt, dass die Verteilungsfunktion von

$$\frac{1}{\sqrt{V(\sum_{i=1}^n X_i)}} \sum_{i=1}^n (X_i - \mathbf{E}X_i) = \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^n (X_i - \mu)$$

punktweise gegen die Verteilungsfunktion Φ einer $N(0,1)$–verteilten ZV konvergiert, d.h., dass für alle $x \in \mathbb{R}$ gilt:

$$\lim_{n\to\infty} \mathbf{P}\left[\frac{1}{\sqrt{n}\sigma} \sum_{i=1}^n (X_i - \mu) \le x\right] = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2}\, dt.$$

(**ohne Beweis**)

Sprechweise: $\frac{1}{\sqrt{n}\sigma} \sum_{i=1}^n (X_i - \mu)$ konvergiert nach Verteilung gegen eine $N(0,1)$–verteilte ZV.

Bemerkungen:

a) Die Aussage des obigen Satzes lässt sich wie folgt leicht merken: Betrachtet wird eine Summe unabhängiger identisch verteilter Zufallsvariablen. Gemäß obigem Satz lässt sich diese asymptotisch durch eine Normalverteilung approximieren. Dazu renormalisiert man diese Summe so, dass sie Erwartungswert Null und Varianz Eins hat, d.h., man ersetzt $\sum_{i=1}^n X_i$ durch

$$\frac{1}{\sqrt{V(\sum_{i=1}^n X_i)}} \sum_{i=1}^n (X_i - \mathbf{E}X_i).$$

Anschliessend kann man die Werte der Verteilungsfunktion der obigen normalisierten Summe durch die einer $N(0,1)$-Verteilung approximativ berechnen.

b) Aus obigem Satz folgt für $-\infty \leq \alpha < \beta \leq \infty$:

$$\mathbf{P}\left[\alpha < \frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(X_i-\mu) \leq \beta\right]$$

$$= \mathbf{P}\left[\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(X_i-\mu) \leq \beta\right] - \mathbf{P}\left[\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(X_i-\mu) \leq \alpha\right]$$

$$\stackrel{(n\to\infty)}{\longrightarrow} \Phi(\beta) - \Phi(\alpha) = \frac{1}{\sqrt{2\pi}}\int_{\alpha}^{\beta} e^{-t^2/2}\,dt.$$

Anwendung im Beispiel 4.36:

Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit

$$\mathbf{P}[X_i = 0] = \mathbf{P}[X_i = 1] = \frac{1}{2} \quad (i = 1, \ldots, n).$$

Dann gilt

$$\mathbf{E}X_1 = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2},$$

$$\mathbf{E}(X_1^2) = 0^2 \cdot \frac{1}{2} + 1^2 \cdot \frac{1}{2} = \frac{1}{2},$$

und

$$V(X_1) = \mathbf{E}(X_1^2) - (\mathbf{E}X_1)^2 = \frac{1}{4}.$$

Damit

$$
\begin{aligned}
p &= \mathbf{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{2} > -\frac{r}{2n}\right] \\
&= \mathbf{P}\left[\frac{1}{\sqrt{n}\sqrt{V(X_1)}}\sum_{i=1}^{n}\left(X_i - \frac{1}{2}\right) > -\frac{r}{2\sqrt{n}\sqrt{V(X_1)}}\right] \\
&= 1 - \mathbf{P}\left[\frac{1}{\sqrt{n}\sqrt{V(X_1)}}\sum_{i=1}^{n}\left(X_i - \frac{1}{2}\right) \leq -\frac{r}{2\sqrt{n}\sqrt{V(X_1)}}\right] \\
&\approx 1 - \Phi\left(-\frac{r}{2\sqrt{n}\sqrt{V(X_1)}}\right) \\
&= 1 - \Phi\left(-\frac{3000}{2\cdot 1000\cdot \frac{1}{2}}\right) \\
&= 1 - \Phi(-3) = \Phi(3) \approx 0,9986.
\end{aligned}
$$

Beispiel 4.37 *Ein Flugunternehmen weiß aus Erfahrung, dass im Mittel* 7% *derjenigen Personen, die ein Flugticket erworben haben, nicht bzw. zu spät zum Abflug erscheinen. Um die Zahl der somit ungenutzten Plätze nicht zu groß werden zu lassen, werden daher für einen Flug, bei dem* 240 *Plätze zu Verfügung stehen, mehr als* 240 *Flugtickets verkauft.*

Wieviele Flugscheine dürfen höchstens verkauft werden, dass mit Wahrscheinlichkeit mindestens 0.99 *alle zum Abflug erschienenen Personen, die ein Flugticket haben, auch einen Platz im Flugzeug bekommen ?*

Zur stochastischen Modellierung des obigen Beispiels betrachten wir unabhängige $b(1,p)$-verteilte Zufallsvariablen $X_1,\ldots,X_n$. Dabei gelte $X_i = 1$ genau dann, falls die Person, die das i-te Flugticket gekauft hat, (rechtzeitig) zum Abflug erscheint. $p = 1 - 0.07 = 0.93$ ist die Wahrscheinlichkeit, dass der Käufer des i-ten Flugtickets (rechtzeitig) zum Abflug erscheint, und n ist die Anzahl der verkauften Flugtickets.

Dann gibt $\sum_{i=1}^{n} X_i$ die Anzahl der zum Abflug erschienenen Personen, die ein Flugticket haben, an, und damit ist die Wahrscheinlichkeit, dass alle zum Abflug erschienenen Personen, die ein Flugticket haben, auch einen Platz im Flugzeug bekommen, gegeben gemäß:

$$\mathbf{P}\left[\sum_{i=1}^{n} X_i \leq 240\right].$$

Gesucht ist dass größte $n \in \mathbb{N}$ mit

$$\mathbf{P}\left[\sum_{i=1}^{n} X_i \leq 240\right] \geq 0.99.$$

Es gilt:

$$\begin{aligned}
&\mathbf{P}\left[\sum_{i=1}^{n} X_i \leq 240\right] \\
&= \mathbf{P}\left[\sum_{i=1}^{n} (X_i - \mathbf{E}X_1) \leq 240 - n \cdot \mathbf{E}X_1\right] \\
&= \mathbf{P}\left[\frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{V(X_1)}} \sum_{i=1}^{n} (X_i - \mathbf{E}X_1) \leq \frac{240 - n \cdot \mathbf{E}X_1}{\sqrt{n} \cdot \sqrt{V(X_1)}}\right].
\end{aligned}$$

Nach dem Zentralen Grenzwertsatz stimmt die letzte Wahrscheinlichkeit approximativ mit

$$\Phi\left(\frac{240 - n \cdot \mathbf{E}X_1}{\sqrt{n} \cdot \sqrt{V(X_1)}}\right)$$

überein, wobei Φ die Verteilungsfunktion der $N(0,1)$-Verteilung ist.

Mit

$$\mathbf{E}X_1 = p,\ V(X_1) = p(1-p) \text{ und } p = 0.93$$

folgt, dass die obige Bedingung approximativ äquivalent ist zu

$$\Phi\left(\frac{240 - n \cdot p}{\sqrt{n} \cdot \sqrt{p \cdot (1-p)}}\right) \geq 0.99.$$

Wegen $\Phi(2.4) \approx 0.99$ ist die äquivalent zu

$$\frac{240 - n \cdot p}{\sqrt{n} \cdot \sqrt{p \cdot (1-p)}} \geq 2.4$$

Quadrieren der letzten Gleichung liefert die notwendige Bedingung

$$\frac{(240 - n \cdot p)^2}{n \cdot p \cdot (1-p)} \geq 2.4^2$$

Diese impliziert aber nur dann die vorige Bedingung, wenn gleichzeitig

$$240 - n \cdot p \geq 0, \text{ d.h. } n \leq \frac{240}{p} = \frac{240}{0.93} \approx 258.1 \tag{4.4}$$

gilt.

Gilt dies, so führt die obige Bedingung auf

$$(240 - n \cdot p)^2 \geq 2.4^2 n \cdot p \cdot (1 - p)$$

bzw. auf

$$240^2 - (480p + 2.4^2 p \cdot (1 - p)) \cdot n + p^2 n^2 \geq 0.$$

Bestimmt man die Nullstellen des quadratischen Polynoms auf der linken Seite, so erhält man

$$n_1 \approx 247.7 \text{ und } n_2 \approx 268.8$$

Also ist die obige Ungleichung erfüllt für $n \leq 247$ oder $n \geq 269$.

Unter Berücksichtigung von $n \leq 258.1$ (vgl. (4.4)) erhält man als Resultat:

Es dürfen höchstens 247 Flugtickets verkauft werden, damit mit Wahrscheinlichkeit ≥ 0.99 nicht zu viele Passagiere beim Abflug erscheinen.

Kapitel 5

Induktive Statistik

5.1 Einführung

Aufgabenstellung der *induktiven* (oder auch *schließenden*) *Statistik* ist es, aufgrund von Beobachtungen eines zufälligen Vorgangs Rückschlüsse auf die zugrundeliegenden Gesetzmäßigkeiten, d.h. auf Eigenschaften des zugrundeliegenden W–Raumes, zu ziehen. Die verschiedenen Arten der dabei auftretenden Fragestellungen werden anhand des folgenden Beispiels erläutert.

Beispiel 5.1 *Ein Produzent stellt Sicherungen her. Beim Produktionsprozess lässt es sich nicht vermeiden, dass einige der produzierten Sicherungen defekt sind. Wie kann man feststellen, wie groß der Ausschussanteil* $p \in [0, 1]$ *ist ?*

Im Prinzip ist das keine stochastische Fragestellung. Man kann z.B. soviele Sicherungen herstellen, wie man insgesamt herstellen möchte, dann alle testen, ob sie defekt sind oder nicht, und kann daraus den relativen Anteil der defekten Sicherungen genau bestimmen. Dies ist aber aus zweierlei Gründen nicht sinnvoll: Zum einen ist das Testen aller Sicherungen sehr aufwendig, zum anderen könnte das Ergebnis sein, dass sehr viele defekt sind, so dass man eine große Zahl defekter Sicherungen hergestellt hätte. Wünschenswert wäre, das schon früher festzustellen, um dann noch Einfluss auf den Produktionsprozess nehmen zu können.

Eine naheliegende Idee ist, nur eine kleine Menge von Sicherungen zu testen, und daraus Rückschlüsse zu ziehen auf den Ausschussanteil in einer großen Menge von Sicherungen, die später nach der gleichen Methode hergestellt werden. Dazu entnimmt man der laufenden Produktion n Sicherungen und setzt

$$x_i = \begin{cases} 1, & \text{falls die } i\text{–te Sicherung defekt ist,} \\ 0, & \text{sonst,} \end{cases}$$

für $i = 1, \ldots, n$. Man versucht dann, ausgehend von $(x_1, \ldots, x_n) \in \{0,1\}^n$ Rückschlüsse auf p zu ziehen.

Der gesamte Vorgang kann stochastisch wie folgt modelliert werden: Man fasst die $x_1, \ldots, x_n$ als Realisierungen (d.h. als beobachtete Werte) von Zufallsvariablen $X_1, \ldots, X_n$ mit

$$X_i = \begin{cases} 1, & \text{falls die } i\text{–te Sicherung defekt ist,} \\ 0, & \text{sonst,} \end{cases}$$

auf, wobei diese ZVen unabhängig identisch verteilt sind mit

$$\mathbf{P}[X_1 = 1] = p, \quad \mathbf{P}[X_1 = 0] = 1 - p. \tag{5.1}$$

$x_1, \ldots, x_n$ wird dann als *Stichprobe* der Verteilung von X_1 bezeichnet.

In diesem Modell beschäftigt man sich mit den folgenden drei Fragestellungen:

Fragestellung 1: *Wie kann man ausgehend von* $(x_1, \ldots, x_n) \in \{0,1\}^n$ *den Wert von* p *schätzen ?*

Gesucht ist hier eine Funktion $T_n : \{0,1\}^n \to [0,1]$, für die $T_n(x_1, \ldots, x_n)$ eine "möglichst gute" Schätzung von p ist. Hierbei wird $T_n : \{0,1\}^n \to [0,1]$ als *Schätzfunktion* und $T_n(X_1, \ldots, X_n)$ als *Schätzstatistik* bezeichnet.

Beachtet man, dass $p = \mathbf{E}X_1$ gilt, so ist

$$T_n(x_1, \ldots, x_n) = \frac{x_1 + \cdots + x_n}{n}$$

eine naheliegende Schätzung von p. Diese hat die folgenden beiden Eigenschaften:

- Nach dem starken Gesetz der großen Zahlen gilt

  $$\frac{1}{n} \sum_{i=1}^{n} X_i \to \mathbf{E}X_1 = p \quad (n \to \infty) \quad f.s.,$$

  d.h. für großen Stichprobenumfang n nähert sich der geschätzte Wert mit Wahrscheinlichkeit Eins immer mehr dem "richtigen" Wert an.

  Man bezeichnet T_n daher als *stark konsistente Schätzung für* p.

- Weiter gilt
$$\mathbf{E}T_n(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}X_i = \mathbf{E}X_1 = p,$$
d.h. für festen Stichprobenumfang n ergibt sich im Mittel der "richtige" Wert, so dass der "richtige" Wert durch $T_n(x_1, \ldots, x_n)$ weder systematisch über- noch unterschätzt wird. Schätzfunktionen mit dieser Eigenschaft werden als *erwartungstreue* Schätzfunktionen bezeichnet.

Fragestellung 2: *Wie kann man ausgehend von* $(x_1, \ldots, x_n) \in \{0,1\}^n$ *ein (möglichst kleines) Intervall angeben, in dem* p *mit (möglichst großer) Wahrscheinlichkeit liegt ?*

Hierbei möchte man $x_1, \ldots, x_n$ zur Konstruktion eines Intervalls
$$[U(x_1, \ldots, x_n), O(x_1, \ldots, x_n)] \subseteq \mathbb{R}$$
verwenden, in dem der wahre Wert p mit möglichst großer Wahrscheinlichkeit liegt.

$[U(X_1, \ldots, X_n), O(X_1, \ldots, X_n)]$ heißt *Konfidenzintervall* zum *Konfidenzniveau* $1 - \alpha$, falls gilt
$$\mathbf{P}\left[p \in [U(X_1, \ldots, X_n), O(X_1, \ldots, X_n)]\right] \geq 1 - \alpha$$
für alle $p \in [0,1]$. Aufgrund von (5.1) hängen die bei der Berechnung der obigen Wahrscheinlichkeit verwendeten Zufallsvariablen von p ab.

Häufig wird hier $\alpha = 0.05$ bzw. $\alpha = 0.01$ gewählt, d.h. man fordert, dass der wahre Wert p mit Wahrscheinlichkeit $1 - \alpha = 0.95$ bzw. $1 - \alpha = 0.99$ im Konfidenzintervall liegt.

Fragestellung 3: *Wie kann man ausgehend von* $(x_1, \ldots, x_n) \in \{0,1\}^n$ *feststellen, ob der wahre Ausschussanteil* p *einen gewissen Wert* p_0 *überschreitet ?*

Hierbei möchte man zwischen zwei *Hypothesen*
$$H_0 : p \leq p_0 \quad \text{und} \quad H_1 : p > p_0$$
entscheiden. Ein *statistischer Test* dazu wird festgelegt durch Angabe eines *Ablehnungsbereichs* $K \subseteq \{0,1\}^n$ für H_0: Man lehnt H_0 ab, falls $(x_1, \ldots, x_n) \in K$, und man lehnt H_0 nicht ab, falls $(x_1, \ldots, x_n) \notin K$.

5.2 Punktschätzverfahren

Im Folgenden werden Verfahren vorgestellt, mit deren Hilfe man ausgehend von einer Stichprobe einer unbekannten Verteilung Kennzahlen (wie z.B. Erwartungswert oder Varianz) sowie Parameter eines angenommenen Verteilungsmodells (wie z.B. den Parameter λ der Exponentialverteilung) schätzen kann.

Ausgangspunkt sind Realisierungen $x_1, \ldots, x_n \in \mathbb{R}$ von unabhängigen identisch verteilten reellen ZVen $X_1, \ldots, X_n$. $x_1, \ldots, x_n$ wird als *Stichprobe* der Verteilung von X_1 bezeichnet. Die Verteilung $\mathbf{P}_{X_1}$ von X_1 sei unbekannt, es sei aber bekannt, dass diese aus einer vorgegebenen Klasse

$$\{w_\theta \ : \ \theta \in \Theta\}$$

von Verteilungen stammt.

Beispiel 5.2 *Gegeben sei eine Stichprobe einer Normalverteilung mit unbekanntem Erwartungswert und unbekannter Varianz. In diesem Fall ist $\Theta = \mathbb{R} \times \mathbb{R}_+$, und für $\theta = (\mu, \sigma) \in \Theta$ ist w_θ die Normalverteilung mit Erwartungswert μ und Varianz σ^2.*

Es sei g eine Funktion $g : \Theta \to \mathbb{R}$. Gesucht ist eine **Schätzfunktion** $T_n : \mathbb{R}^n \to \mathbb{R}$, mit deren Hilfe man ausgehend von der Stichprobe $x_1, \ldots, x_n$ den unbekannten Wert $g(\theta)$ durch $T_n(x_1, \ldots, x_n)$ schätzen kann.

Fortsetzung von Beispiel 5.2: *Interessiert man sich in Beispiel 5.2 für die Varianz der unbekannten Verteilung, so ist $g : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ definiert durch $g(\mu, \sigma) = \sigma^2$.*

Definition 5.1 **a)** *T_n heißt* **erwartungstreue** *Schätzung von $g(\theta)$, falls für alle $\theta \in \Theta$ gilt:*

$$\mathbf{E}_\theta T_n(X_1, \ldots, X_n) = g(\theta).$$

Dabei seien bei der Bildung des Erwartungswertes $\mathbf{E}_\theta$ die ZVen $X_1, \ldots, X_n$ unabhängig identisch verteilt mit $\mathbf{P}_{X_1} = w_\theta$.

b) *Eine Folge von Schätzfunktionen T_n heißt* **stark konsistente** *Schätzung von $g(\theta)$, falls für alle $\theta \in \Theta$ gilt:*

$$\mathbf{P}_\theta \left[\lim_{n\to\infty} T_n(X_1, \ldots, X_n) \neq g(\theta) \right] = 0.$$

Dabei seien bei der Bildung der Wahrscheinlichkeit $\mathbf{P}_\theta$ die ZVen $X_1, X_2, \ldots$ wieder unabhängig identisch verteilt mit $\mathbf{P}_{X_1} = w_\theta$.

Wir betrachten nun zunächst die Schätzung von Kennzahlen der zugrunde liegenden Verteilung wie z.B. Erwartungswert und Varianz.

Beispiel 5.3 *Wie schätzt man den Erwartungswert einer unbekannten Verteilung ?*

Die Schätzung

$$T_n(x_1,\ldots,x_n) = \frac{x_1+\cdots+x_n}{n}$$

ist erwartungstreu, da

$$\mathbf{E}T_n(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}X_i = \frac{1}{n}\sum_{i=1}^n \mathbf{E}X_1 = \mathbf{E}X_1.$$

Sie ist stark konsistent, da nach dem starken Gesetz der großen Zahlen gilt

$$T_n(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^n X_i \to \mathbf{E}X_1 \quad (n\to\infty) \quad f.s.$$

Beispiel 5.4 *Wie schätzt man die Varianz einer unbekannten Verteilung ?*

Die Idee bei der Konstruktion einer Schätzung von

$$V(X_1) = \mathbf{E}[(X_1-\mathbf{E}X_1)^2]$$

ist zunächst den Erwartungswert von $(X_1-\mathbf{E}X_1)^2$ wie oben durch

$$\frac{1}{n}\sum_{i=1}^n (X_i-\mathbf{E}X_1)^2$$

zu schätzen, und dann für den darin auftretenden Wert $\mathbf{E}X_1$ die Schätzung von oben zu verwenden. Dies führt auf

$$\bar{T}_n(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2 \quad \text{mit} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i.$$

$\bar{T}_n$ ist stark konsistent, da nach dem starken Gesetz der großen Zahlen gilt

$$\begin{aligned}
\bar{T}_n(X_1,\ldots,X_n) &= \frac{1}{n}\sum_{i=1}^n X_i^2 - 2\cdot\bar{X}\cdot\frac{1}{n}\sum_{i=1}^n X_i + \frac{1}{n}\sum_{i=1}^n \bar{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^n X_i^2 - 2\cdot\bar{X}\cdot\bar{X} + \bar{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 \\
&\to \mathbf{E}(X_1^2) - (\mathbf{E}X_1)^2 = V(X_1) \quad (n\to\infty) \quad f.s.
\end{aligned}$$

$\bar{T}_n$ ist aber nicht erwartungstreu, da

$$\begin{aligned}
\mathbf{E}\bar{T}_n(X_1,\ldots,X_n) &= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2\right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{E}X_i^2 - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbf{E}(X_iX_j) \\
&= \mathbf{E}X_1^2 - \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}(X_iX_i) - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1,\ldots,n,j\neq i}\mathbf{E}(X_iX_j) \\
&= \mathbf{E}X_1^2 - \frac{n}{n^2}\mathbf{E}(X_1^2) - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1,\ldots,n,j\neq i}\mathbf{E}(X_i)\cdot\mathbf{E}(X_j) \\
&\qquad\qquad \text{(vergleiche Satz 4.12)} \\
&= \left(1-\frac{1}{n}\right)\mathbf{E}X_1^2 - \left(1-\frac{1}{n}\right)(\mathbf{E}X_1)^2 \\
&= \frac{n-1}{n}\cdot V(X_1).
\end{aligned}$$

Aus obigem folgt aber, dass die Schätzung

$$\tilde{T}_n(x_1,\ldots,x_n) = \frac{n}{n-1}\cdot\bar{T}_n(x_1,\ldots,x_n) = \frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})^2 \quad \text{mit} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$$

stark konsistent und erwartungstreu ist.

Als nächstes wird eine systematische Methode zur Konstruktion von Schätzfunktionen spezifischer Parameter eines angenommenen Verteilungsmodels vorgestellt.

$X_1, \ldots, X_n$ seien unabhängige identisch verteilte reelle ZVen.

Fall 1: X_1 *sei eine diskrete Zufallsvariable* mit Werten z_1, z_2, ... Die Verteilung von X_1 sei w_θ, wobei $\theta \in \Theta \subseteq \mathbb{R}^l$ gelte.

Geschätzt werden soll θ aufgrund einer Realisierung $x_1, \ldots, x_n$ von $X_1, \ldots, X_n$.

Beispiel 5.5 *Eine Supermarktkette interessiert sich für die (zufällige) Zahl der Kunden, die in einer Niederlassung während der Mittagszeit (d.h. zwischen 12:30 Uhr und 13:30 Uhr) einkaufen.*

Geht man davon aus, dass im Einzugsbereich des Supermarktes insgesamt n Kunden leben, die sich unbeeinflusst voneinander mit Wahrscheinlichkeit $p \in (0,1)$ enstcheiden, um die Mittagszeit einzukaufen, so ist es naheliegend, die zufällige Zahl von Kunden durch eine Binomialverteilung mit Parametern n und p zu modellieren. Da n hier eher groß sein wird, bietet es sich an, diese Binomialverteilung durch eine Poisson-Verteilung mit Parameter $\theta = n \cdot p$ zu approximieren (vgl. Lemma 4.4). Daher wird im Folgenden angenommen, dass die zufällige Zahl X von Kunden während der Mittagszeit $\pi(\theta)$-verteilt ist, d.h. dass gilt

$$\mathbf{P}[X = k] = \frac{\theta^k}{k!} \cdot e^{-\theta} \quad (k \in \mathbb{N}_0).$$

In den vergangenen $n = 5$ Tagen kamen während der Mittagszeit $x_1 = 10$, $x_2 = 25$, $x_3 = 3$, $x_4 = 15$ und $x_5 = 7$ Kunden.

Wie schätzt man ausgehend von dieser Stichprobe den Wert von θ ?

In Beispiel 5.5 ist X_1 eine diskrete Zufallsvariable mit Werten 0, 1, 2, ..., für $\theta \in \Theta$ ist die Verteilung von X_1 eine $\pi(\theta)$-Verteilung, d.h.,

$$\mathbf{P}[X_1 = k] = w_\theta(\{k\}) = \frac{\theta^k}{k!} \cdot e^{-\theta}.$$

Geschätzt werden soll θ ausgehend von $x_1, \ldots, x_n$.

Für jeden festen Wert von θ kann man die Wahrscheinlichkeit bestimmen, dass gerade $x_1, \ldots, x_n$ als Realisierungen von $X_1, \ldots, X_n$ auftreten, falls diese Zufallsvariablen wirklich die Verteilung haben. Die Idee beim **Maximum–Likelihood–Prinzip** ist, als Schätzer für θ denjenigen Wert zu nehmen, bei dem die Wahrscheinlichkeit, dass gerade die beobachteten $x_1, \ldots, x_n$ als Realisierung der Zufallsvariablen $X_1, \ldots, X_n$ auftreten, maximal ist, d.h. bei dem

$$\mathbf{P}_\theta \left[X_1 = x_1, \ldots, X_n = x_n\right]$$

maximal ist.

Unter Ausnützung der Unabhängigkeit der ZVen $X_1, \ldots, X_n$ lässt sich die obige Wahrscheinlichkeit umschreiben zu

$$\begin{aligned}
\mathbf{P}_\theta \left[X_1 = x_1, \ldots, X_n = x_n\right] &= \mathbf{P}_\theta \left[X_1 = x_1\right] \cdot \ldots \cdot \mathbf{P}_\theta \left[X_n = x_n\right] \\
&= \prod_{i=1}^{n} w_\theta(\{x_i\}) \\
&=: L(\theta; x_1, \ldots, x_n).
\end{aligned}$$

$L(\theta; x_1, \ldots, x_n)$ ist die sogenannte *Likelihood–Funktion.*

Bei der **Maximum–Likelihood–Methode** verwendet man als Schätzer

$$\hat{\theta}(x_1, \ldots, x_n) = \arg\max_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n),$$

d.h., man verwendet als Schätzung dasjenige

$$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) \in \Theta,$$

für das gilt:

$$L\left(\hat{\theta}(x_1, \ldots, x_n); x_1, \ldots, x_n\right) = \max_{\theta \in \Theta} L\left(\theta; x_1, \ldots, x_n\right).$$

Fortsetzung von Beispiel 5.5: In Beispiel 5.5 sind $X_1, \ldots, X_n$ unabhängig identisch $\pi(\theta)$–verteilt, d.h. es gilt

$$\mathbf{P}[X_1 = k] = \frac{\theta^k}{k!} \cdot e^{-\theta} \quad (k \in \mathbb{N}_0).$$

Bestimmt werden soll der Maximum–Likelihood Schätzer (kurz: ML Schätzer) für θ.

Dazu muss die Likelihood–Funktion

$$L\left(\theta; x_1, \ldots, x_n\right) = \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} \cdot e^{-\theta} = e^{-n \cdot \theta} \cdot \frac{\theta^{x_1 + \ldots + x_n}}{x_1! \cdot \ldots \cdot x_n!}$$

bezüglich $\theta \in \mathbb{R}_+$ maximiert werden.

Beachtet man, dass für $x > 0$ die Funktion $\ln(x)$ streng monoton wachsend ist, so sieht man, dass

$$L(\theta; x_1, \ldots, x_n)$$

genau dann maximal wird, wenn

$$\ln L(\theta; x_1, \ldots, x_n)$$

maximal wird. Die Anwendung des Logarithmus führt hier zu einer Vereinfachung der Rechnung, da das Produkt

$$\prod_{i=1}^{n} w_\theta(\{x_i\})$$

in die Summe

$$\ln\left(\prod_{i=1}^{n} w_\theta(\{x_i\})\right) = \sum_{i=1}^{n} \ln w_\theta(\{x_i\})$$

umgewandelt wird. Sie ist aber nur möglich, sofern $L(\theta; x_1, \ldots, x_n)$ für alle θ ungleich Null ist.

Es genügt also im Folgenden,

$$\ln L(\theta; x_1, \ldots, x_n) = -n \cdot \theta + (x_1 + \ldots + x_n) \cdot \ln(\theta) - \ln(x_1! \cdot \ldots \cdot x_n!)$$

bezüglich θ zu maximieren.

Da diese Funktion im hier vorliegenden Beispiel differenzierbar ist, ist eine notwendige Bedingung dafür

$$\frac{\partial}{\partial\theta} \ln L(\theta; x_1, \ldots, x_n) = 0.$$

Mit

$$\frac{\partial}{\partial\theta} \ln L(\theta; x_1, \ldots, x_n) = -n + \frac{x_1 + \ldots + x_n}{\theta} - 0$$

folgt unter Beachtung von

$$L(\theta; x_1, \ldots, x_n) \to -\infty \quad \text{für} \quad \theta \to 0 \quad \text{oder} \quad \theta \to \infty,$$

dass der ML Schätzer gegeben ist durch

$$\hat{\theta}(x_1, \ldots, x_n) = \frac{x_1 + \ldots + x_n}{n}$$

Für die Daten aus Beispiel 5.5 ergibt sich damit

$$\hat{\theta}(x_1, \ldots, x_n) = \frac{10 + 25 + 3 + 15 + 7}{5} = 12$$

als Schätzung für θ.

Fall 2: *X_1 habe Dichte $f_\theta : \mathbb{R} \to [0,1]$, $\theta \in \Theta \subseteq \mathbb{R}^l$.*

In diesem Fall ist es nicht sinnvoll, θ durch Maximierung der Wahrscheinlichkeit

$$\mathbf{P}_\theta\left[X_1 = x_1, \ldots, X_n = x_n\right] = \prod_{i=1}^{n} \mathbf{P}_\theta\left[X_i = x_i\right]$$

zu bestimmen, da diese Wahrscheinlichkeit für alle Werte $x_1, \ldots, x_n$ Null ist. Statt dessen definiert man die Likelihood–Funktion durch

$$L(\theta; x_1, \ldots, x_n) := \prod_{i=1}^{n} f_\theta(x_i),$$

d.h. anstelle des Produktes der Wahrscheinlichkeiten betrachtet man das Produkt der Werte der Dichten an den Stellen $x_1, \ldots, x_n$, und wählt den Maximum–Likelihood–Schätzer wieder durch Maximierung der Likelihood–Funktion bzgl. θ.

Beispiel 5.6 *Student S. fährt immer mit dem Auto zur Universität. Auf dem Weg dorthin passiert er eine Ampelanlage. In der Vergangenheit war diese mehrfach rot, wobei die letzten $n = 6$ Wartezeiten $x_1 = 10$, $x_2 = 60$, $x_3 = 45$, $x_4 = 50$, $x_5 = 5$ und $x_6 = 30$ Sekunden betrugen. Da das Eintreffen von Student S. an der Ampel als rein zufällig innerhalb der Rotphase der Ampel (vorausgesetzt die Ampel ist nicht grün!) betrachtet werden kann, ist es naheliegend, die zufällige Wartezeit X von Student S. an der roten Ampel durch eine auf einem Intervall $[0, a]$ gleichverteilte Zufallsvariable X zu modellieren, d.h. durch eine stetig verteilte Zufallsvariable X mit Dichte*

$$f_a(x) = \begin{cases} \frac{1}{a} & \text{für} \quad 0 \le x \le a, \\ 0 & \text{für} \quad x < 0 \text{ oder } x > a. \end{cases}$$

Wie schätzt man ausgehend von den obigen Daten die Dauer a der Rotphase ?

Anwendung des Maximum-Likelihood Prinzips erfordert hier Maximierung von

$$L(a) = \prod_{i=1}^{n} f_a(x_i).$$

Zur Bestimmung der Werte von $L(a)$ bietet sich die folgende Überlegung an: $L(a)$ ist Null, falls einer der Faktoren Null ist. Ist dies nicht der Fall, sind alle $f_a(x_i)$ gleich $1/a$ und damit ist $L(a) = 1/a^n$.

Da $f_a(x_i)$ für $x_i \geq 0$ genau dann Null ist, falls $a < x_i$ gilt, folgt, dass $L(a)$ genau dann Null ist, falls $a < \max\{x_1, \ldots, x_n\}$ ist.

Insgesamt ist damit gezeigt:

$$L(a) = \begin{cases} \frac{1}{a^n} & \text{für} \quad a \geq \max\{x_1, \ldots, x_n\}, \\ 0 & \text{für} \quad a < \max\{x_1, \ldots, x_n\}. \end{cases}$$

Damit wird $L(a)$ maximal für

$$\hat{a} = \max\{x_1, \ldots, x_n\},$$

und im Falle der Daten aus Beispiel 5.6 liefert das Maximum-Likelihood Prinzip die Schätzung

$$\hat{a} = \max\{10, 60, 45, 50, 5, 30\} = 60.$$

Beispiel 5.7 *$X_1, \ldots, X_n$ seien unabhängig identisch $N(a, \sigma^2)$–verteilt, d.h. X_1 hat die Dichte*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Hierbei sei $a \in \mathbb{R}$ bekannt und $\sigma > 0$ unbekannt. Geschätzt werden soll $\theta = \sigma^2$.

In diesem Fall ist die Likelihood–Funktion gegeben durch

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \cdot \theta}} \cdot e^{-\frac{(x_i-a)^2}{2\theta}} = (2\pi)^{-n/2} \theta^{-n/2} e^{-\frac{\sum_{i=1}^n (x_i-a)^2}{2\theta}}.$$

Maximierung von $\ln L(\theta; x_1, \ldots, x_n)$ bzgl. θ führt auf

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial}{\partial \theta} [\ln L(\theta; x_1, \ldots, x_n)] \\ &= \frac{\partial}{\partial \theta} \left[\ln\left((2\pi)^{-n/2}\right) - \frac{n}{2} \cdot \ln(\theta) - \frac{\sum_{i=1}^n (x_i - a)^2}{2\theta} \right] \\ &= -\frac{n}{2} \cdot \frac{1}{\theta} + \frac{\sum_{i=1}^n (x_i - a)^2}{2\theta^2}. \end{aligned}$$

Unter Beachtung von

$$L(\theta; x_1, \ldots, x_n) \to 0 \quad \text{für} \quad \theta \to 0 \quad \text{oder} \quad \theta \to \infty$$

ergibt sich damit der Maximum–Likelihood–Schätzer zu

$$\hat{\theta}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - a)^2.$$

Beispiel 5.8 *Wir betrachten nochmals Beipiel 5.7, d.h., $X_1, \ldots, X_n$ seien wieder unabhängig identisch $N(\mu, \sigma^2)$ verteilt, aber diesmal seien sowohl $\mu \in \mathbb{R}$ als auch $\sigma > 0$ unbekannt. Geschätzt werden soll $\theta = (\mu, \sigma^2)$.*

Die Likelihood–Funktion ist hier gegeben durch

$$\begin{aligned} L(\theta; x_1, \ldots, x_n) &= L((\mu, \sigma^2); x_1, \ldots, x_n) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} e^{-\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}}. \end{aligned}$$

Maximierung von

$$\ln L((\mu, \sigma^2); x_1, \ldots, x_n) = \ln\left((2\pi)^{-n/2}\right) - \frac{n}{2} \cdot \ln(\sigma^2) - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}$$

führt auf

$$0 \stackrel{!}{=} \frac{\partial \ln(L((\mu, \sigma^2); x_1, \ldots, x_n))}{\partial \mu} = \frac{\sum_{i=1}^{n} (x_i - \mu)}{\sigma^2},$$

was äquivalent ist zu

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

sowie auf

$$0 \stackrel{!}{=} \frac{\partial \ln(L((\mu, \sigma^2); x_1, \ldots, x_n))}{\partial \sigma^2} = -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2(\sigma^2)^2},$$

woraus folgt

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Damit ergibt sich der Maximum–Likelihood–Schätzer zu

$$(\hat{\mu}, \hat{\sigma}^2) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2 \right).$$

5.3 Statistische Testverfahren

Statistische Testverfahren werden anhand der folgenden Fragestellung eingeführt.

Beispiel 5.9 *Wie kann man feststellen, ob eine geplante Vereinfachung des Steuerrechts zu Mindereinnahmen des Staates führt oder nicht ?*

Eine naheliegende Idee ist, für n zufällig ausgewählte Steuererklärungen des vergangenen Jahres die Differenzen

$$\begin{aligned} x_i &= \text{Steuer im Fall } i \text{ bei neuem Steuerrecht} \\ &\quad -\text{Steuer im Fall } i \text{ bei altem Steuerrecht} \end{aligned}$$

$(i = 1, \ldots, n)$ zu berechnen. Ist hierbei $x_i > 0$, so erhält der Staat bei der i–ten Steuererklärung nach dem neuen Recht mehr Geld; im Falle $x_i < 0$ erhält er weniger Geld.

Ein naiver Zugang ist nun,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

zu betrachten und im Falle $\bar{x} < 0$ zu schließen, dass die Steuerreform die Einnahmen des Staates verringern wird, und im Fall $\bar{x} \geq 0$ zu schließen, dass dies nicht der Fall ist. Es stellt sich aber sofort die Frage, ob das Ergebnis hier nicht aufgrund der zufälligen Auswahl der n betrachteten Steuererklärungen (statt aufgrund des Einflusses der Steuerreform) enstanden ist. Diese zufällige Auswahl hat vor allem dann einen großen Einfluss, wenn $\bar{x}$ "nahe bei" Null ist, n "klein" ist (für große n würden sich zufällige Schwankungen bei den Werten der x_i bei der Bildung des arithmetischen Mittels $\bar{x}$ "herausmitteln") und wenn die Differenzen der zu zahlenden Steuern in der Menge aller Steuerpflichtigen stark schwanken. Letzteres kann durch Betrachtung der Streuung

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

beurteilt werden, die (aufgrund der zufälligen Auswahl der n Steuerpflichtigen) eine Schätzung für die Streuung der Differenzen der zu zahlenden Steuern in der Menge aller Steuerpflichtigen darstellt.

Man steht dann vor dem Problem, ausgehend von der Größe von $\bar{x}$, s^2 und vom Stichprobenumfang n zu entscheiden, ob die Steuerreform die Einnahmen des Staates vermindert oder nicht.

Zahlenbeispiel zu Beispiel 5.9: $n = 100$, $\bar{x} = 120$ und $s = 725$. Was folgt daraus ?

Wir modellieren die Fragestellung stochastisch wie folgt: Wir fassen die $x_1, \ldots, x_n$ als Realisierungen von unabhängigen identisch verteilten reellen ZVen $X_1, \ldots, X_n$ auf. Aufgrund dieser Realisierungen möchten wir entscheiden, ob $\mathbf{E}X_1$ kleiner als Null ist oder nicht.

Zwecks Vereinfachung der Problemstellung schränken wir die Klasse der betrachteten Verteilungen ein. Wir nehmen an, dass die Verteilung $\mathbf{P}_{X_1}$ von X_1 aus einer gegebenen Klasse

$$\{w_\theta \ : \ \theta \in \Theta\}$$

von Verteilungen stammt.

Im Beispiel 5.9 könnte man z.B. annehmen, dass X_1 normalverteilt ist mit unbekanntem Erwartungswert μ und bekannter Varianz $\sigma^2 = s^2$, oder dass X_1 normalverteilt ist mit unbekanntem Erwartungswert μ und unbekannter Varianz σ^2.

Wir betrachten eine Aufteilung der Parametermenge Θ in zwei Teile:

$$\Theta = \Theta_0 \cup \Theta_1 \quad \text{wobei } \Theta_0 \neq \emptyset, \Theta_1 \neq \emptyset \text{ und } \Theta_0 \cap \Theta_1 = \emptyset.$$

Die Aufgabe ist, aufgrund von $x_1, \ldots, x_n$ zu entscheiden ("testen"), ob die sogenannte *Nullhypothese*

$$H_0 : \theta \in \Theta_0$$

abgelehnt, d.h. die sogenannte *Alternativhypothese*

$$H_1 : \theta \in \Theta_1$$

angenommen werden kann, oder nicht.

In Beispiel 5.9 wollen wir uns zwischen den *Hypothesen*

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0$$

mit $\mu_0 = 0$ entscheiden. Dabei bedeutet $\mu \leq 0$, dass die Steuerreform die Steuereinnahmen vermindert, während $\mu > 0$ bedeutet, dass dies nicht der Fall ist.

Andere häufig auftretende Beispiele für das Aufteilen der Parametermenge Θ in zwei Mengen Θ_0 und Θ_1 sind

$$H_0 : \mu \geq \mu_0 \quad \text{und} \quad H_1 : \mu < \mu_0$$

oder

$$H_0 : \mu = \mu_0 \quad \text{und} \quad H_1 : \mu \neq \mu_0.$$

Bei Letzterem interessieren sowohl Abweichungen von μ_0 nach oben als auch Abweichungen nach unten und man spricht daher von einem *zweiseitigen Testproblem*. Bei den anderen beiden Beispielen handelt es sich um sogenannte *einseitige Testprobleme*. Hier möchte man entweder eine Abweichung von μ_0 nach oben oder eine Abweichung nach unten feststellen.

Durch Angabe eines *Ablehnungsbereichs* (oder *kritischen Bereichs*) $K \subseteq \mathbb{R}^n$ ist ein **statistischer Test** festgelegt:

H_0 wird abgelehnt, falls $(x_1, \ldots, x_n) \in K$. Ist dagegen $(x_1, \ldots, x_n) \notin K$, so wird H_0 nicht abgelehnt.

Bei einem solchen Test können zwei Arten von Fehlern auftreten:

Ein *Fehler 1. Art* ist die Entscheidung für H_1, obwohl H_0 richtig ist. Ein *Fehler 2. Art* ist die Entscheidung für H_0, obwohl H_1 richtig ist.

In Beispiel 5.9 bedeutet das Auftreten eines Fehlers 1. Art, dass wir zu dem Schluss kommen, dass die Steuerreform die Steuereinnahmen nicht vermindert, obwohl sie dass in Wahrheit tut. Dagegen bedeutet ein Fehler 2. Art, dass wir bei Vorliegen einer Steuerreform, die die Steuereinnahmen nicht vermindert, zum Schluss kommen, dass die Steuerreform die Steuereinnahmen verringert.

Die Funktion $g : \Theta \to [0, 1]$ mit

$$g(\theta) = \mathbf{P}_\theta \left[(X_1, \ldots, X_n) \in K \right]$$

heißt *Gütefunktion* des Tests. Hierbei gibt $\mathbf{P}_\theta \left[(X_1, \ldots, X_n) \in K \right]$ die Wahrscheinlichkeit an, dass H_0 abgelehnt wird; die obige Wahrscheinlichkeit wird berechnet für unabhängig identisch verteilte ZVen $X_1, \ldots, X_n$ mit $\mathbf{P}_{X_1} = w_\theta$.

Im Fall $\theta \in \Theta_0$ gilt:

$$\begin{aligned} g(\theta) &= \text{Wahrscheinlichkeit, } H_0 \text{ abzulehnen obwohl } H_0 \text{ richtig ist} \\ &=: \textit{Fehlerwahrscheinlichkeit 1. Art.} \end{aligned}$$

Im Fall $\theta \in \Theta_1$ gilt:

$$\begin{aligned} 1 - g(\theta) &= \mathbf{P}_\theta \left[(X_1, \ldots, X_n) \notin K \right] \\ &= \text{Wahrscheinlichkeit, } H_0 \text{ nicht abzulehnen obwohl } H_1 \text{ richtig ist} \\ &=: \textit{Fehlerwahrscheinlichkeit 2. Art.} \end{aligned}$$

Die ideale Gütefunktion ist gegeben durch

$$g(\theta) = \begin{cases} 0, & \text{falls } \theta \in \Theta_0, \\ 1, & \text{falls } \theta \in \Theta_1. \end{cases}$$

Leider existieren nur in trivialen Fällen Tests mit dieser Gütefunktion. Darüberhinaus existieren im allgemeinen auch keine Tests, die die Fehlerwahrscheinlichkeiten 1. und 2. Art gleichmäßig bzgl. $\theta \in \Theta$ minimieren.

Als Ausweg bietet sich eine asymmetrische Betrachtungsweise der Fehler erster und zweiter Art an. In vielen Anwendungen ist eine der beiden Fehlerarten als schwerwiegender zu betrachten als die andere. Z.B. führt in Beispiel 5.9 ein Fehler erster Art (Entscheidung für $\mu > 0$ obwohl $\mu \leq 0$ gilt) zur Durchführung einer Steuerreform, die die Steuereinnahmen vermindert. Aus Sicht des Finanzministers ist dies ein deutlich schwerwiegender Fehler als ein Fehler zweiter Art, der dazu führt, dass eine Steuerreform, die die Einnahmen des Staates nicht vermindert, nicht durchgeführt wird.

Was man daher macht, ist eine Schranke für eine der beiden Arten von Fehlerwahrscheinlichkeiten vorzugeben und unter dieser Nebenbedingung die andere Art von Fehlerwahrscheinlichkeiten zu minimieren. OBdA gibt man hierbei eine Schranke für die Fehlerwahrscheinlichkeit erster Art vor.

Dazu gibt man ein $\alpha \in (0,1)$ vor (sog. *Niveau*, meist wählt man $\alpha = 0.05$ oder $\alpha = 0.01$) und betrachtet nur noch Tests mit Fehlerwahrscheinlichkeiten 1. Art $\leq \alpha$, d.h. mit

$$g(\theta) \leq \alpha \quad \text{für alle } \theta \in \Theta_0$$

(sog. *Tests zum Niveau* α).

Unter allen Tests zum Niveau α sucht man dann denjenigen Test, für den für *alle* $\theta \in \Theta_1$ die zugehörige Fehlerwahrscheinlichkeit 2. Art $1 - g(\theta)$ am kleinsten ist.

Der Ablehnungsbereich solcher Tests hat häufig die Form

$$K = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \;:\; T(x_1, \ldots, x_n) > c\}$$

(evt. mit $> c$ ersetzt durch $< c$) für eine Funktion $T : \mathbb{R}^n \to \mathbb{R}$ und ein $c \in \mathbb{R}$. Die Zufallsvariable $T(X_1, \ldots, X_n)$ heißt in diesem Fall *Testgröße* oder *Teststatistik*, c heißt *kritischer Wert*.

Bemerkungen:

a) Bei den obigen Tests werden die Fehlerwahrscheinlichkeiten 1. und 2. Art unsymmetrisch behandelt. Als Konsequenz sollte man die Hypothesen so wählen, dass der Fehler erster Art als schlimmer angesehen wird als der Fehler zweiter Art, bzw. dass das statistisch zu sichernde Resultat als Alternativhypothese formuliert wird.

b) Aufgrund der Konstruktion der obigen Tests wird bei einem Test zum Niveau $\alpha = 5\%$ bei wiederholtem Durchführen des Tests für unabhängige Daten bei Gültigkeit von H_0 in bis zu 5% der Fälle H_0 fälschlicherweise abgelehnt.

c) Führt man mehrere verschiedene Tests zum Niveau α hintereinander aus, und gelten jeweils die Nullhypothesen, so ist die Wahrscheinlichkeit, mindestens bei einem dieser Tests die Nullhypothese abzulehnen, im allgemeinen größer als α. Sind z.B. die Prüfgrößen der einzelnen Tests unabhängig und ist der Fehler erster Art bei jedem der Tests genau $\alpha = 0.05$, so ist beim Durchführen von $n = 3$ solchen Tests die Wahrscheinlichkeit, kein einziges Mal die Nullhypothese abzulehnen, gegeben durch

$$(1 - \alpha)^n,$$

d.h., die Wahrscheinlichkeit, bei mindestens einen der Tests die Nullhypothese abzulehnen, beträgt

$$1 - (1 - \alpha)^n = 1 - 0.95^3 \approx 0.14$$

(sog. *Problem des multiplen Testens*).

d) Betrachtet man erst die Daten, und wählt dann einen zu diesen Daten passenden Test aus, so führt dies analog zu b) eventuell zu einem Verfälschen des Niveaus.

e) Häufig betrachtet man den sogenannten *p–Wert*

$$p = \max_{\theta \in \Theta_0} \mathbf{P}[T(X_1, \ldots, X_n) > T(x_1, \ldots, x_n)]$$

eines Tests. Dieser gibt dasjenige Niveau an, bei dem die Nullhypothese H_0 bei den gegebenen Daten gerade noch abgelehnt werden kann. Ist das vorgegebene Niveau α größer oder gleich dem p–Wert, so kann H_0 zum Niveau α abgelehnt werden, andernfalls kann H_0 nicht abgelehnt werden.

Man beachte, dass der p–Wert *nicht* die Wahrscheinlichkeit angibt, dass die Nullhypothese falsch ist. Denn in dem oben beschriebenen Modell für statistische Tests ist diese entweder richtig oder falsch, daher gibt es keine Wahrscheinlichkeit zwischen Null und Eins, mit der diese richtig ist.

Beispiele:

a) Einseitiger Gauß-Test

Hier wird davon ausgegangen, dass die ZVen $X_1, \ldots, X_n$ unabhängig identisch $N(\mu, \sigma_0^2)$–verteilt sind, wobei $\mu \in \mathbb{R}$ unbekannt ist und $\sigma_0 > 0$ bekannt ist.

Zu testen sei

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0.$$

Als Testgröße wird verwendet

$$T(X_1, \ldots, X_n) = \frac{\sqrt{n}}{\sigma_0} \left(\bar{X}_n - \mu_0 \right)$$

mit

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

Da $\bar{X}_n$ ein Schätzer für μ ist, werden die Werte von $T(X_1, \ldots, X_n)$ (mit großer Wk.) umso größer sein, je größer μ ist. Sinnvollerweise entscheidet man sich daher vor allem dann für eine Ablehnung von $H_0 : \mu \leq \mu_0$, wenn der Wert von $T(X_1, \ldots, X_n)$ groß ist.

Beim einseitigen Gauß-Test wird H_0 abgelehnt, falls $(x_1, \ldots, x_n)$ im Ablehnungsbereich

$$K = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \; : \; T(x_1, \ldots, x_n) > c\}$$

enthalten ist.

Zur Bestimmung von c wird wie folgt vorgegangen:

Man kann zeigen, dass Linearkombinationen von unabhängigen normalverteilten ZVen normalverteilt sind. Daher ist für $\mu = \mu_0$ die Testgröße $T(X_1, \ldots, X_n)$ $N(0, 1)$–verteilt, da

$$\mathbf{E}_{\mu_0} T(X_1, \ldots, X_n) = \frac{\sqrt{n}}{\sigma_0} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{E}_{\mu_0} X_i - \mu_0 \right) = \frac{\sqrt{n}}{\sigma_0} \left(\frac{1}{n} \sum_{i=1}^{n} \mu_0 - \mu_0 \right) = 0$$

und

$$V_{\mu_0} \left(T(X_1, \ldots, X_n) \right) = \left(\frac{\sqrt{n}}{\sigma_0} \right)^2 \frac{1}{n^2} \sum_{i=1}^{n} \sigma_0^2 = 1.$$

Sei $\alpha \in (0,1)$ das vorgegebene Niveau. Dann wählt man c so, dass die Fehlerwahrscheinlichkeit erster Art des Tests im Falle $\mu = \mu_0$ gerade gleich α ist, d.h., dass gilt:

$$\mathbf{P}_{\mu_0}\left[(X_1, \ldots, X_n) \in K\right] = \alpha,$$

bzw.

$$\mathbf{P}_{\mu_0}\left[\frac{\sqrt{n}}{\sigma_0}\left(\bar{X}_n - \mu_0\right) > c\right] = \alpha.$$

Die linke Seite oben ist gleich $1 - \Phi(c)$, wobei Φ die Verteilungsfunktion zur $N(0,1)$–Verteilung ist. Also ist die obige Forderung äquivalent zu

$$1 - \Phi(c) = \alpha \text{ bzw. } \Phi(c) = 1 - \alpha$$

(d.h. c ist das sogenannte α*–Fraktil* der $N(0,1)$–Verteilung).

Aus dieser Beziehung kann man c z.B. unter Zuhilfenahme von Tabellen für die Verteilungsfunktion bzw. die Fraktile der $N(0,1)$–Verteilung bestimmen.

Für diese Wahl von c gilt, dass der resultierende Test ein Test zum Niveau α ist. Ist nämlich $\mu = \bar{\mu}$ für ein $\bar{\mu} \in \mathbb{R}$, so ist

$$\frac{\sqrt{n}}{\sigma_0}(\bar{X}_n - \bar{\mu})$$

$N(0,1)$-verteilt, und daher gilt für die Gütefunktion des obigen Tests:

$$\begin{aligned}
g(\bar{\mu}) &= \mathbf{P}_{\bar{\mu}}\left[\frac{\sqrt{n}}{\sigma_0}(\bar{X}_n - \mu_0) > c\right] \\
&= \mathbf{P}_{\bar{\mu}}\left[\frac{\sqrt{n}}{\sigma_0}(\bar{X}_n - \bar{\mu}) + \frac{\sqrt{n}}{\sigma_0}(\bar{\mu} - \mu_0) > c\right] \\
&= \mathbf{P}_{\bar{\mu}}\left[\frac{\sqrt{n}}{\sigma_0}(\bar{X}_n - \bar{\mu}) > c + \frac{\sqrt{n}}{\sigma_0}(\mu_0 - \bar{\mu})\right] \\
&= 1 - \Phi\left(c + \frac{\sqrt{n}}{\sigma_0}(\mu_0 - \bar{\mu})\right).
\end{aligned}$$

Also ist für $\bar{\mu} \le \mu_0$ die Fehlerwahrscheinlichkeit erster Art des einseitigen Gauß-Tests wegen

$$c + \frac{\sqrt{n}}{\sigma_0}(\mu_0 - \bar{\mu}) \ge c$$

und Φ monoton wachsend gegeben durch

$$g(\bar{\mu}) = 1 - \Phi\left(c + \frac{\sqrt{n}}{\sigma_0}(\mu_0 - \bar{\mu})\right) \le 1 - \Phi(c) = \alpha,$$

d.h. alle Fehlerwahrscheinlichkeiten erster Art sind kleiner oder gleich α.

Aus der obigen Überlegung sieht man auch, dass für $\bar{\mu} > \mu_0$ die Fehlerwahrscheinlichkeit zweiter Art gleich

$$1 - g(\bar{\mu}) = \Phi\left(c + \frac{\sqrt{n}}{\sigma_0}(\mu_0 - \bar{\mu})\right)$$

ist, d.h. für $\bar{\mu}$ nahe bei μ_0 nahe bei

$$\Phi(c) = 1 - \alpha$$

sowie für $\bar{\mu}$ sehr groß nahe bei

$$\lim_{x \to -\infty} \Phi(x) = 0$$

ist.

Anwendung in Beispiel 5.9 mit $\mu_0 = 0$ und $\alpha = 5\%$ ergibt $1 - \Phi(c) = 0.05$ bzw. $\Phi(c) = 0.95$, woraus $c \approx 1.645$ folgt. In Beispiel 5.9 war $n = 100$, $\bar{x} = 120$ und $\sigma_0 = s = 725$. Wegen

$$\frac{\sqrt{n}}{\sigma_0}(\bar{x} - \mu_0) = \frac{\sqrt{100}}{725}(120 - 0) \approx 1.655 > c$$

kann hier H_0 abgelehnt werden, d.h. man kommt zur Schlussfolgerung, dass die Steuerreform die Steuereinnahmen vermutlich nicht vermindert.

Der obige einseitige Gauß-Test kann nach naheliegender Modifikation auch zum Testen der Hypothesen

$$H_0 : \mu \geq \mu_0 \quad \text{versus} \quad H_1 : \mu < \mu_0$$

verwendet werden. Dazu beachte man, dass bei der obigen Testgröße große (bzw. kleine) Werte eine Entscheidung für große (bzw. kleine) Werte von μ nahelegen. Daher entscheidet man sich jetzt für Ablehnung von $H_0 : \mu \geq \mu_0$, falls $(x_1, \ldots, x_n)$ im Ablehnungsbereich

$$K = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \ : \ T(x_1, \ldots, x_n) < c\}$$

enthalten ist. c wird dabei wieder so gewählt, dass für $\mu = \mu_0$ die Fehlerwahrscheinlichkeit erster Art gleich α ist, d.h. dass gilt:

$$\mathbf{P}_{\mu_0}\left[\frac{\sqrt{n}}{\sigma_0}\left(\bar{X}_n - \mu_0\right) < c\right] = \alpha.$$

Analog zu oben folgt daraus

$$\Phi(c) = \alpha,$$

d.h. c wird hier als $(1-\alpha)$-Fraktil der $N(0,1)$-Verteilung gewählt.

Problematisch bei Anwendung des Gauß-Tests in Beispiel 5.9 ist, dass die Varianz eigentlich unbekannt war und aus den Daten geschätzt wurde und damit die Voraussetzungen des Gauß-Tests nicht erfüllt waren.

Daher ist eigentlich eine Anwendung des sogenannten t–Tests nötig, der als nächstes behandelt wird.

b) Einseitiger t–Test

Hier wird davon ausgegangen, dass die ZVen $X_1, \ldots, X_n$ unabhängig identisch $N(\mu, \sigma)$–verteilt sind, wobei $\mu \in \mathbb{R}$ und $\sigma > 0$ **beide** unbekannt sind.

Zu testen sei wieder

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0.$$

Als Testgröße wird

$$T(X_1, \ldots, X_n) = \sqrt{n} \cdot \frac{(\bar{X}_n - \mu_0)}{S_n}$$

verwendet, wobei

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{und} \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} |X_i - \bar{X}|^2.$$

Die Testgröße wird also analog zum Gauß-Test bestimmt, nur dass jetzt anstelle der Varianz σ_0 eine Schätzung derselbigen verwendet wird.

Wie bei der Testgröße des einseitigen Gauß-Tests gilt auch hier, dass die Werte von $T(X_1, \ldots, X_n)$ (mit großer Wk.) umso größer sind, je größer μ ist.

H_0 wird wieder abgelehnt, falls $(x_1, \ldots, x_n)$ im Ablehnungsbereich

$$K = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \ : \ T(x_1, \ldots, x_n) > c\}$$

enthalten ist.

Ausgangspunkt zur Bestimmung des Wertes von c ist, dass für $\mu = \mu_0$ die Testgröße

$$\sqrt{n} \cdot \frac{(\bar{X}_n - \mu_0)}{S_n}$$

t_{n-1}–verteilt ist, wobei man eine t–verteilte Zufallsvariable mit $n-1$ Freiheitsgraden (kurz: eine t_{n-1}–verteilte ZV) erhält, indem man ausgehend von unabhängig identisch $N(0,1)$–verteilten ZVen $Y_1, \ldots, Y_n$ die ZV

$$\frac{Y_n}{\sqrt{(Y_1^2 + \cdots + Y_{n-1}^2)/(n-1)}}$$

bildet. Die Verteilungsfunktion der t_{n-1}–Verteilung ist tabelliert.

Man wählt nun c so, dass

$$\mathbf{P}_{\mu_0}\left[\frac{\sqrt{n}}{S_n}\left(\bar{X}_n - \mu_0\right) > c\right] = \alpha$$

gilt.

Anwendung in Beispiel 5.9 mit $\mu_0 = 0$ und $\alpha = 5\%$ ergibt $c = 1.660$. Mit $n = 100$, $\bar{x} = 120$ und $s = 725$ folgt

$$\sqrt{n}\frac{(\bar{x} - \mu_0)}{s} = \sqrt{100}\frac{(120 - 0)}{725} \approx 1.655 < c,$$

d.h. H_0 kann nicht abgelehnt werden und man kommt nun zur Schlussfolgerung, dass die Steuerreform die Steuereinnahmen vermutlich vermindert.

Im Vergleich mit der Anwendung des einseitigen Gauß-Test fällt auf, dass der kritische Wert c jetzt größer ist und daher die Nullhypothese seltener abgelehnt wird. Dies liegt daran, dass beim t–Test die Varianz als unbekannt vorausgesetzt wird, damit weniger Informationen über die zugrundeliegende Verteilung bekannt sind und man sich daher seltener für die Ablehnung der Nullhypothese entscheiden muss, um sicherzustellen, dass eine fälschliche Ablehung der Nullhypothese nur mit Wahrscheinlichkeit α erfolgt.

Der einseitige t-Test kann analog zum einseitigen Gauß-Test auch zum Testen der Hypothesen

$$H_0 : \mu \geq \mu_0 \quad \text{und} \quad H_1 : \mu \leq \mu_0.$$

verwendet werden.

c) Zweiseitiger Gauß- bzw. t-Test.

Zu testen ist hier

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0,$$

wobei die Stichprobe wieder normalverteilt mit unbekanntem Erwartungswert μ und bekannter bzw. unbekannter Varianz σ_0^2 bzw. σ^2 ist. Die Teststatistik T

wird wie beim einseitigen Gauß- bzw. t-Test gebildet. H_0 wird abgelehnt, falls $(x_1, \ldots, x_n)$ im Ablehnungsbereich

$$K = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \;:\; |T(x_1, \ldots, x_n)| > c\}$$

enthalten ist, wobei c durch die Forderung

$$\mathbf{P}_{\mu_0}\left[\left|\frac{\sqrt{n}}{\sigma_0}\left(\bar{X}_n - \mu_0\right)\right| > c\right] = \alpha.$$

bestimmt wird. Da hier $T(X_1, \ldots, X_n)$ die gleiche Verteilung hat wie $(-1) \cdot T(X_1, \ldots, X_n)$, ist dies äquivalent zu

$$\mathbf{P}_{\mu_0}\left[\frac{\sqrt{n}}{\sigma_0}\left(\bar{X}_n - \mu_0\right) > c\right] = \frac{\alpha}{2},$$

und c ergibt sich im Falle des zweiseitigen Gauß-Test, bei dem die Varianz als bekannt vorausgesetzt wird, als $\alpha/2$-Fraktil der $N(0,1)$–Verteilung, und im Falle des zweiseitigen t-Tests, bei dem die Varianz unbekannt ist, als $\alpha/2$-Fraktil der t_{n-1}–Verteilung.

Eine Übersicht über die bisher eingeführten Tests findet man in Tabelle 5.1.

Hypothesen	Varianz	$T(x_1, \ldots, x_n)$	Ablehnung von H_0, falls
$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$	bekannt	$\sqrt{n} \cdot \frac{\bar{x}_n - \mu_0}{\sigma_0}$	$T(x_1, \ldots, x_n) > u_\alpha$
$H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$	bekannt	$\sqrt{n} \cdot \frac{\bar{x}_n - \mu_0}{\sigma_0}$	$T(x_1, \ldots, x_n) < u_{1-\alpha}$
$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$	bekannt	$\sqrt{n} \cdot \frac{\bar{x}_n - \mu_0}{\sigma_0}$	$\lvert T(x_1, \ldots, x_n)\rvert > u_{\alpha/2}$
$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$	unbekannt	$\sqrt{n} \cdot \frac{\bar{x}_n - \mu_0}{s_n}$	$T(x_1, \ldots, x_n) > t_{n-1,\alpha}$
$H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$	unbekannt	$\sqrt{n} \cdot \frac{\bar{x}_n - \mu_0}{s_n}$	$T(x_1, \ldots, x_n) < t_{n-1,1-\alpha}$
$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$	unbekannt	$\sqrt{n} \cdot \frac{\bar{x}_n - \mu_0}{s_n}$	$\lvert T(x_1, \ldots, x_n)\rvert > t_{n-1,\alpha/2}$

Tabelle 5.1: Gauß- und t–Test für eine Stichprobe. Vorausgesetzt ist jeweils, dass $x_1, \ldots, x_n$ eine Stichprobe einer Normalverteilung mit unbekanntem Erwartungswert μ und bekannter Varianz σ_0^2 bzw. unbekannter Varianz σ^2 sind. u_α bzw. $t_{n-1,\alpha}$ ist das α-Fraktil der $N(0,1)$– bzw. der t_{n-1}–Verteilung. Es werden die Abkürzungen $\bar{x}_n = 1/n \sum_{i=1}^n x_i$ und $s_n^2 = 1/(n-1) \sum_{i=1}^n (x_i - \bar{x}_n)^2$ verwendet.

Bei den obigen Test wurde der Erwartungswert mit einem festen Wert verglichen. Manchmal ist allerdings kein solcher Wert vorgegeben, statt dessen hat man Stichproben zweier unterschiedlicher Verteilungen gegeben und möchte deren (unbekannte) Erwartungswerte vergleichen. Die zugehörigen Tests bezeichnet man als *Tests für zwei Stichproben* (im Gegensatz zu den oben vorgestellten *Tests für eine Stichprobe*).

Beispiel 5.10 *Im Rahmen einer prospektiv kontrollierten Studie mit Randomisierung soll die Wirksamkeit eines Medikamentes überprüft werden. Dazu werden die Überlebenszeiten $x_1, \ldots, x_n$ der Studiengruppe (die mit dem neuen Medikament behandelt wurde) sowie die Überlebenszeiten $y_1, \ldots, y_m$ der Kontrollgruppe (die aus Personen besteht, die nicht mit dem neuen Medikament behandelt wurden) ermittelt. Durch Vergleich dieser Überlebenszeiten möchte man feststellen, ob die Einnahme des neuen Medikaments eine Wirkung auf die Überlebenszeit hat oder nicht.*

Zur stochastischen Modellierung fassen wir $x_1, \ldots, x_n$ bzw. $y_1, \ldots, y_m$ als Realisierungen von Zufallsvariablen $X_1, \ldots, X_n$ bzw. $Y_1, \ldots, Y_m$ auf. Hierbei seien die Zufallsvariablen

$$X_1, \ldots, X_n, Y_1, \ldots, Y_m$$

unabhängig, wobei $X_1, \ldots, X_n$ identisch verteilt seien mit Erwartungswert μ_X und $Y_1, \ldots, Y_m$ identisch verteilt seien mit Erwartungswert μ_Y.

Aufgrund der obigen Stichprobe wollen wir uns zwischen der Nullhypothese

$$H_0 : \mu_X = \mu_Y$$

und der Alternativhypothese

$$H_1 : \mu_X \neq \mu_Y.$$

Eine Möglichkeit dafür ist der sogenannte *zweiseitige Gauß-Test für zwei Stichproben* . Bei diesem geht man davon aus, dass die $X_1, \ldots, X_n$ unabhängig identisch $N(\mu_X, \sigma_0^2)$-verteilt sind, und dass die $Y_1, \ldots, Y_m$ unabhängig identisch $N(\mu_Y, \sigma_0^2)$-verteilt sind. Hierbei sind μ_X, μ_Y unbekannt, die Varianz σ_0^2 wird aber als bekannt vorausgesetzt. Man beachte, dass hier insbesondere vorausgesetzt wird, dass die $X_1, \ldots, X_n$ die gleiche Varianz wie die $Y_1, \ldots, Y_m$ haben.

Betrachtet wird hier die Testgröße

$$T(x_1, \ldots, x_n, y_1, \ldots, y_m) = \frac{\bar{x} - \bar{y}}{\sigma_0 \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

wobei

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{und} \quad \bar{x} = \frac{1}{m}\sum_{j=1}^{m} y_j.$$

Ist die Differenz von μ_X und μ_Y betragsmäßig groß, so wird, da $\bar{x}$ und $\bar{y}$ Schätzungen von μ_X bzw. μ_Y sind, auch $T(x_1, \ldots, y_m)$ betragsmäßig groß sein. Dies legt nahe, H_0 abzulehnen, sofern $T(x_1, \ldots, y_m)$ betragsmäßig einen kritischen Wert c übersteigt.

Ausgangspunkt zur Bestimmung von c ist, dass bei Gültigkeit von H_0 (d.h. für $\mu_X = \mu_Y$)

$$T(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_j}{\sigma_0 \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$N(0,1)$-verteilt ist. Dazu beachte man, dass $T(X_1, \ldots, Y_m)$ normalverteilt ist, da Linearkombinationen unabhängiger normalverteilter Zufallsvariablen immer normalverteilt sind. Desweiteren gilt

$$\mathbf{E}T(X_1, \ldots, Y_m) = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}X_i - \frac{1}{m}\sum_{j=1}^{m} \mathbf{E}Y_j}{\sigma_0 \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\mu_X - \mu_Y}{\sigma_0 \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} = 0$$

für $\mu_X = \mu_Y$, sowie

$$\begin{aligned}
V(T(X_1, \ldots, Y_m)) &= \frac{V\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_j\right)}{\sigma_0^2 \cdot \left(\frac{1}{n} + \frac{1}{m}\right)} \\
&= \frac{V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + V\left(\frac{1}{m}\sum_{j=1}^{m} Y_j\right)}{\sigma_0^2 \cdot \left(\frac{1}{n} + \frac{1}{m}\right)} \\
&= \frac{\frac{1}{n^2}\sum_{i=1}^{n} V(X_i) + \frac{1}{m^2}\sum_{j=1}^{m} V(Y_j)}{\sigma_0^2 \cdot \left(\frac{1}{n} + \frac{1}{m}\right)} \\
&= \frac{\frac{\sigma_0^2}{n} + \frac{\sigma_0^2}{m}}{\sigma_0^2 \cdot \left(\frac{1}{n} + \frac{1}{m}\right)} = 1.
\end{aligned}$$

Man wählt nun c als $\alpha/2$-Fraktil der $N(0,1)$-Verteilung. Es gilt dann bei Gültigkeit von H_0: Die Wahrscheinlichkeit, H_0 fälschlicherweise abzulehnen, ist gegeben durch

$$\mathbf{P}[|T(X_1, \ldots, X_n, Y_1, \ldots, Y_m)| > c] = 2 \cdot \mathbf{P}[T(X_1, \ldots, X_n, Y_1, \ldots, Y_m) > c] = \alpha.$$

Damit erhält man als Vorschrift für den zweiseitigen Gauß-Test für zwei Stichproben: Lehne H_0 ab, falls

$$\left| \frac{\bar{x} - \bar{y}}{\sigma_0 \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} \right| > c,$$

wobei $c \in \mathbb{R}$ so gewählt ist, dass für eine $N(0, 1)$ verteilte Zufallsvariable Z gilt:

$$\mathbf{P}[Z > c] = \frac{\alpha}{2},$$

d.h. man wählt c als $\alpha/2$–Fraktil der $N(0, 1)$-Verteilung.

Beim zweiseitigen Gauß-Test für zwei Stichproben wird vorausgesetzt, dass die Varianz σ_0^2 bekannt ist. In Anwendungen ist diese aber üblicherweise unbekannt und muss aus den Daten geschätzt werden.

Beim *zweiseitigen t-Test für zwei Stichproben* geht man davon aus, dass die X_1, $\dots$, X_n, Y_1, $\dots$, Y_m unabhängig sind, wobei die X_1, $\dots$, X_n $N(\mu_X, \sigma^2)$-verteilt und die $Y_1, \dots, Y_m$ $N(\mu_Y, \sigma^2)$-verteilt sind. Hierbei sind μ_X, μ_Y und σ^2 unbekannt. Man beachte, dass wieder vorausgesetzt wird, dass die Varianz der X_i mit der der Y_j übereinstimmt.

Zu testen ist wieder

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y.$$

In einem ersten Schritt schätzt man σ^2 durch die sogenannte *gepoolte Stichprobenvarianz*

$$S_p^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{j=1}^m (Y_j - \bar{Y}_m)^2}{m + n - 2}.$$

Wegen

$$\begin{aligned}
\mathbf{E}[S_p^2] &= \frac{1}{m+n-2} \Bigg((n-1) \cdot \mathbf{E}\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right] \\
&\qquad + (m-1) \cdot \mathbf{E}\left[\frac{1}{m-1} \sum_{j=1}^m (Y_j - \bar{Y}_m)^2 \right] \Bigg) \\
&= \frac{1}{m+n-2} \left((n-1) \cdot \sigma^2 + (m-1) \cdot \sigma^2 \right) = \sigma^2
\end{aligned}$$

(vgl. Beispiel 5.4) handelt es sich hierbei um eine erwartungstreue Schätzung der Varianz.

Man bildet dann analog zum zweiseitigen Gauß-Test für zwei Stichproben die Teststatistik

$$T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{S_p^2} \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

Man kann zeigen, dass bei Gültigkeit von $\mu_X = \mu_Y$ diese Teststatistik t-verteilt ist mit $m+n-2$-Freiheitsgraden. Daher lehnt man beim *zweiseitigen Gauß-Test für zwei Stichproben* $H_0 : \mu_X = \mu_Y$ genau dann ab, falls

$$|T| > t_{m+n-2,\alpha},$$

wobei $t_{m+n-2,\alpha}$ das $\alpha/2$-Fraktil der t-Verteilung mit $m + n - 2$-Freiheitsgraden ist.

Index

Symbols

A

B

D

R

S

T

U

V

W

Z